aws

Developer Guide

Amazon Comprehend

Amazon Comprehend: Developer Guide

Table of Contents

What is Amazon Comprehend?

Amazon Comprehend uses natural language processing (NLP) to extract insights about the content of documents. It develops insights by recognizing the entities, key phrases, language, sentiments, and other common elements in a document. Use Amazon Comprehend to create new products based on understanding the structure of documents. For example, using Amazon Comprehend you can search social networking feeds for mentions of products or scan an entire document repository for key phrases.

You can access Amazon Comprehend document analysis capabilities using the Amazon Comprehend console or using the Amazon Comprehend APIs. You can run real-time analysis for small workloads or you can start asynchronous analysis jobs for large document sets. You can use the pre-trained models that Amazon Comprehend provides, or you can train your own custom models for classification and entity recognition.

Amazon Comprehend may store your content to continuously improve the quality of its pre-trained models. See the Amazon Comprehend FAQ to learn more.

All of the Amazon Comprehend features accept UTF-8 text documents as the input. In addition, custom classification and custom entity recognition accept image files, PDF files, and Word files as input.

Amazon Comprehend can examine and analyze documents in a variety of languages, depending on the specific feature. For more information, see Languages supported in Amazon Comprehend. Amazon Comprehend's Dominant language capability can examine documents and determine the dominant language for a far wider selection of languages.

Topics

Amazon Comprehend insights

Amazon Comprehend uses a pre-trained model to examine and analyze a document or set of documents to gather insights about it. This model is continuously trained on a large body of text so that there is no need for you to provide training data.

Amazon Comprehend analyzes the following types of insights:

- **Entities** – References to the names of people, places, items, and locations contained in a document.
- **Key phrases** – Phrases that appear in a document. For example, a document about a basketball game might return the names of the teams, the name of the venue, and the final score.
- **Personally Identifiable Information (PII)** – Personal data that can identify an individual, such as an address, bank account number, or phone number.
- **Language** – The dominant language of a document.
- **Sentiment** – The dominant sentiment of a document, which can be positive, neutral, negative, or mixed.
- **Targeted sentiment** – The sentiments associated with specific entities in a document. The sentiment for each entity occurrence can be positive, negative, neutral or mixed.
- **Syntax** – The parts of speech for each word in the document.

For more information, see Insights.

Amazon Comprehend Custom

You can customize Amazon Comprehend for your specific requirements without the skillset required to build machine learning-based NLP solutions. Using automatic machine learning, or AutoML, Amazon Comprehend Custom builds customized NLP models on your behalf, using data you already have.

Custom classification – Create custom classification models (classifiers) to organize your documents into your own categories.

Custom entity recognition – Create custom entity recognition models (recognizers) that can analyze text for your specific terms and noun-based phrases.

For more information, see Amazon Comprehend Custom.

Flywheels

Use flywheels to simplify the process of training and managing custom model versions over time. A flywheel helps to orchestrate the tasks associated with training and evaluating new versions of a model. Flywheels support plain-text custom models for custom classification and custom entity recognition. For more information, see Flywheels.

Document clustering (topic modeling)

You can also use Amazon Comprehend to examine a corpus of documents to organize them based on similar keywords within them. Document clustering (topic modeling) is useful to organize a large corpus of documents into topics or clusters that are similar based on word frequency. For more information, see Topic modeling.

Examples

The following examples show how you might use the Amazon Comprehend operations in your applications.

Example 1: Find documents about a subject

Find the documents about a particular subject using Amazon Comprehend topic modeling. Scan a set of documents to determine the topics discussed, and to find the documents associated with each topic. You can specify the number of topics that Amazon Comprehend should return from the document set.

Example 2: Find out how customers feel about your products

If your company publishes a catalog, let Amazon Comprehend tell you what customers think of your products. Send each customer comment to the `DetectSentiment` operation and it will tell you whether customers feel positive, negative, neutral, or mixed about a product.

Example 3: Discover what matters to your customers

Use Amazon Comprehend topic modeling to discover the topics that your customers are talking about on your forums and message boards, then use entity detection to determine the people, places, and things that they associate with the topic. Use sentiment analysis to determine how your customers feel about a topic.

Benefits

Benefits of using Amazon Comprehend include:

- **Integrate powerful natural language processing into your apps** – Amazon Comprehend removes the complexity of building text analysis capabilities into your applications by making powerful and accurate natural language processing available with a simple API. You don't need textual analysis expertise to take advantage of the insights that Amazon Comprehend produces.
- **Deep learning based natural language processing** – Amazon Comprehend uses deep learning technology to accurately analyze text. Our models are constantly trained with new data across multiple domains to improve accuracy.
- **Scalable natural language processing** – Amazon Comprehend enables you to analyze millions of documents so that you can discover the insights that they contain.
- **Integrated with other AWS services** – Amazon Comprehend is designed to work seamlessly with other AWS services like Amazon S3, AWS KMS, and AWS Lambda. Store your documents in Amazon S3, or analyze real-time data with Firehose. Support for AWS Identity and Access Management (IAM) makes it easy to securely control access to Amazon Comprehend operations. Using IAM, you can create and manage users and groups to grant the appropriate access to your developers and end users.
- **Encryption of output results and volume data** – Amazon S3 already enables you to encrypt your input documents, and Amazon Comprehend extends this even farther. By using your own KMS key, you can encrypt the output results of your job and the data on the storage volume attached to the compute instance that processes the analysis job. The result is significantly enhanced security.
- **Low cost** – With Amazon Comprehend, there are no minimum fees or upfront commitments. You pay for the documents that you analyze and custom models that you train.

Amazon Comprehend pricing

With Amazon Comprehend, you pay only for the resources that you use. If you are a new AWS customer, you can get started with Amazon Comprehend for free. For more information, see AWS free usage tier.

There is a usage charge for running real-time or asynchronous analysis jobs. You pay to train custom models, and you pay for custom model management. For real-time requests using custom models, you pay for the endpoint from the time that you start your endpoint until you delete the

endpoint. There is no additional charge for using flywheels. However, when you run a flywheel iteration, you incur the standard charges for training a new model version and storing the model data.

For the rates and additional detailed information, see Amazon Comprehend Pricing.

Are you a first-time user of Amazon Comprehend?

If you are a first-time user of Amazon Comprehend, we recommend that you read the following sections in order:

1. **How it works** – This section introduces Amazon Comprehend concepts.
2. **Setting up** – In this section, you create an account and set up the AWS CLI.
3. **Getting started with Amazon Comprehend** – In this section, you run a Amazon Comprehend analysis job.
4. **Tutorial: Analyzing insights from customer reviews with Amazon Comprehend** – In this section, you perform sentiment and entities analysis and visualize the results.
5. **Amazon Comprehend API Reference** – Reference documentation for Amazon Comprehend operations.

AWS provides the following resources for learning about the Amazon Comprehend service:

- The AWS Machine Learning Blog includes useful articles about Amazon Comprehend.
- Amazon Comprehend Resources provides useful videos and tutorials about Amazon Comprehend.

How it works

Amazon Comprehend uses a pre-trained model to gather **insights** about a document or a set of documents. This model is continuously trained on a large body of text so that there is no need for you to provide training data.

You can use Amazon Comprehend to build your own **custom models** for custom classification and custom entity recognition. You can use Flywheels to help manage the custom models.

Amazon Comprehend provides **topic modeling** using a built-in model. Topic modeling examines a corpus of documents and organizes the documents based on similar keywords within them.

Amazon Comprehend provides synchronous and asynchronous **document processing modes**. Use synchronous mode for processing one document or a batch of up to 25 documents. Use an asynchronous job to process a large number of documents.

Amazon Comprehend works with AWS Key Management Service (AWS KMS) to provide enhanced encryption for your data. For more information, see KMS encryption in Amazon Comprehend.

Key concepts

- Insights
- Amazon Comprehend Custom
- Topic modeling
- Document processing modes

Insights

Amazon Comprehend can analyze a document or set of documents to gather insights about it. Some of the insights that Amazon Comprehend develops about a document include:

- Entities – Amazon Comprehend returns a list of entities, such as people, places, and locations, identified in a document.
- Events – Amazon Comprehend detects specific types of events and related details.
- Key phrases – Amazon Comprehend extracts key phrases that appear in a document. For example, a document about a basketball game might return the names of the teams, the name of the venue, and the final score.

- Personally identifiable information (PII) – Amazon Comprehend analyzes documents to detect personal data that identify an individual, such as an address, bank account number, or phone number.
- Dominant language – Amazon Comprehend identifies the dominant language in a document. Amazon Comprehend can identify 100 languages.
- Sentiment – Amazon Comprehend determines the dominant sentiment of a document. Sentiment can be positive, neutral, negative, or mixed.
- Targeted Sentiment – Amazon Comprehend determines the sentiment of specific entities mentioned in a document. The sentiment of each mention can be positive, neutral, negative, or mixed.
- Syntax analysis – Amazon Comprehend parses each word in your document and determines the part of speech for the word. For example, in the sentence "It is raining today in Seattle," "it" is identified as a pronoun, "raining" is identified as a verb, and "Seattle" is identified as a proper noun.

Entities

An *entity* is a textual reference to the unique name of a real-world object such as people, places, and commercial items, and to precise references to measures such as dates and quantities.

For example, in the text "John moved to 1313 Mockingbird Lane in 2012," "John" might be recognized as a `PERSON`, "1313 Mockingbird Lane" might be recognized as a `LOCATION`, and "2012" might be recognized as a `DATE`.

Each entity also has a score that indicates the level of confidence that Amazon Comprehend has that it correctly detected the entity type. You can filter out the entities with lower scores to reduce the risk of using incorrect detections.

The following table lists the entity types.

Type	Description
COMMERCIAL_ITEM	A branded product
DATE	A full date (for example, 11/25/2017), day (Tuesday), month (May), or time (8:30 a.m.)

Type	Description
EVENT	An event, such as a festival, concert, election, etc.
LOCATION	A specific location, such as a country, city, lake, building, etc.
ORGANIZATION	Large organizations, such as a government, company, religion, sports team, etc.
OTHER	Entities that don't fit into any of the other entity categories
PERSON	Individuals, groups of people, nicknames, fictional characters
QUANTITY	A quantified amount, such as currency, percentages, numbers, bytes, etc.
TITLE	An official name given to any creation or creative work, such as movies, books, songs, etc.

Detect entities operations can be performed using any of the primary languages supported by Amazon Comprehend. This includes only predefined (non-custom) entity detection. All documents must be in the same language.

You can use any of the following API operations to detect entities in a document or set of documents.

- DetectEntities
- BatchDetectEntities
- StartEntitiesDetectionJob

The operations return a list of API Entity objects, one for each entity in the document. The `BatchDetectEntities` operation returns a list of `Entity` objects, one list for each document in the batch. The `StartEntitiesDetectionJob` operation starts an asynchronous job that produces a file containing a list of `Entity` objects for each document in the job.

The following example is the response from the `DetectEntities` operation.

```
{
```

```
    "Entities": [
        {
            "Text": "today",
            "Score": 0.97,
            "Type": "DATE",
            "BeginOffset": 14,
            "EndOffset": 19
        },
        {
            "Text": "Seattle",
            "Score": 0.95,
            "Type": "LOCATION",
            "BeginOffset": 23,
            "EndOffset": 30
        }
    ],
    "LanguageCode": "en"
}
```

Events

Use *event detection* to analyze text documents for specific types of events and their related entities. Amazon Comprehend supports event detection across large collections of documents using asynchronous analysis jobs. For more information about events, including example event analysis jobs, see Announcing the launch of Amazon Comprehend Events

Entities

From the input text, Amazon Comprehend extracts a list of entities that are related to the detected event. An *entity* can be a real-world object, such as a person, place, or location; an entity can also be a concept, such as a measurement, date, or quantity. Each occurrence of an entity is identified by a *mention*, which is a textual reference to the entity in the input text. For each unique entity, all mentions are grouped into a list. This list provides details for each location in the input text where the entity occurs. Amazon Comprehend detects only the entities associated with supported event types.

Each entity associated with a supported event type returns with the following related details:

- **Mentions**: Details for each occurrence of the same entity in the input text.
  - **BeginOffset**: A character offset in the input text that shows where the mention begins (the first character is at position 0).

- **EndOffset**: A character offset in the input text that shows where the mention ends.
- **Score**: The level of confidence that Amazon Comprehend has in the accuracy of the entity's type.
- **GroupScore**: The level of confidence from Amazon Comprehend that the mention is correctly grouped with other mentions of the same entity.
- **Text**: The text of the entity.
- **Type**: The entity's type. For all supported entity types, see Entity types.

Events

Amazon Comprehend returns the list of events (of supported event types) that it detects in the input text. Each event returns with the following related details:

- **Type**: The event's type. For all supported event types, see Event types.
- **Arguments**: A list of arguments that are related to the detected event. An *argument* consists of an entity that is related to the detected event. The argument's role describes the relationship, such as *who* did *what*, *where* and *when*.
  - **EntityIndex**: An index value that identifies an entity from the list of entities that Amazon Comprehend returned for this analysis.
  - **Role**: The argument type, which describes how the entity for this argument is related to the event. For all supported argument types, see Argument types.
  - **Score**: The level of confidence that Amazon Comprehend has in the accuracy of the role detection.
- **Triggers**: A list of triggers for the detected event. A *trigger* is a single word or phrase that indicates the occurrence of the event.
  - **BeginOffset**: A character offset in the input text that shows where the trigger begins (the first character is at position 0).
  - **EndOffset**: A character offset in the input text that shows where the trigger ends.
  - **Score**: The level of confidence that Amazon Comprehend has in the accuracy of the detection.
  - **Text**: The text of the trigger.
  - **GroupScore**: The level of confidence from Amazon Comprehend that the trigger is correctly grouped with other triggers for the same event.
  - **Type**: The type of event that this trigger indicates.

Detect events results format

When your event detection job completes, Amazon Comprehend writes the analysis results to the Amazon S3 output location that you specified when you started the job.

For each detected event, the output provides details in the following format:

```
{
   "Entities": [
     {
       "Mentions": [
         {
           "BeginOffset": number,
           "EndOffset": number,
           "Score": number,
           "GroupScore": number,
           "Text": "string",
           "Type": "string"
         }, ...
       ]
     }, ...
   ],
   "Events": [
     {
       "Type": "string",
       "Arguments": [
         {
           "EntityIndex": number,
           "Role": "string",
           "Score": number
         }, ...
       ],
       "Triggers": [
         {
           "BeginOffset": number,
           "EndOffset": number,
           "Score": number,
           "Text": "string",
           "GroupScore": number,
           "Type": "string"
         }, ...
       ]
     }, ...
```

```
    ]
}
```

Supported types for entities, events, and arguments

Entity types

Type	Description
DATE	Any reference to a date or time, whether specific or general.
FACILITY	Buildings, airports, highways, bridges, and other permanent man-made structures and real estate improvements.
LOCATION	Physical locations such as streets, cities, states, countries, bodies of water, or geographic coordinates.
MONETARY_VALUE	The value of something in US or other currency. The value can be specific or approximate.
ORGANIZATION	Companies and other groups of people defined by an established organizational structure.
PERSON	The names or nicknames of individuals or fictional characters.
PERSON_TITLE	Any title which describes a person, which is usually an employment category (such as CEO) or honorific (such as Mr.).
QUANTITY	A number or value and the unit of measureme nt.
STOCK_CODE	A stock ticker symbol, such as AMZN, an International Securities Identification Number

Type	Description
	(ISIN), Committee on Uniform Securities Identification Procedures (CUSIP), or Stock Exchange Daily Official List (SEDOL).

Event types

Type	Description
BANKRUPTCY	A legal proceeding involving a person or company unable to repay outstanding debts.
EMPLOYMENT	Occurs when an employee is hired, fired, retired, or otherwise changes employment state.
CORPORATE_ACQUISTION	Occurs when a company obtains the possessio n of most or all of another company's shares or physical assets to gain control of that company.
INVESTMENT_GENERAL	Occurs when a person or company purchases an asset with the prospect of generating future income or appreciation.
CORPORATE_MERGER	Occurs when two or more companies unite to create a new legal entity.
IPO	An initial public offering (IPO) of shares of a private corporation to the public in a new stock issuance.
RIGHTS_ISSUE	A group of rights offered to existing sharehold ers to purchase additional stock shares, known as subscription warrants, in proportion to their existing holdings.

Type	Description
SECONDARY_OFFERING	An offer of securities by a shareholder of a company.
SHELF_OFFERING	A Securities and Exchange Commission (SEC) provision that allows an issuer to register a new issue of security and sell portions of the issue over a period of time without re-regist ering the security or incurring penalties. Also known as a shelf registration.
TENDER_OFFERING	An offer to purchase some or all of sharehold ers' shares in a company.
STOCK_SPLIT	Occurs when a company's board of directors increases the number of shares that are outstanding by issuing more shares to current shareholders. This event also applies to reverse stock splits.

Argument types

Argument types for BANKRUPTCY

Argument type	Description
FILER	The person or company filing the bankruptcy.
DATE	The date or time of bankruptcy.
PLACE	Location or facility where (or nearest to where) the bankruptcy took place.

Argument types for EMPLOYMENT

Type	Description
EMPLOYEE	The person employed by a company.
EMPLOYEE_TITLE	The title of the employee.
EMPLOYER	The person or company employing the employee.
START_DATE	The start date or time of the employment.
END_DATE	The end date or time of the employment.

Argument types for CORPORATE_ACQUISTION, INVESTMENT_GENERAL

Type	Description
AMOUNT	The monetary value associated with the transaction.
INVESTEE	The person or company associated with the investment.
INVESTOR	The person or company investing in the asset.
DATE	The date or time of the acquisition or investment.
PLACE	Location where (or nearest to where) the acquisition or investment took place.

Argument types for CORPORATE_MERGER

Type	Description
DATE	The date or time of the merger.

Type	Description
NEW_COMPANY	The new legal entity resulting from the merger.
PARTICIPANT	The company involved in the merger.

Argument types for IPO, RIGHTS_ISSUE, SECONDARY_OFFERING, SHELF_OFFERING, TENDER_OFFERING

Type	Description
EXPIRE_DATE	The expiration date or time of the offering.
INVESTOR	The person or company investing in the asset.
OFFEREE	The person or company receiving the offering.
OFFERING_AMOUNT	The monetary value associated with the offering.
OFFERING_DATE	The date or time of the offering.
OFFEROR	The person or company initiating the offering.
OFFEROR_TOTAL_VALUE	The total monetary value associated with the offering.
RECORD_DATE	The record date or time of the offering.
SELLING_AGENT	The person or company facilitating the sale of the offering.
SHARE_PRICE	The monetary value associated with the share price.
SHARE_QUANTITY	The number of shares associated with the offering.

Type	Description
UNDERWRITERS	The company associated with the underwriting of the offering.

Argument types for STOCK_SPLIT

Type	Description
COMPANY	The company issuing shares of the stock split.
DATE	The date or time of the stock split.
SPLIT_RATIO	The ratio of the increased new number of shares outstanding to the current number of shares before the stock split.

Key phrases

A *key phrase* is a string containing a noun phrase that describes a particular thing. It generally consists of a noun and the modifiers that distinguish it. For example, "day" is a noun; "a beautiful day" is a noun phrase that includes an article ("a") and an adjective ("beautiful"). Each key phrase includes a score that indicates the level of confidence that Amazon Comprehend has that the string is a noun phrase. You can use the score to determine if the detection has high enough confidence for your application.

Detect key phrases operations can be performed using any of the primary languages supported by Amazon Comprehend. All documents must be in the same language.

You can use any of the following operations to detect key phrases in a document or set of documents.

- DetectKeyPhrases
- BatchDetectKeyPhrases
- StartKeyPhrasesDetectionJob

The operations return a list of KeyPhrase objects, one for each key phrase in the document. The `BatchDetectKeyPhrases` operation returns a list of `KeyPhrase` objects, one for each document in the batch. The `StartKeyPhrasesDetectionJob` operation starts an asynchronous job that produces a file containing a list of `KeyPhrase` objects for each document in the job.

The following example is the response from the `DetectKeyPhrases` operation.

```
{
    "LanguageCode": "en",
    "KeyPhrases": [
        {
            "Text": "today",
            "Score": 0.89,
            "BeginOffset": 14,
            "EndOffset": 19
        },
        {
            "Text": "Seattle",
            "Score": 0.91,
            "BeginOffset": 23,
            "EndOffset": 30
        }
    ]
}
```

Dominant language

You can use Amazon Comprehend to examine text to determine the dominant language. Amazon Comprehend identifies the language using identifiers from RFC 5646 — if there is a 2-letter ISO 639-1 identifier, with a regional subtag if necessary, it uses that. Otherwise, it uses the ISO 639-2 3-letter code.

For more information about RFC 5646, see Tags for identifying languages on the *IETF Tools* web site.

The response includes a score that indicates the confidence level that Amazon Comprehend has that a particular language is the dominant language in the document. Each score is independent of the other scores. The score doesn't indicate that a language makes up a particular percentage of a document.

If a long document (such as a book) contains multiple languages, you can break the long document into smaller pieces and run the `DetectDominantLanguage` operation on the individual pieces. You can then aggregate the results to determine the percentage of each language in the longer document.

Amazon Comprehend language detection has the following limitations:

- It doesn't support phonetic language detection. For example, it doesn't detect "arigato" as Japanese or "nihao" as Chinese.
- It may have diffuculty distinguishing close language pairs, such as Indonesian and Malay; or Bosnian, Croatian, and Serbian.
- For best results, provide at least 20 characters of input text.

Amazon Comprehend detects the following languages.

Code	Language
af	Afrikaans
am	Amharic
ar	Arabic
as	Assamese
az	Azerbaijani
ba	Bashkir
be	Belarusian
bn	Bengali
bs	Bosnian
bg	Bulgarian
ca	Catalan
ceb	Cebuano

Code	Language
cs	Czech
cv	Chuvash
cy	Welsh
da	Danish
de	German
el	Greek
en	English
eo	Esperanto
et	Estonian
eu	Basque
fa	Persian
fi	Finnish
fr	French
gd	Scottish Gaelic
ga	Irish
gl	Galician
gu	Gujarati
ht	Haitian
he	Hebrew
ha	Hausa

Code	Language
hi	Hindi
hr	Croatian
hu	Hungarian
hy	Armenian
ilo	Iloko
id	Indonesian
is	Icelandic
it	Italian
jv	Javanese
ja	Japanese
kn	Kannada
ka	Georgian
kk	Kazakh
km	Central Khmer
ky	Kirghiz
ko	Korean
ku	Kurdish
lo	Lao
la	Latin
lv	Latvian

Code	Language
lt	Lithuanian
lb	Luxembourgish
ml	Malayalam
mt	Maltese
mr	Marathi
mk	Macedonian
mg	Malagasy
mn	Mongolian
ms	Malay
my	Burmese
ne	Nepali
new	Newari
nl	Dutch
no	Norwegian
or	Oriya
om	Oromo
pa	Punjabi
pl	Polish
pt	Portuguese
ps	Pushto

Code	Language
qu	Quechua
ro	Romanian
ru	Russian
sa	Sanskrit
si	Sinhala
sk	Slovak
sl	Slovenian
sd	Sindhi
so	Somali
es	Spanish
sq	Albanian
sr	Serbian
su	Sundanese
sw	Swahili
sv	Swedish
ta	Tamil
tt	Tatar
te	Telugu
tg	Tajik
tl	Tagalog

Code	Language
th	Thai
tk	Turkmen
tr	Turkish
ug	Uighur
uk	Ukrainian
ur	Urdu
uz	Uzbek
vi	Vietnamese
yi	Yiddish
yo	Yoruba
zh	Chinese (Simplified)
zh-TW	Chinese (Traditional)

You can use any of the following operations to detect the dominant language in a document or set of documents.

- DetectDominantLanguage
- BatchDetectDominantLanguage
- StartDominantLanguageDetectionJob

The `DetectDominantLanguage` operation returns a DominantLanguage object. The `BatchDetectDominantLanguage` operation returns a list of `DominantLanguage` objects, one for each document in the batch. The `StartDominantLanguageDetectionJob` operation starts an asynchronous job that produces a file containing a list of `DominantLanguage` objects, one for each document in the job.

The following example is the response from the `DetectDominantLanguage` operation.

```
{
    "Languages": [
        {
            "LanguageCode": "en",
            "Score": 0.9793661236763
        }
    ]
}
```

Sentiment

Use Amazon Comprehend to determine the sentiment of content in UTF-8 encoded text documents. For example, you can use sentiment analysis to determine the sentiments of comments on a blog posting to determine if your readers liked the post.

You can determine sentiment for documents in any of the primary languages supported by Amazon Comprehend. All documents in one job must be in the same language.

Sentiment determination returns the following values:

- **Positive** – The text expresses an overall positive sentiment.
- **Negative** – The text expresses an overall negative sentiment.
- **Mixed** – The text expresses both positive and negative sentiments.
- **Neutral** – The text does not express either positive or negative sentiments.

You can use any of the following API operations to detect the sentiment of a document or a set of documents.

- DetectSentiment
- BatchDetectSentiment
- StartSentimentDetectionJob

The operations return the most likely sentiment for the text and the scores for each of the sentiments. The score represents the likelihood that the sentiment was correctly detected. For example, in the example below it is 95 percent likely that the text has a `Positive` sentiment.

There is a less than 1 percent likelihood that the text has a `Negative` sentiment. You can use the `SentimentScore` to determine if the accuracy of the detection meets the needs of your application.

The `DetectSentiment` operation returns an object that contains the detected sentiment and a SentimentScore object. The `BatchDetectSentiment` operation returns a list of sentiments and `SentimentScore` objects, one for each document in the batch. The `StartSentimentDetectionJob` operation starts an asynchronous job that produces a file containing a list of sentiments and `SentimentScore` objects, one for each document in the job.

The following example is the response from the `DetectSentiment` operation.

```
{
"SentimentScore": {
        "Mixed": 0.030585512690246105,
        "Positive": 0.94992071056365967,
        "Neutral": 0.0141543131828308,
        "Negative": 0.00893945890665054
    },
    "Sentiment": "POSITIVE",
    "LanguageCode": "en"
}
```

Targeted sentiment

Targeted sentiment provides a granular understanding of the sentiments associated with specific entities (such as brands or products) in your input documents.

The difference between targeted sentiment and sentiment is the level of granularity in the output data. Sentiment analysis determines the dominant sentiment for each input document, but doesn't provide data for further analysis. Targeted sentiment analysis determines the entity-level sentiment for specific entities in each input document. You can analyze the output data to determine the specific products and services that get positive or negative feedback.

For example, in a set of restaurant reviews, a customer provides the following review: "The tacos were delicious and the staff was friendly." Analysis of this review produces the following results:

- **Sentiment analysis** determines whether the overall sentiment of each restaurant review is positive, negative, neutral, or mixed. In this example, the overall sentiment is positive.

- **Targeted sentiment analysis** determines sentiment for entities and attributes of the restaurant that customers mention in the reviews. In this example, the customer made positive comments about “tacos” and “staff”.

Targeted sentiment provides the following outputs for each analysis job:

- Identity of the entities mentioned in the documents.
- Classification of the entity type for each entity mention.
- The sentiment and a sentiment score for each entity mention.
- Groups of mentions (co-reference groups) that correspond to a single entity.

You can use the console or the API to run targeted sentiment analysis. The console and the API support real-time analysis and asynchronous analysis for targeted sentiment.

Amazon Comprehend supports targeted sentiment for documents in the English language.

For additional information about targeted sentiment, including a tutorial, see Extract granular sentiment in text with Amazon Comprehend Targeted Sentiment in the AWS machine learning blog.

Topics

Entity types

Targeted sentiment identifies the following entity types. It assigns entity type OTHER if the entity doesn’t belong in any other category. Each entity mention in the output file includes the entity type, such as `"Type": "PERSON"`.

Entity type definitions

Entity Type	Definition
PERSON	Examples include individuals, groups of people, nicknames, fictional characters, and animal names.
LOCATION	Geographical locations such as countries, cities, states, addresses, geological formations, bodies of water, natural landmarks, and astronomical locations.
ORGANIZATION	Examples include governments, companies, sports teams, and religions.
FACILITY	Buildings, airports, highways, bridges, and other permanent man-made structures and real estate improvements.
BRAND	Organization, group, or producer of a specific commercial item or line of products.
COMMERCIAL_ITEM	Any non-generic purchasable or acquirable item, including vehicles, and large products that had only one item produced.
MOVIE	A movie or television show. Entity could be the full name, a nickname, or a subtitle.
MUSIC	A song, full or partial. Also, collections of individual music creations, such as an album or an anthology.
BOOK	A book, published professionally or self-published.
SOFTWARE	An officially released software product.
GAME	A game, such as video games, board games, common games, or sports.
PERSONAL_TITLE	Official titles and honorifics such as President, PhD, or Dr.
EVENT	Examples include festival, concert, election, war, conference, and promotional event.
DATE	Any reference to a date or time, whether specific or general, whether absolute or relative.

Entity Type	Definition
QUANTITY	All measurements along with their units (currency, percent, number, bytes, etc.).
ATTRIBUTE	An attribute, characteristic, or trait of an entity, such as the "quality" of a product, the "price" of a phone, or the "speed" of a CPU.
OTHER	Entities that don't belong in any of the other categories.

Co-reference group

Targeted sentiment identifies co-reference groups in each input document. A co-reference group is a group of mentions in a document that correspond to one real-world entity.

Example

In the following example of a customer review, "spa" is the entity, which has entity type `FACILITY`. The entity has two additional mentions as a pronoun ("it").

Output file organization

The targeted sentiment analysis job creates a JSON text output file. The file contains one JSON object for each of the input documents. Each JSON object contains the following fields:

- **Entities** – An array of entities found in the document.
- **File** – The file name of the input document.

- **Line** – If the input file is one document per line, **Entities** contains the line number of the document in the file.

> **Note**
>
> If targeted sentiment doesn't identify any entities in the input text, it returns an empty array as the Entities result.

The following example shows **Entities** for an input file with three lines of input. The input format is **ONE_DOC_PER_LINE**, so each line of input is a document.

```
{ "Entities":[
    {entityA},
    {entityB},
    {entityC}
    ],
  "File": "TargetSentimentInputDocs.txt",
  "Line": 0
}
{ "Entities": [
    {entityD},
    {entityE}
  ],
  "File": "TargetSentimentInputDocs.txt",
  "Line": 1
}
{ "Entities": [
    {entityF},
    {entityG}
    ],
  "File": "TargetSentimentInputDocs.txt",
  "Line": 2
}
```

An entity in the **Entities** array includes a logical grouping (called a co-reference group) of the entity mentions detected in the document. Each entity has the following overall structure:

```
{"DescriptiveMentionIndex": [0],
  "Mentions": [
```

```
      {mentionD},
      {mentionE}
     ]
}
```

An entity contains these fields:

- **Mentions** – An array of mentions of the entity in the document. The array represents a co-reference group. See the section called "Co-reference group" for an example. The order of mentions in the Mentions array is the order of their location (offset) in the document. Each mention includes the sentiment score and group score for that mention. The group score indicates the confidence level that these mentions belong to the same entity.
- **DescriptiveMentionIndex** – One or more index into the Mentions array that provides the best name for the entity group. For example, an entity could have three mentions with **Text** values "ABC Hotel," "ABC Hotel," and "it." The best name is "ABC Hotel," which has a DescriptiveMentionIndex value of [0,1].

Each mention includes the following fields

- **BeginOffset** – The offset into the document text where the mention begins.
- **EndOffset** – The offset into the document text where the mention ends.
- **GroupScore** – The confidence that all the entities mentioned in the group relate to the same entity.
- **Text** – The text in the document that identifies the entity.
- **Type** – The type of the entity. Amazon Comprehend supports a variety of entity types.
- **Score** – Model confidence that the entity is relevant. Value range is zero to one, where one is highest confidence.
- **MentionSentiment** – Contains the sentiment and sentiment score for the mention.
- **Sentiment** – The sentiment of the mention. Values include: POSITIVE, NEUTRAL, NEGATIVE, and MIXED.
- **SentimentScore** – Provides model confidence for each of the possible sentiments. Value range is zero to one, where one is highest confidence.

The **Sentiment** values have the following meaning:

- **Positive** – The entity mention expresses a positive sentiment.

- **Negative** – The entity mention expresses a negative sentiment.
- **Mixed** – The entity mention expresses both positive and negative sentiments.
- **Neutral** – The entity mention does not express either positive or negative sentiments.

In the following example, an entity has only one mention in the input document, so the DescriptiveMentionIndex is zero (the first mention in the Mentions array). The identified entity is a PERSON with the name "I." The sentiment score is neutral.


```
{"Entities":[
  {
    "DescriptiveMentionIndex": [0],
    "Mentions": [
      {
       "BeginOffset": 0,
       "EndOffset": 1,
       "Score": 0.999997,
       "GroupScore": 1,
       "Text": "I",
       "Type": "PERSON",
       "MentionSentiment": {
         "Sentiment": "NEUTRAL",
         "SentimentScore": {
           "Mixed": 0,
           "Negative": 0,
           "Neutral": 1,
           "Positive": 0
         }
       }
     }
   ]
  }
 ],
 "File": "Input.txt",
 "Line": 0
}
```


Real time analysis using the console

You can use the Amazon Comprehend console to run the section called “Targeted sentiment” in real-time. Use the sample text or paste your own text into the input text box, then choose **Analyze**.

In the **Insights** panel, the console displays three views of the targeted sentiment analysis:

- **Analyzed text** – Displays the analyzed text and underlines each entity. The color of the underline indicates the sentiment value (positive, neutral, negative, or mixed) that the analysis assigned to the entity. The console displays the color mappings at the top right corner of the analzed text box. If you hover your cursor over an entity, the console displays a popup panel containing analysis values (entity type, sentiment score) for the entity.
- **Results** – Displays a table containing a row for each entity mention identified in the text. For each entity, the table shows the entity and entity score. The row also includes the primary sentiment and the score for each sentiment value. If there are multiple mentions of the same entity, known as a the section called "Co-reference group", the table displays these mentions as a collapsible set of rows associated with the main entity.

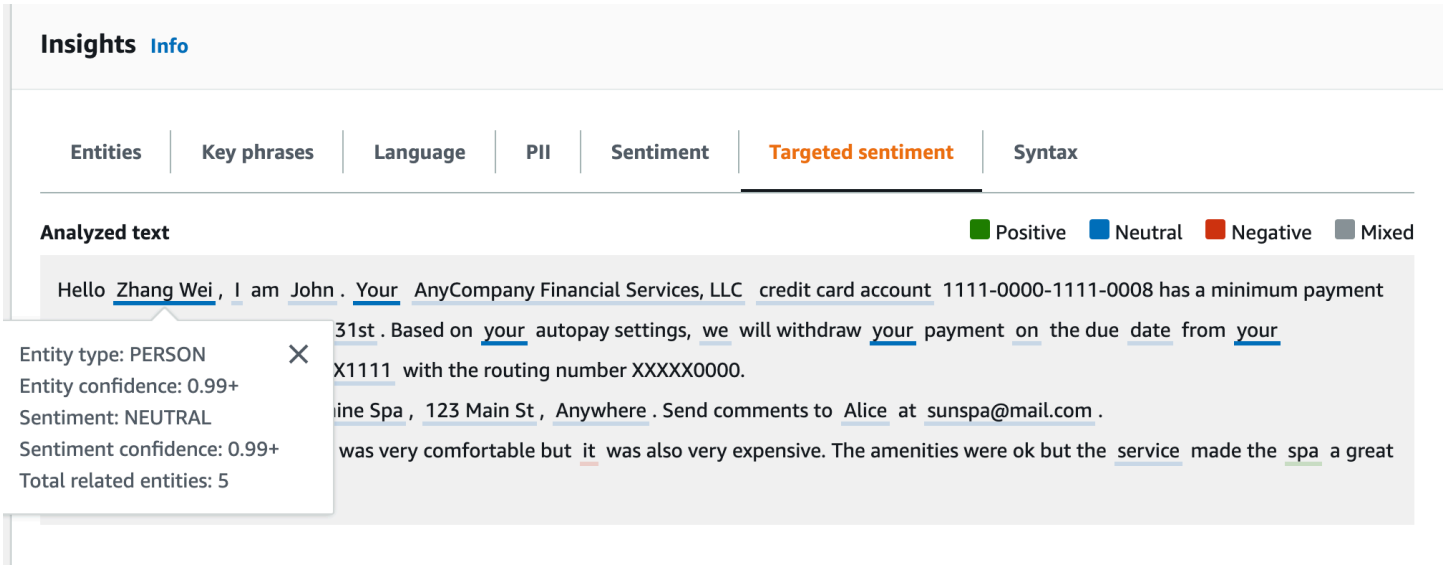
  If you hover over an entity row in the **Results** table, the console highlights the entity mention in the **Analyzed text** panel.
- **Application integration** – Displays the parameter values of the API request and the structure of the JSON object returned in the API response. For a description of the fields in the JSON object, see the section called "Output file organization".

Console real-time analysis example

This example uses the following text as input, which is the default input text that the console provides.

```
Hello Zhang Wei, I am John. Your AnyCompany Financial Services, LLC credit card account
 1111-0000-1111-0008 has a minimum payment
  of $24.53 that is due by July 31st. Based on your autopay settings, we will withdraw
 your payment on the due date from your
  bank account number XXXXXX1111 with the routing number XXXXX0000.
  Customer feedback for Sunshine Spa, 123 Main St, Anywhere. Send comments to Alice at
 sunspa@mail.com.
  I enjoyed visiting the spa. It was very comfortable but it was also very expensive.
 The amenities were ok but the service made
  the spa a great experience.
```

The **Analyzed text** panel shows the following output for this example. Hover your mouse over the text `Zhang Wei` to view the popup panel for this entity.



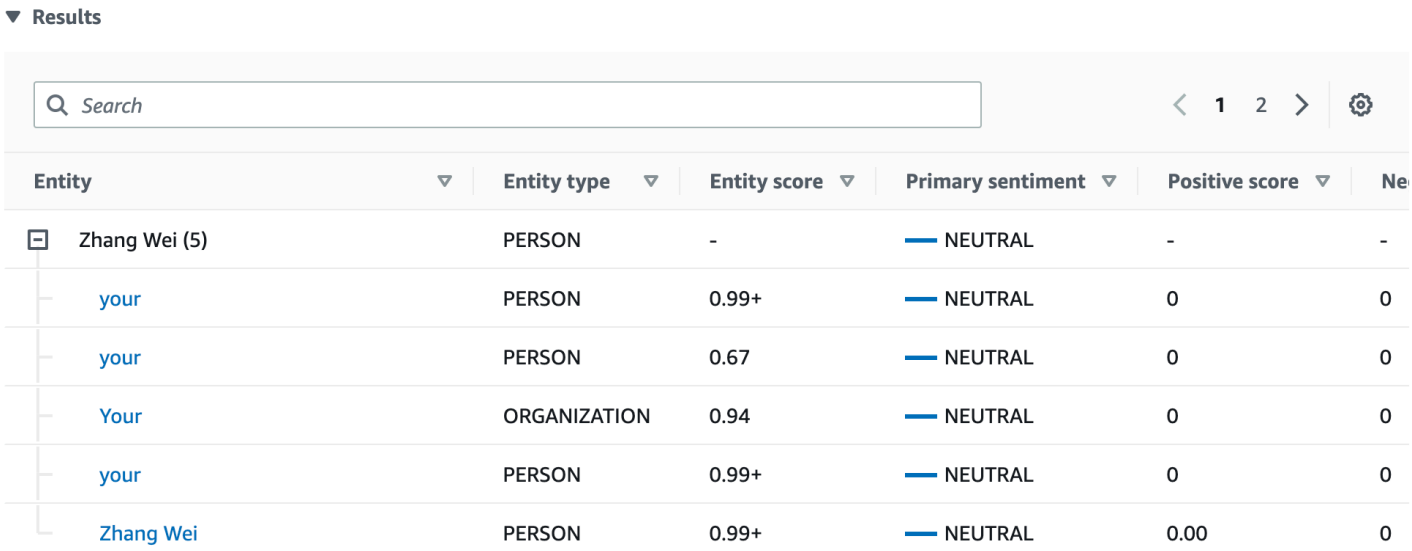
The **Results** table provides additional detail about each entity, including the entity score, the primary sentiment, and the score for each sentiment.

▼ Results

Search 1 2

Entity	Entity type	Entity score	Primary sentiment	Positive score	Ne
Zhang Wei (5)	PERSON	-	NEUTRAL	-	-
John (3)	PERSON	-	NEUTRAL	-	-
AnyCompany Financial Services, LLC (2)	ORGANIZATION	-	NEUTRAL	-	-
credit card account	OTHER	0.99+	NEUTRAL	0.00	0.0
$24.53	QUANTITY	0.99+	NEUTRAL	0.00	0.0
by July 31st (3)	DATE	-	NEUTRAL	-	-
bank account	OTHER	0.99+	NEUTRAL	0.00	0.0
XXXXXX1111	OTHER	0.51	NEUTRAL	0.00	0.0
Customer	PERSON	0.98	NEUTRAL	0	0
Sunshine Spa (5)	FACILITY	-	MIXED	-	-

In our example, targeted sentiment analysis recognizes that each mention of **your** in the input text is a reference to the person entity **Zhang Wei**. The console displays these mentions as a set of collapsible rows associated with the main entity.



The **Application integration** panel displays the JSON object that the DetectTargetedSentiment API generates. See the following section for a full example.

Targeted sentiment output example

The following example shows the output file from a targeted sentiment analysis job. The input file consists of three simple documents:

```
The burger was very flavorful and the burger bun was excellent. However, customer
 service was slow.
My burger was good, and it was warm. The burger had plenty of toppings.
The burger was cooked perfectly but it was cold. The service was OK.
```

The targeted sentiment analysis of this input file produces the following output.

```
{"Entities":[
  {
    "DescriptiveMentionIndex": [
      0
    ],
    "Mentions": [
      {
        "BeginOffset": 4,
        "EndOffset": 10,
        "Score": 0.999991,
        "GroupScore": 1,
        "Text": "burger",
```

```
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "POSITIVE",
            "SentimentScore": {
              "Mixed": 0,
              "Negative": 0,
              "Neutral": 0,
              "Positive": 1
            }
          }
        }
      ]
    },
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 38,
          "EndOffset": 44,
          "Score": 1,
          "GroupScore": 1,
          "Text": "burger",
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "NEUTRAL",
            "SentimentScore": {
              "Mixed": 0.000005,
              "Negative": 0.000005,
              "Neutral": 0.999591,
              "Positive": 0.000398
            }
          }
        }
      ]
    },
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 45,
```

```
          "EndOffset": 48,
          "Score": 0.961575,
          "GroupScore": 1,
          "Text": "bun",
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "POSITIVE",
            "SentimentScore": {
              "Mixed": 0.000327,
              "Negative": 0.000286,
              "Neutral": 0.050269,
              "Positive": 0.949118
            }
          }
        }
      ]
    },
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 73,
          "EndOffset": 89,
          "Score": 0.999988,
          "GroupScore": 1,
          "Text": "customer service",
          "Type": "ATTRIBUTE",
          "MentionSentiment": {
            "Sentiment": "NEGATIVE",
            "SentimentScore": {
              "Mixed": 0.000001,
              "Negative": 0.999976,
              "Neutral": 0.000017,
              "Positive": 0.000006
            }
          }
        }
      ]
    }
  ],
  "File": "TargetSentimentInputDocs.txt",
  "Line": 0
```

```
}
{
  "Entities": [
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 0,
          "EndOffset": 2,
          "Score": 0.99995,
          "GroupScore": 1,
          "Text": "My",
          "Type": "PERSON",
          "MentionSentiment": {
            "Sentiment": "NEUTRAL",
            "SentimentScore": {
              "Mixed": 0,
              "Negative": 0,
              "Neutral": 1,
              "Positive": 0
            }
          }
        }
      ]
    },
    {
      "DescriptiveMentionIndex": [
        0,
        2
      ],
      "Mentions": [
        {
          "BeginOffset": 3,
          "EndOffset": 9,
          "Score": 0.999999,
          "GroupScore": 1,
          "Text": "burger",
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "POSITIVE",
            "SentimentScore": {
              "Mixed": 0.000002,
```

```
            "Negative": 0.000001,
            "Neutral": 0.000003,
            "Positive": 0.999994
          }
        }
      },
      {
        "BeginOffset": 24,
        "EndOffset": 26,
        "Score": 0.999756,
        "GroupScore": 0.999314,
        "Text": "it",
        "Type": "OTHER",
        "MentionSentiment": {
          "Sentiment": "POSITIVE",
          "SentimentScore": {
            "Mixed": 0,
            "Negative": 0.000003,
            "Neutral": 0.000006,
            "Positive": 0.999991
          }
        }
      },
      {
        "BeginOffset": 41,
        "EndOffset": 47,
        "Score": 1,
        "GroupScore": 0.531342,
        "Text": "burger",
        "Type": "OTHER",
        "MentionSentiment": {
          "Sentiment": "POSITIVE",
          "SentimentScore": {
            "Mixed": 0.000215,
            "Negative": 0.000094,
            "Neutral": 0.00008,
            "Positive": 0.999611
          }
        }
      }
    ]
  },
  {
    "DescriptiveMentionIndex": [
```

```
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 52,
          "EndOffset": 58,
          "Score": 0.965462,
          "GroupScore": 1,
          "Text": "plenty",
          "Type": "QUANTITY",
          "MentionSentiment": {
            "Sentiment": "NEUTRAL",
            "SentimentScore": {
              "Mixed": 0,
              "Negative": 0,
              "Neutral": 1,
              "Positive": 0
            }
          }
        }
      ]
    },
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 62,
          "EndOffset": 70,
          "Score": 0.998353,
          "GroupScore": 1,
          "Text": "toppings",
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "NEUTRAL",
            "SentimentScore": {
              "Mixed": 0,
              "Negative": 0,
              "Neutral": 0.999964,
              "Positive": 0.000036
            }
          }
        }
```

```
      ]
    }
  ],
  "File": "TargetSentimentInputDocs.txt",
  "Line": 1
}
{
  "Entities": [
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 4,
          "EndOffset": 10,
          "Score": 1,
          "GroupScore": 1,
          "Text": "burger",
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "POSITIVE",
            "SentimentScore": {
              "Mixed": 0.001515,
              "Negative": 0.000822,
              "Neutral": 0.000243,
              "Positive": 0.99742
            }
          }
        },
        {
          "BeginOffset": 36,
          "EndOffset": 38,
          "Score": 0.999843,
          "GroupScore": 0.999661,
          "Text": "it",
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "NEGATIVE",
            "SentimentScore": {
              "Mixed": 0,
              "Negative": 0.999996,
              "Neutral": 0.000004,
              "Positive": 0
```

```
            }
          }
        }
      ]
    },
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 53,
          "EndOffset": 60,
          "Score": 1,
          "GroupScore": 1,
          "Text": "service",
          "Type": "ATTRIBUTE",
          "MentionSentiment": {
            "Sentiment": "NEUTRAL",
            "SentimentScore": {
              "Mixed": 0.000033,
              "Negative": 0.000089,
              "Neutral": 0.993325,
              "Positive": 0.006553
            }
          }
        }
      ]
    }
  ],
  "File": "TargetSentimentInputDocs.txt",
  "Line": 2
}
  }
```


Syntax analysis

Use syntax analysis to parse the words from the document and return the part of speech, or syntactic function, for each word in the document. You can identify the nouns, verbs, adjectives and so on in your document. Use this information to gain a richer understanding of the content of your documents, and to understand the relationship of the words in the document.

For example, you can look for the nouns in a document and then look for the verbs related to those nouns. In a sentence like "My grandmother moved her couch" you can see the nouns, "grandmother" and "couch," and the verb, "moved." You can use this information to build applications for analyzing text for word combinations that you are interested in.

To start the analysis, Amazon Comprehend parses the source text to find the individual words in the text. After the text is parsed, each word is assigned the part of speech that it takes in the source text.

Amazon Comprehend can identify the following parts of speech.

Token	Part of speech
ADJ	Adjective Words that typically modify nouns.
ADP	Adposition The head of a prepositional or postpositional phrase.
ADV	Adverb Words that typically modify verbs. They may also modify adjectives and other adverbs.
AUX	Auxiliary Function words that accompanies the verb of a verb phrase.
CCONJ	Coordinating conjunction A coordinating conjunction connects words, phrases, or clauses in a sentence without

Token	Part of speech
	subordinating one to the other.
CONJ	Conjunction A conjunction connects words, phrases, or clauses in a sentence.
DET	Determiner Articles and other words that specify a particular noun phrase.
INTJ	Interjection Words used as an exclamation or part of an exclamation.
NOUN	Noun Words that specify a person, place, thing, animal, or idea.
NUM	Numeral Words, typically determiners, adjectives, or pronouns, that express a number.
O	Other Words that can't be assigned a part of speech category.

Token	Part of speech
PART	Particle Function words associated with another word or phrase to impart meaning.
PRON	Pronoun Words that substitute for nouns or noun phrases.
PROPN	Proper noun A noun that is the name of a specific individual, place or object.
PUNCT	Punctuation Non-alphabetical characters that delimit text.
SCONJ	Subordinating conjunction A conjunction that joins a dependent clause to a sentence. An example of a subordinating conjunction is "because".
SYM	Symbol Word-like entities such as the dollar sign ($) or mathemati cal symbols.

Token	Part of speech
VERB	Verb Words that signal events and actions.

For more information about the parts of speech, see Universal POS tags at the *Universal Dependencies* website.

The operations return tokens that identify the word and the part of speech that the word represents in the text. Each token represents a word in the source text. It provides the location of the word in the source, the part of speech that the word takes in the text, the confidence that Amazon Comprehend has that the part of speech was correctly identified, and the word that was parsed from the source text.

The following is the structure of the list of syntax tokens. One syntax token is generated for each word in the document.

```
{
   "SyntaxTokens": [
      {
         "BeginOffset": number,
         "EndOffset": number,
         "PartOfSpeech": {
            "Score": number,
            "Tag": "string"
         },
         "Text": "string",
         "TokenId": number
      }
   ]
}
```

Each token provides the following information:

- `BeginOffset` and `EndOffset`—Provides the location of the word in the input text.

- `PartOfSpeech`—Provides two pieces of information, the `Tag` that identifies the part of speech and the `Score` that represents the confidence that Amazon Comprehend Syntax has that the part of speech was correctly identifies.
- `Text`—Provides the word that was identified.
- `TokenId`—Provides an identifier for the token. The identifier is the position of the token in the list of tokens.

Amazon Comprehend Custom

You can customize Amazon Comprehend for your specific requirements without the skillset required to build machine learning-based NLP solutions. Using automatic machine learning, or AutoML, Comprehend Custom builds customized NLP models on your behalf, using training data that you provide.

Input document processing – Amazon Comprehend supports one-step document processing for custom classification and custom entity recognition. For example, you can input a mix of plain text documents and semi-structured documents (such as PDF documents, Microsoft Word documents, and images) to a custom analysis job. For more information, see Document processing.

Custom classification – Create custom classification models (classifiers) to organize your documents into your own categories. For each classification label, provide a set of documents that best represent that label and train your classifier on it. Once trained, a classifier can be used on any number of unlabeled document sets. You can use the console for a code-free experience or install the latest AWS SDK. For more information, see Custom classification.

Custom entity recognition – Create custom entity recognition models (recognizers) that can analyze text for your specific terms and noun-based phrases. You can train recognizers to extract terms like policy numbers, or phrases that imply a customer escalation. To train the model, you provide a list of the entities and a set of documents that contain them. Once the model is trained, you can submit analysis jobs against it to extract their custom entities. For more information, see Custom entity recognition.

Topic modeling

You can use Amazon Comprehend to examine the content of a collection of documents to determine common themes. For example, you can give Amazon Comprehend a collection of news

articles, and it will determine the subjects, such as sports, politics, or entertainment. The text in the documents doesn't need to be annotated.

Amazon Comprehend uses a Latent dirichlet allocation-based learning model to determine the topics in a set of documents. It examines each document to determine the context and meaning of a word. The set of words that frequently belong to the same context across the entire document set make up a topic.

A word is associated to a topic in a document based on how prevalent that topic is in a document and how much affinity the topic has to the word. The same word can be associated with different topics in different documents based on the topic distribution in a particular document.

For example, the word "glucose" in an article that talks predominantly about sports can be assigned to the topic "sports," while the same word in an article about "medicine" will be assigned to the topic "medicine."

Each word associated with a topic is given a weight that indicates how much the word helps define the topic. The weight is an indication of how many times the word occurs in the topic compared to other words in the topic, across the entire document set.

For the most accurate results you should provide Amazon Comprehend with the largest possible corpus to work with. For best results:

- You should use at least 1,000 documents in each topic modeling job.
- Each document should be at least 3 sentences long.
- If a document consists of mostly numeric data, you should remove it from the corpus.

Topic modeling is an asynchronous process. You submit your list of documents to Amazon Comprehend from an Amazon S3 bucket using the StartTopicsDetectionJob operation. The response is sent to an Amazon S3 bucket. You can configure both the input and output buckets. Get a list of the topic modeling jobs that you have submitted using the ListTopicsDetectionJobs operation and view information about a job using the DescribeTopicsDetectionJob operation. Content delivered to Amazon S3 buckets might contain customer content. For more information about removing sensitive data, see How Do I Empty an S3 Bucket? or How Do I Delete an S3 Bucket?.

Documents must be in UTF-8 formatted text files. You can submit your documents two ways. The following table shows the options.

Format	Description
One document per file	Each file contains one input document. This is best for collections of large documents.
One document per line	The input is a single file. Each line in the file is considered a document. This is best for short documents, such as social media postings. Each line must end with a line feed (LF, \n), a carriage return (CR, \r), or both (CRLF, \r\n). The Unicode line separator (u+2028) can't be used to end a line.

For more information, see the InputDataConfig data type.

After Amazon Comprehend processes your document collection, it returns a compressed archive containing two files, `topic-terms.csv` and `doc-topics.csv`. For more information about the output file, see OutputDataConfig.

The first output file, `topic-terms.csv`, is a list of topics in the collection. For each topic, the list includes, by default, the top terms by topic according to their weight. For example, if you give Amazon Comprehend a collection of newspaper articles, it might return the following to describe the first two topics in the collection:

Topic	Term	Weight
000	team	0.118533
000	game	0.106072
000	player	0.031625
000	season	0.023633
000	play	0.021118
000	yard	0.024454

Topic	Term	Weight
000	coach	0.016012
000	games	0.016191
000	football	0.015049
000	quarterback	0.014239
001	cup	0.205236
001	food	0.040686
001	minutes	0.036062
001	add	0.029697
001	tablespoon	0.028789
001	oil	0.021254
001	pepper	0.022205
001	teaspoon	0.020040
001	wine	0.016588
001	sugar	0.015101

The weights represent a probability distribution over the words in a given topic. Since Amazon Comprehend returns only the top 10 words for each topic the weights won't sum to 1.0. In the rare cases where there are less than 10 words in a topic, the weights will sum to 1.0.

The words are sorted by their discriminative power by looking at their occurrence across all topics. Typically this is the same as their weight, but in some cases, such as the words "play" and "yard" in the table, this results in an order that is not the same as the weight.

You can specify the number of topics to return. For example, if you ask Amazon Comprehend to return 25 topics, it returns the 25 most prominent topics in the collection. Amazon Comprehend

can detect up to 100 topics in a collection. Choose the number of topics based on your knowledge of the domain. It may take some experimentation to arrive at the correct number.

The second file, `doc-topics.csv`, lists the documents associated with a topic and the proportion of the document that is concerned with the topic. If you specified `ONE_DOC_PER_FILE` the document is identified by the file name. If you specified `ONE_DOC_PER_LINE` the document is identified by the file name and the 0-indexed line number within the file. For example, Amazon Comprehend might return the following for a collection of documents submitted with one document per file:

Document	Topic	Proportion
sample-doc1	000	0.999330137
sample-doc2	000	0.998532187
sample-doc3	000	0.998384574
...		
sample-docN	000	3.57E-04

Amazon Comprehend utilizes information from the *Lemmatization Lists Dataset by MBM*, which is made available here under the Open database license (ODbL) v1.0.

Document processing modes

Amazon Comprehend supports three document processing modes. Your choice of mode depends on the number documents you need to process and how immediately you need to view the results:

- **Single-document synchronous** – You call Amazon Comprehend with a single document and receive a synchronous response, delivered to your application (or the console) right away.
- **Multi-document synchronous** – You call the Amazon Comprehend API with a collection of up to 25 documents and receive a synchronous response.
- **Asynchronous batch** – For a large collection of documents, put the documents into an Amazon S3 bucket and start an asynchronous job (using console or API operations) to analyze the documents. Amazon Comprehend stores the results of the analysis in the S3 bucket/folder that you specify in the request.

Topics

Single-document processing

Single-document operations are synchronous operations that return the results of the document analysis directly to your application. Use single-document synchronous operations when you are creating an interactive application that works on one document at a time.

For more information about the synchronous API operations, see Real-time analysis using the built-in models (for console) and Real-time analysis using the API.

Multiple document synchronous processing

When you have multiple documents that you want to process, you can use the `Batch*` API operations to send more than one document to Amazon Comprehend at a time. You can send up to 25 documents in each request. Amazon Comprehend sends back a list of responses, one for each document in the request. Requests made with these operations are synchronous. Your application calls the operation and then waits for the response from the service.

Using the `Batch*` operations is identical to calling the single document APIs for each of the documents in the request. Using these APIs can result in better performance for your applications.

The input to each of the APIs is a JSON structure containing the documents to process. For all operations except `BatchDetectDominantLanguage`, you must set the input language. You can set only one input language for each request. For example, the following is the input to the `BatchDetectEntities` operation. It contains two documents and is in English.

```
{
   "LanguageCode": "en",
   "TextList": [
      "I have been living in Seattle for almost 4 years",
      "It is raining today in Seattle"
   ]
}
```

The response from a `Batch*` operation contains two lists, the `ResultList` and the `ErrorList`. The `ResultList` contains one record for each document that was successfully processed. The result for each document in the request is identical to the result you would get if you ran a single document operation on the document. The results for each document are assigned an index based on the order of the documents in the input file. The response from the `BatchDetectEntities` operation is:

```
{
   "ResultList"  : [
      {
         "Index": 0,
         "Entities": [
            {
               "Text": "Seattle",
               "Score": 0.95,
               "Type": "LOCATION",
               "BeginOffset": 22,
               "EndOffset": 29
            },
            {
               "Text": "almost 4 years",
               "Score": 0.89,
               "Type": "QUANTITY",
               "BeginOffset": 34,
               "EndOffset": 48
            }
         ]
      },
      {
         "Index": 1,
         "Entities": [
            {
              "Text": "today",
              "Score": 0.87,
              "Type": "DATE",
              "BeginOffset": 14,
              "EndOffset": 19
            },
            {
               "Text": "Seattle",
               "Score": 0.96,
               "Type": "LOCATION",
               "BeginOffset": 23,
```

```
                "EndOffset": 30
            }
          ]
       }
   ],
   "ErrorList": []
}
```

When an error occurs in the request the response contains an `ErrorList` that identifies the documents that contained an error. The document is identified by its index in the input list. For example, the following input to the `BatchDetectLanguage` operation contains a document that cannot be processed:

```
{
   "TextList": [
     "hello friend",
     "$$$$$$",
     "hola amigo"
   ]
}
```

The response from Amazon Comprehend includes an error list that identifies the document that contained an error:

```
{
    "ResultList": [
        {
          "Index": 0,
          "Languages":[
            {
              "LanguageCode":"en",
              "Score": 0.99
            }
          ]
        },
        {
          "Index": 2
          "Languages":[
            {
              "LanguageCode":"es",
              "Score": 0.82
            }
```

```
          ]
        }
    ],
    "ErrorList": [
      {
        "Index": 1,
        "ErrorCode": "InternalServerException",
        "ErrorMessage": "Unexpected Server Error. Please try again."
      }
    ]
}
```

For more information about the synchronous batch API operations, see Real-time batch APIs.

Asynchronous batch processing

To analyze large documents and large collections of documents, use the Amazon Comprehend asynchronous operations.

To analyze a collection of documents, you typically perform the following steps:

1. Store the documents in an Amazon S3 bucket.
2. Start one or more analysis jobs to analyze the documents.
3. Monitor the progress of the analysis jobs.
4. Retrieve the results of the analysis from an S3 bucket when the job is complete.

For more information about using the asynchronous API operations, see Running analysis jobs using the console (console) and Async analysis jobs using the API.

Languages supported in Amazon Comprehend

Amazon Comprehend supports a wide variety of languages for its various features. The languages supported and the features that support them can be seen in the following tables.

Topics

Supported languages

Amazon Comprehend (except the **detect dominant language** feature) supports the following languages for one or more features.

Code	Language
de	German
en	English
es	Spanish
it	Italian
pt	Portuguese
fr	French
ja	Japanese
ko	Korean
hi	Hindi
ar	Arabic
zh	Chinese (simplified)
zh-TW	Chinese (traditional)

Note

Amazon Comprehend identifies the language using identifiers from RFC 5646 — if there is a 2-letter ISO 639-1 identifier, with a regional subtag./ If necessary, it uses that. Otherwise, it uses the ISO 639-2 3-letter code.
For more information about RFC 5646, see Tags for identifying languages on the *IETF Tools* web site.

Languages supported by Amazon Comprehend features

Feature	Supported languages
Dominant language	See Dominant language.
Entities	All supported languages.
Key phrases	All supported languages.
Detecting PII entities	English and Spanish.
Labeling PII entities	English and Spanish.
Sentiment	All supported languages.
Targeted sentiment	English.
Syntax analysis	German (de), English (en), Spanish (es), French (fr), Italian (it), and Portuguese (pt).
Topic modeling	Not dependent on the language used. Doesn't support character-based languages such as Chinese, Japanese, and Korean.
Custom classification	Plain-text models support the following languages: German (de), English (en), Spanish (es), French (fr), Italian (it), and Portuguese (pt).

Feature	Supported languages
	Native document models support English documents only.
Custom entity recognition	German (de), English (en), Spanish (es), French (fr), Italian (it), and Portuguese (pt). Custom Entity Recognition for PDF and Word supports English documents only.

Setting up

Before you use Amazon Comprehend for the first time, complete the following tasks.

Setting up tasks

- Sign up for an AWS account
- Create a user with administrative access
- Set up the AWS Command Line Interface (AWS CLI)
- Grant programmatic access

Sign up for an AWS account

If you do not have an AWS account, complete the following steps to create one.

To sign up for an AWS account

1. Open https://portal.aws.amazon.com/billing/signup.
2. Follow the online instructions.

   Part of the sign-up procedure involves receiving a phone call and entering a verification code on the phone keypad.

   When you sign up for an AWS account, an *AWS account root user* is created. The root user has access to all AWS services and resources in the account. As a security best practice, assign administrative access to a user, and use only the root user to perform tasks that require root user access.

AWS sends you a confirmation email after the sign-up process is complete. At any time, you can view your current account activity and manage your account by going to https://aws.amazon.com/ and choosing **My Account**.

Create a user with administrative access

After you sign up for an AWS account, secure your AWS account root user, enable AWS IAM Identity Center, and create an administrative user so that you don't use the root user for everyday tasks.

Secure your AWS account root user

1. Sign in to the AWS Management Console as the account owner by choosing **Root user** and entering your AWS account email address. On the next page, enter your password.

   For help signing in by using root user, see Signing in as the root user in the *AWS Sign-In User Guide*.

2. Turn on multi-factor authentication (MFA) for your root user.

   For instructions, see Enable a virtual MFA device for your AWS account root user (console) in the *IAM User Guide*.

Create a user with administrative access

1. Enable IAM Identity Center.

   For instructions, see Enabling AWS IAM Identity Center in the *AWS IAM Identity Center User Guide*.

2. In IAM Identity Center, grant administrative access to a user.

   For a tutorial about using the IAM Identity Center directory as your identity source, see Configure user access with the default IAM Identity Center directory in the *AWS IAM Identity Center User Guide*.

Sign in as the user with administrative access

- To sign in with your IAM Identity Center user, use the sign-in URL that was sent to your email address when you created the IAM Identity Center user.

  For help signing in using an IAM Identity Center user, see Signing in to the AWS access portal in the *AWS Sign-In User Guide*.

Assign access to additional users

1. In IAM Identity Center, create a permission set that follows the best practice of applying least-privilege permissions.

   For instructions, see Create a permission set in the *AWS IAM Identity Center User Guide*.

2. Assign users to a group, and then assign single sign-on access to the group.

   For instructions, see Add groups in the *AWS IAM Identity Center User Guide*.

Set up the AWS Command Line Interface (AWS CLI)

You don't need the AWS CLI to perform the steps in the Getting Started exercises. However, some of the other exercises in this guide do require it. If you prefer, you can skip this step and go to Getting started with Amazon Comprehend, and set up the AWS CLI later.

To install and configure the AWS CLI

1. Install the AWS CLI. For instructions, see the following topic in the *AWS Command Line Interface User Guide*:

   Installing or updating the latest version of the AWS Command Line Interface
2. Configure the AWS CLI. For instructions, see the following topic in the *AWS Command Line Interface User Guide*:

   Configuring the AWS Command Line Interface

Grant programmatic access

Users need programmatic access if they want to interact with AWS outside of the AWS Management Console. The way to grant programmatic access depends on the type of user that's accessing AWS.

To grant users programmatic access, choose one of the following options.

Which user needs programmatic access?	To	By
Workforce identity (Users managed in IAM Identity Center)	Use temporary credentials to sign programmatic requests to the AWS CLI, AWS SDKs, or AWS APIs.	Following the instructions for the interface that you want to use. • For the AWS CLI, see Configuring the AWS

Which user needs programmatic access?	To	By
		CLI to use AWS IAM Identity Center in the *AWS Command Line Interface User Guide*. • For AWS SDKs, tools, and AWS APIs, see IAM Identity Center authentication in the *AWS SDKs and Tools Reference Guide*.
IAM	Use temporary credentials to sign programmatic requests to the AWS CLI, AWS SDKs, or AWS APIs.	Following the instructions in Using temporary credentials with AWS resources in the *IAM User Guide*.
IAM	(Not recommended) Use long-term credentials to sign programmatic requests to the AWS CLI, AWS SDKs, or AWS APIs.	Following the instructions for the interface that you want to use. • For the AWS CLI, see Authenticating using IAM user credentials in the *AWS Command Line Interface User Guide*. • For AWS SDKs and tools, see Authenticate using long-term credentials in the *AWS SDKs and Tools Reference Guide*. • For AWS APIs, see Managing access keys for IAM users in the *IAM User Guide*.

Getting started with Amazon Comprehend

The following exercise uses the Amazon Comprehend console to create and run an asynchronous entity detection job. This exercise assumes that you are familiar with Amazon Simple Storage Service (Amazon S3). For a simpler example, see Real-time analysis using the built-in models.

To create a entity detection job

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Analysis Jobs** and then choose **Create job**.
3. Under **Job settings**, give the job a name. The name must be unique within the Region and account.
4. For **Analysis Type**, choose **Entities**.
5. For **Language**, choose the language of the input documents.
6. Under **Input data**, for **Data source**, choose **Example documents**. The console sets **S3 location** to be the folder containing the public samples.
7. Under **Output data**, in **S3 location**, paste the URL or folder location in Amazon S3 for the output files.
8. Under **Access permissions** section, select **Create an IAM role**. The console creates a new IAM role with the proper permissions for Amazon Comprehend to access the input and output buckets.
9. When you have finished filling out the form, choose **Create job** to create and start the topic detection job.

   The new job appears in the job list with the status field showing the status of the job. The field can be `IN_PROGRESS` for a job that is processing, `COMPLETED` for a job that has finished successfully, and `FAILED` for a job that has an error.
10. Choose the job to open the **Job details panel**.
11. Under **Output**, in **Output data location** choose the link to open the Amazon S3 console.
12. In the Amazon S3 console, choose **Download** and save the `output.tar.gz` file.
13. Decompress the file and save it as a Json file.
14. See the section called “Entities” for a description of the entity types and the fields for each detected entity.

Analysis using the Amazon Comprehend console

You can use the Amazon Comprehend console to analyze documents in real-time or to run asynchronous analysis jobs.

Using real-time analysis with built-in models, you can recognize entities, extract key phrases, detect primary language, detect PII, determine sentiment, analyze targeted sentiment, and analyze syntax.

You can run analysis jobs using the built-in models to find insights such as entities, events, phrases, primary language, sentiment, targeted sentiment, and personally identifiable information (PII). You can also run topic-modeling jobs.

The console also supports real-time and asynchronous analysis using custom models. For more information, see Custom classification and Custom entity recognition.

Topics

Real-time analysis using the built-in models

You can use the Amazon Comprehend console to run real-time analysis of a UTF-8 encoded text document. The document can be English or one of the other languages supported by Amazon Comprehend. The results are shown in the console so that you can review the analysis.

To start analyzing documents, sign in to the AWS Management Console and open the Amazon Comprehend console.

You can replace the sample text with your own text and then choose **Analyze** to get an analysis of your text. Below the text being analyzed, the **Results** pane shows more information about the text.

Run real-time analysis using the built-in model

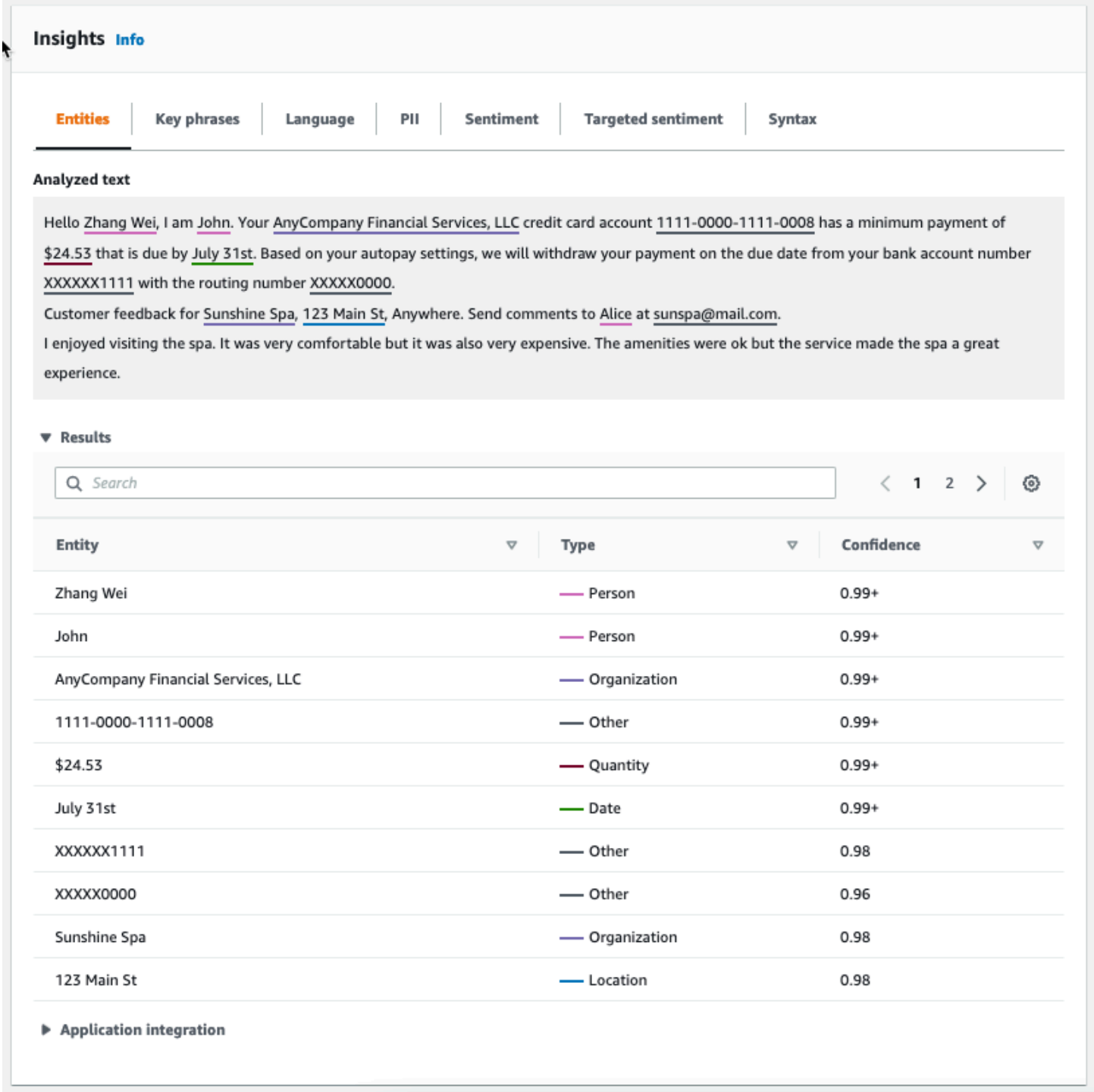
1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Real-time analysis**.

3. Under **Input type**, choose **Built-in** for **Analysis type**.
4. Enter the text you want to analyze.
5. Choose **Analyze**. The console displays the text analysis results in the **Insights** panel. The **Insights** panel includes a tab for each of the insight types. The following sections describe the results for insight type.

Topics

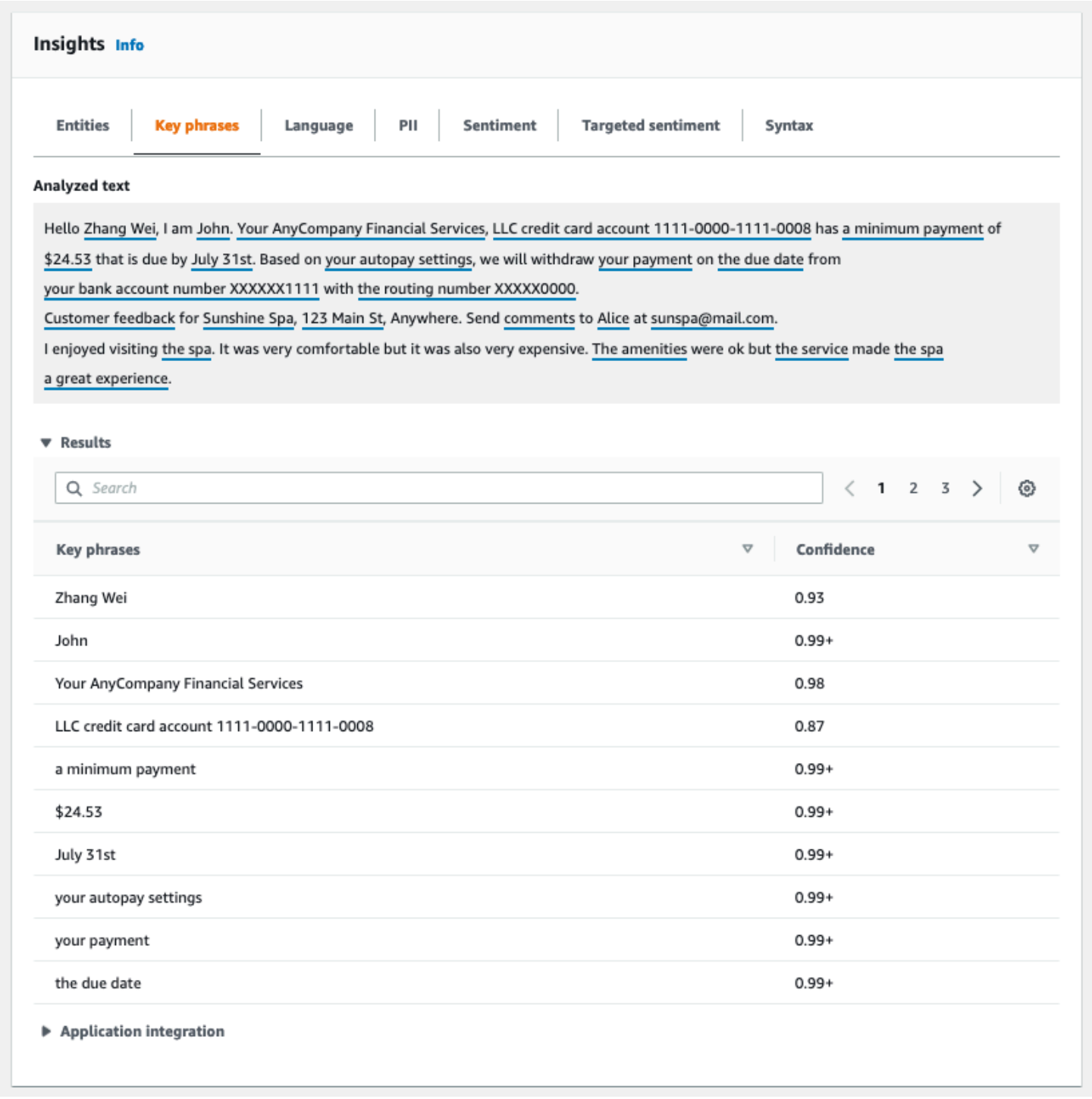
Entities

The **Entities** tab lists each entity, its category, and the level of confidence that Amazon Comprehend has detected in the input text. The results are color-coded to indicate different entity types such as organizations, locations, dates, and persons. For more information, see Entities.



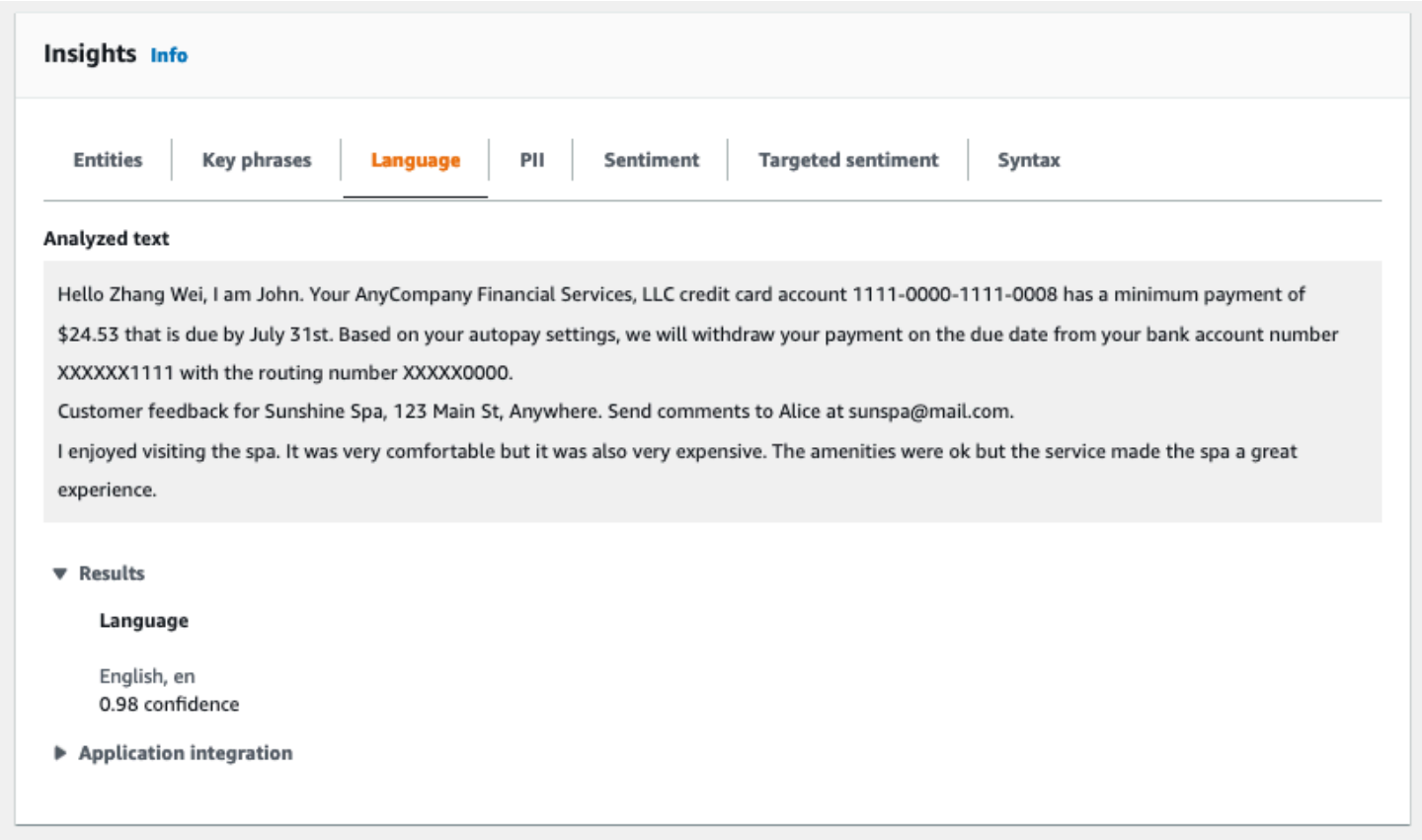
Key phrases

The **Key phrases** tab lists key noun phrases that Amazon Comprehend detected in the input text and the associated confidence level. For more information, see Key phrases.



Language

The **Language** tab shows the dominant language of the text and Amazon Comprehend's level of confidence that it has detected the dominant language correctly. Amazon Comprehend can recognize 100 languages. For more information, see Dominant language.



Personally identifiable information (PII)

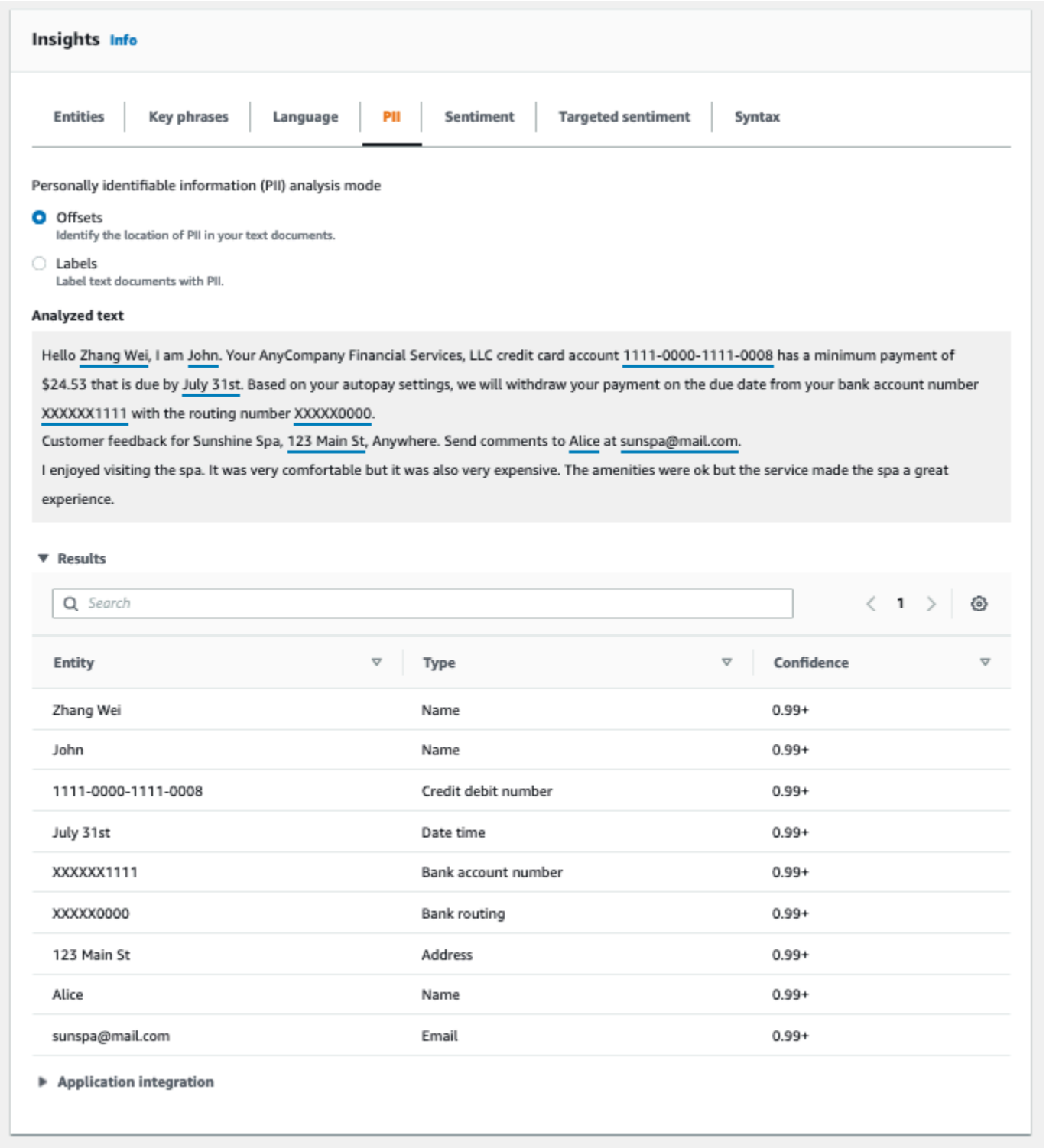
The **PII** tab lists entities in your input text that contain personally identifiable information (PII). A PII entity is a textual reference to personal data that could be used to identify an individual, such as an address, bank account number, or phone number. For more information, see Detecting PII entities.

The **PII** tab provides two analysis modes:

- Offsets
- Labels

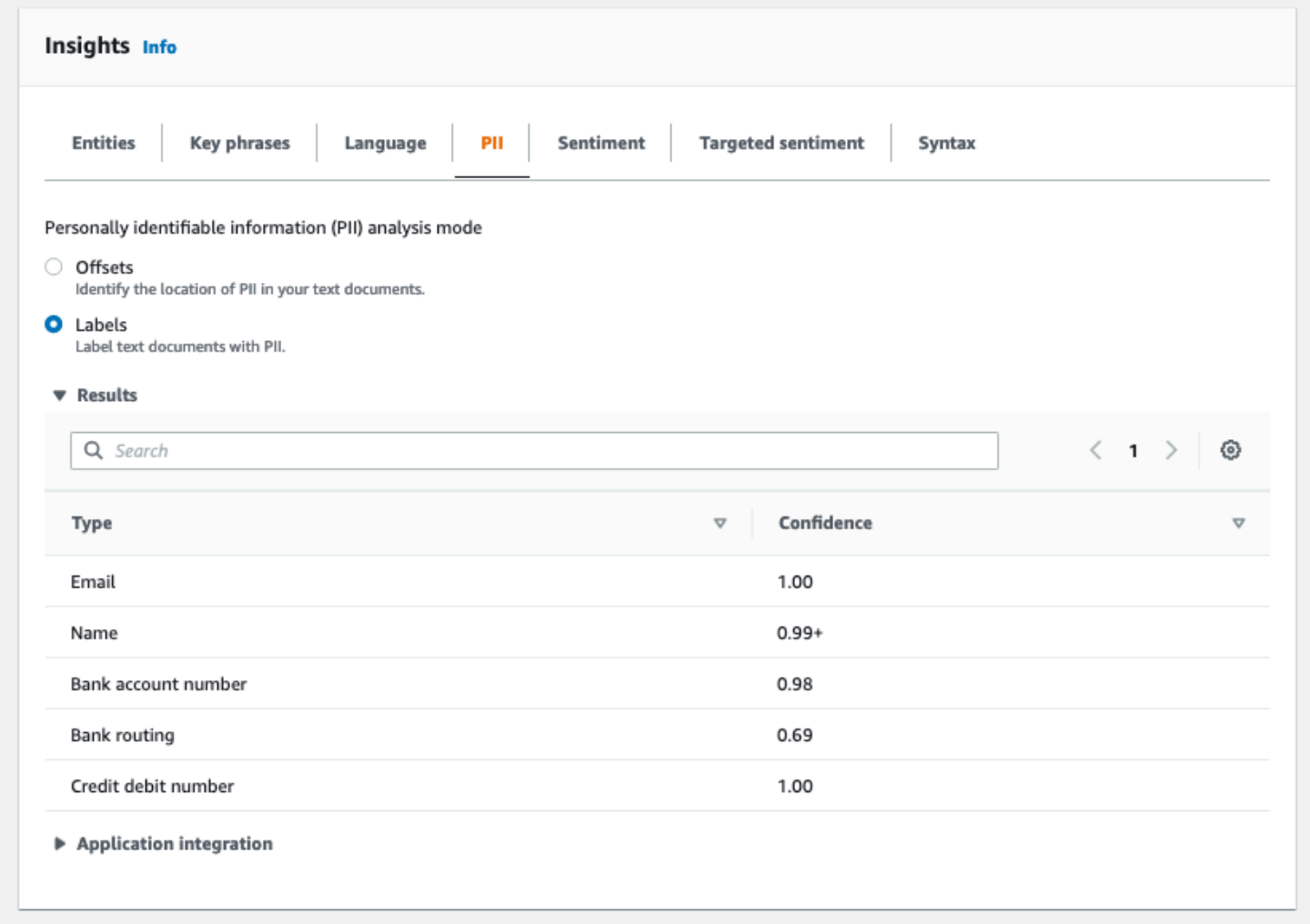
Offsets

The **Offsets** analysis mode identifies the location of PII in your text documents. For more information, see Locate PII entities.



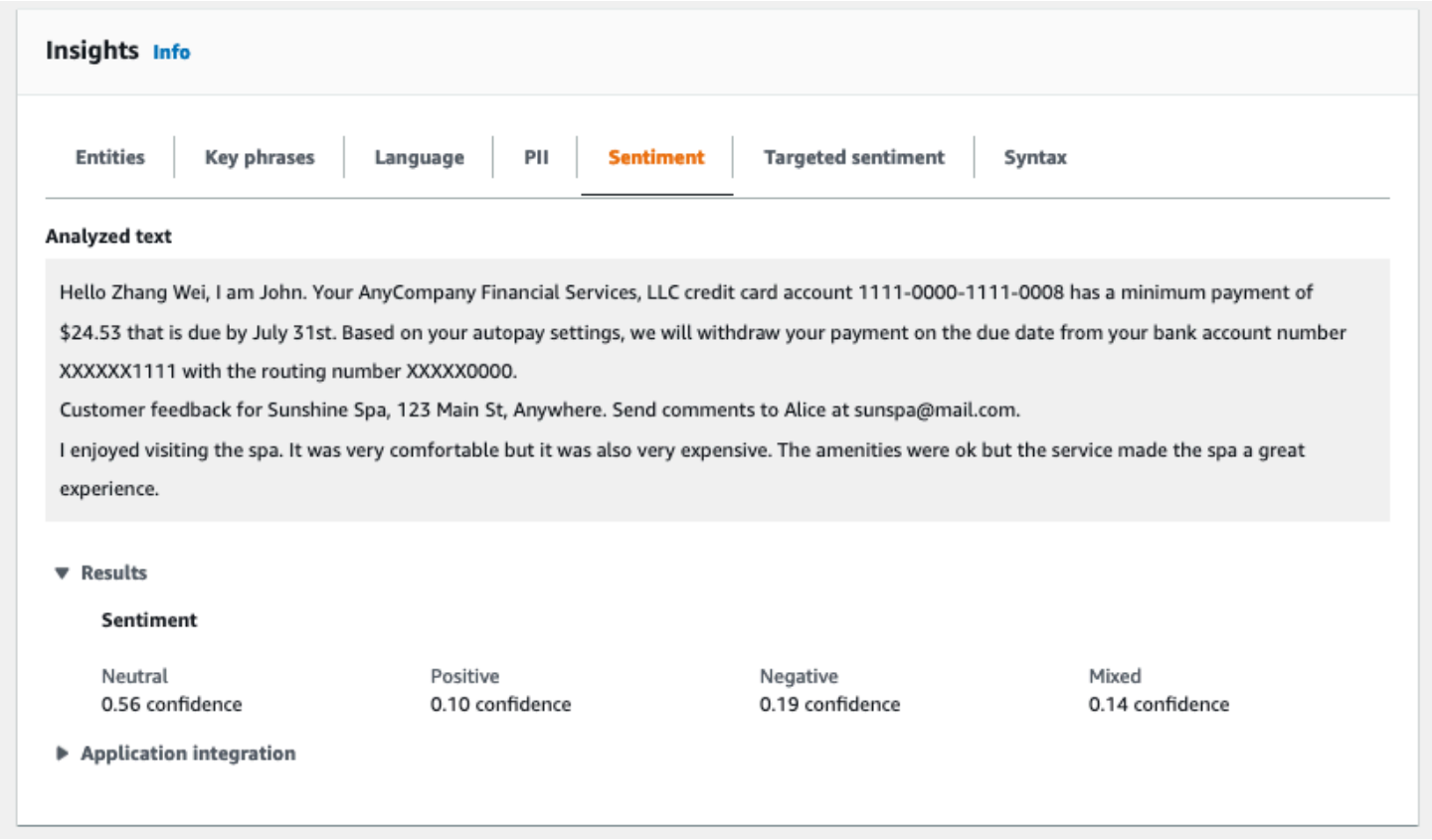
Labels

The **Labels** analysis mode checks for the presence of PII in your text document and returns the labels of identified PII entity types. For more information, see Labeling PII entities.



Sentiment

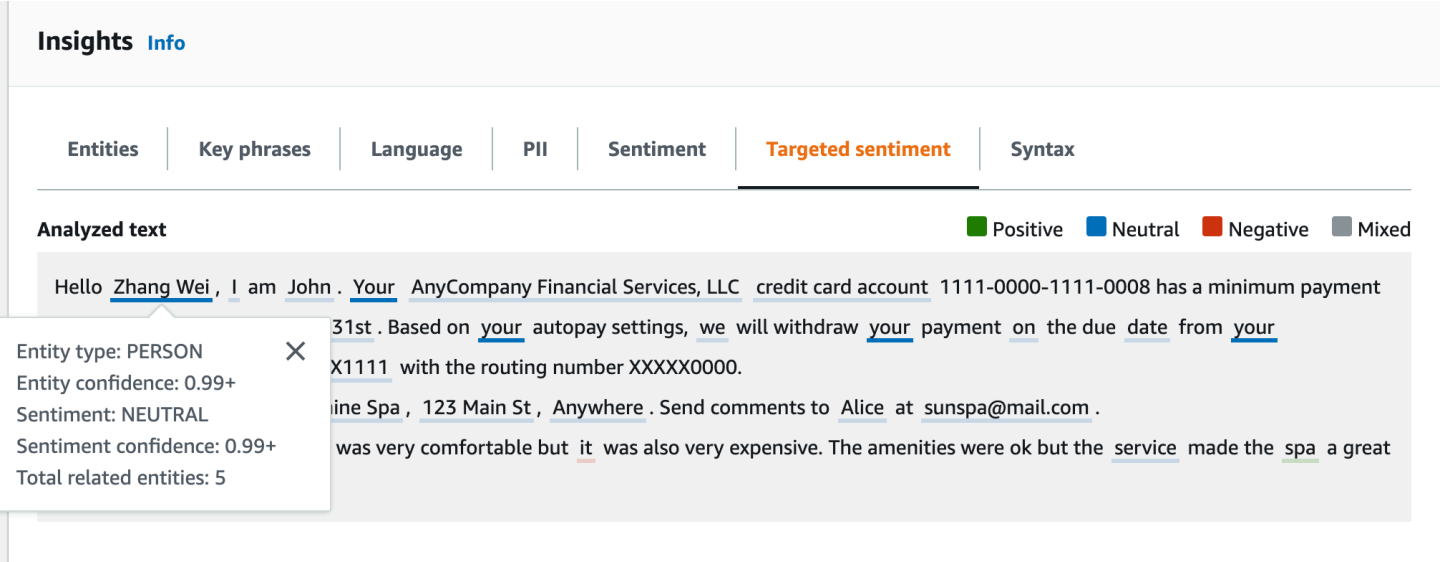
The **Sentiment** tab shows the dominant sentiment of the text. Sentiment can be rated neutral, positive, negative, or mixed. In this case, each sentiment has a confidence rating, providing an estimate by Amazon Comprehend for that sentiment being dominant. For more information, see Sentiment.



Targeted sentiment

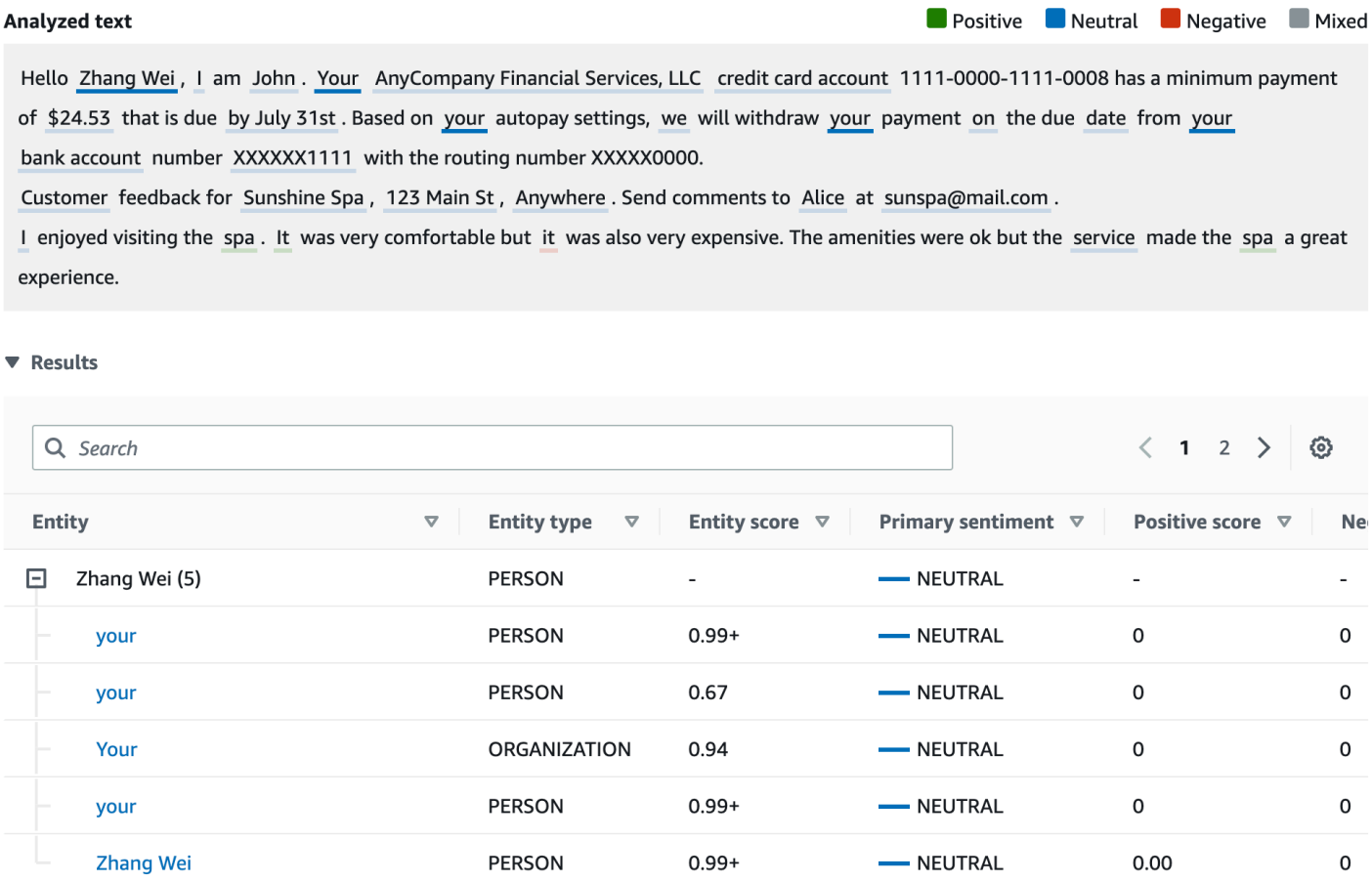
Targeted sentiment analysis identifies the sentiments expressed about entities mentioned in the text. Amazon Comprehend assigns a sentiment rating to each mention of an entity, along with a confidence rating and other information. A sentiment rating can be neutral, positive, negative, or mixed.

In the **Analyzed text** panel, the console underlines each of analyzed entities. The color of the underlined text indicates the overall sentiment of the entity. If you hover your cursor over an entity, the console displays additional information in a pop-up window.



The **Results** table provides additional detail about each entity. If there are multiple mentions of the same entity, called a co-reference group, the table displays these mentions as a collapsible set of rows associated with the main entity.

In the following example, the entity is a person named **Zhang Wei**. The targeted sentiment analysis recognizes that each mention of **your** is a reference to the same person. The console displays these mentions as sub-entries of the main entity.

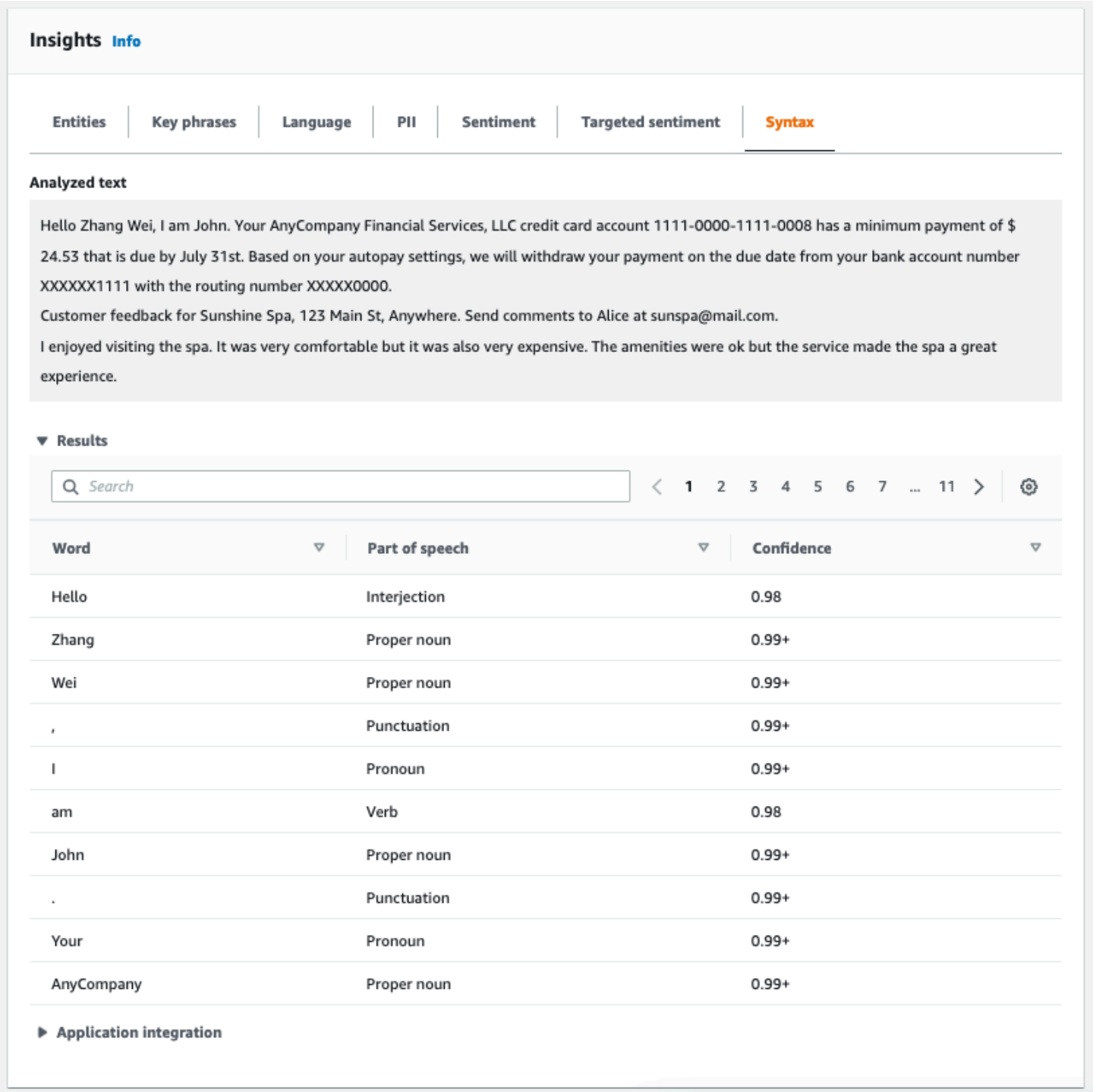


If the text you are analyzing doesn't include any targeted sentiment Entity types, the targeted sentiment analysis displays an empty results field.

For more information about how to use the console for targeted sentiment real-time analysis, see Real time analysis using the console.

Syntax

The **Syntax** tab shows a breakdown of each element in the text, along with its part of speech and the associated confidence score. For more information, see Syntax analysis.



Running analysis jobs using the console

You can use the Amazon Comprehend console to create and manage asynchronous analysis jobs. Your job analyzes documents stored in Amazon S3 to find entities such as events, phrases, primary language, sentiment, or personally identifiable information (PII).

To create an analysis job

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Analysis jobs** and then choose **Create job**.
3. Under **Job settings**, give the analysis job a unique name.
4. For **Analysis type**, choose one of the **Built-in** analysis types.

   If you choose **Primary langugage** or **Topic modeling**, you can skip the next step.
5. Depending on the **Analysis type** that you choose, the console displays one or more of the following additional fields:

   - **Language** is required for all built-in analysis types except **Primary langugage** and **Topic modeling**.

     Choose the language of your input documents.
   - **Target event types** is required for the **Events** analysis type.

     Select the types of events to detect in your input documents. For more information about supported event types, see Event types.
   - **PII detection settings** is required for the **PII** analysis type.

     Select the output mode. For more information about PII detection settings, see Detecting PII entities.
6. Under **Input data**, specify where the input documents are located in Amazon S3:
   - To analyze your own documents, choose **My documents**, and choose **Browse S3** to provide the path to the bucket or folder that contains your files.
   - To analyze samples that are provided by Amazon Comprehend, choose **Example documents**. In this case, Amazon Comprehend uses a bucket that is managed by AWS, and you don't specify the location.
7. (Optional) For **Input format**, specify one of the following formats for your input files:

   - **One document per file** – Each file contains one input document. This is best for collections of large documents.
   - **One document per line** – The input is one or more files. Each line in a file is considered a document. This is best for short documents, such as social media postings. Each line must

end with a line feed (LF, \n), a carriage return (CR, \r), or both (CRLF, \r\n). You can't use the UTF-8 line separator (u+2028) to end a line.

8. Under **Output data**, choose **Browse S3**. Choose the Amazon S3 bucket or folder where you want Amazon Comprehend to write the output data that is produced by the analysis.
9. (Optional) To encrypt the output result from your job, choose **Encryption**. Then, choose whether to use a KMS key associated with the current account or one from another account:
   - If you are using a key associated with the current account, choose the key alias or ID for **KMS key ID**.
   - If you are using a key associated with a different account, enter the ARN for the key alias or ID under **KMS key ID**.

   > **Note**
   >
   > For more information on creating and using KMS keys and the associated encryption, see Key management service (KMS).

10. Under **Access permissions**, provide an IAM role that:
    - Grants read access to the Amazon S3 location of your input documents.
    - Grants write access to the Amazon S3 location of your output documents.
    - Includes a trust policy that allows the `comprehend.amazonaws.com` service principal to assume the role and gain its permissions.

    If you don't already have an IAM role with these permissions and an appropriate trust policy, choose **Create an IAM** role to create one.
11. When you have finished filling out the form, choose **Create job** to create and start the topic detection job.

The new job appears in the job list with the status field showing the status of the job. The field can be `IN_PROGRESS` for a job that is processing, `COMPLETED` for a job that has finished successfully, and `FAILED` for a job that has an error. You can click on a job to get more information about the job, including any error messages.

When the job is completed, Amazon Comprehend stores the analysis results in the output Amazon S3 location that you specified for the job. For a description of the analysis results for each insight type, see Insights.

Using the Amazon Comprehend API

The Amazon Comprehend API supports operations to perform real-time (synchronous) analysis and operations to start and manage asynchronous analysis jobs.

You can use the Amazon Comprehend API operators directly, or you can use the CLI or one of the SDKs. The examples in this chapter use the CLI, the Python SDK, and Java SDK.

To run the AWS CLI and Python examples, you must install the AWS CLI. For more information, see Set up the AWS Command Line Interface (AWS CLI).

To run the Java examples, you must install the AWS SDK for Java. For instructions for installing the SDK for Java, see Set up the AWS SDK for Java.

Topics

Using Amazon Comprehend with an AWS SDK

AWS software development kits (SDKs) are available for many popular programming languages. Each SDK provides an API, code examples, and documentation that make it easier for developers to build applications in their preferred language.

SDK documentation	Code examples
AWS SDK for C++	AWS SDK for C++ code examples
AWS CLI	AWS CLI code examples
AWS SDK for Go	AWS SDK for Go code examples
AWS SDK for Java	AWS SDK for Java code examples
AWS SDK for JavaScript	AWS SDK for JavaScript code examples

SDK documentation	Code examples
AWS SDK for Kotlin	AWS SDK for Kotlin code examples
AWS SDK for .NET	AWS SDK for .NET code examples
AWS SDK for PHP	AWS SDK for PHP code examples
AWS Tools for PowerShell	Tools for PowerShell code examples
AWS SDK for Python (Boto3)	AWS SDK for Python (Boto3) code examples
AWS SDK for Ruby	AWS SDK for Ruby code examples
AWS SDK for Rust	AWS SDK for Rust code examples
AWS SDK for SAP ABAP	AWS SDK for SAP ABAP code examples
AWS SDK for Swift	AWS SDK for Swift code examples

Example availability

Can't find what you need? Request a code example by using the **Provide feedback** link at the bottom of this page.

Real-time analysis using the API

The following examples demonstrate how to use Amazon Comprehend API for real-time analysis, using the AWS CLI, and the AWS SDKs for .NET, Java, and Python. Use the examples to learn about the Amazon Comprehend synchronous operations and as building blocks for your own applications.

The .NET examples in this section use the AWS SDK for .NET. You can use the AWS Toolkit for Visual Studio to develop AWS applications using .NET. It includes helpful templates and the AWS Explorer for deploying applications and managing services. For a .NET developer perspective of AWS, see the AWS guide for .NET developers.

Topics

- Detecting the dominant language

- Detecting named entities
- Detecting key phrases
- Determining sentiment
- Real-time analysis for targeted sentiment
- Detecting syntax
- Real-time batch APIs

Detecting the dominant language

To determine the dominant language used in text, use the DetectDominantLanguage operation. To detect the dominant language in up to 25 documents in a batch, use the BatchDetectDominantLanguage operation. For more information, see Real-time batch APIs.

Topics

- Using the AWS Command Line Interface
- Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

Using the AWS Command Line Interface

The following example demonstrates using the `DetectDominantLanguage` operation with the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend detect-dominant-language \
    --region region \
    --text "It is raining today in Seattle."
```

Amazon Comprehend responds with the following:

```
{
    "Languages": [
        {
            "LanguageCode": "en",
            "Score": 0.9793661236763
        }
```

```
    ]
}
```

Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

For SDK examples of how to determine the dominant language, see Use `DetectDominantLanguage` with an AWS SDK or CLI.

Detecting named entities

To determine the named entities in a document, use the DetectEntities operation. To detect entities in up to 25 documents in a batch, use the BatchDetectEntities operation. For more information, see Real-time batch APIs.

Topics

- Using the AWS Command Line Interface
- Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

Using the AWS Command Line Interface

The following example demonstrates using the `DetectEntities` operation using the AWS CLI. You must specify the language of the input text.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend detect-entities \
    --region region \
    --language-code "en" \
    --text "It is raining today in Seattle."
```

Amazon Comprehend responds with the following:

```
{
    "Entities": [
        {
            "Text": "today",
            "Score": 0.97,
            "Type": "DATE",
```

```
            "BeginOffset": 14,
            "EndOffset": 19
        },
        {
            "Text": "Seattle",
            "Score": 0.95,
            "Type": "LOCATION",
            "BeginOffset": 23,
            "EndOffset": 30
        }
    ],
    "LanguageCode": "en"
}
```

Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

For SDK examples of how to determine the dominant language, see Use `DetectEntities` with an AWS SDK or CLI.

Detecting key phrases

To determine the key noun phrases used in text, use the DetectKeyPhrases operation. To detect the key noun phrases in up to 25 documents in a batch, use the BatchDetectKeyPhrases operation. For more information, see Real-time batch APIs.

Topics

- Using the AWS Command Line Interface
- Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

Using the AWS Command Line Interface

The following example demonstrates using the `DetectKeyPhrases` operation with the AWS CLI. You must specify the language of the input text.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend detect-key-phrases \
    --region region \
    --language-code "en" \
```

```
    --text "It is raining today in Seattle."
```

Amazon Comprehend responds with the following:

```
{
    "LanguageCode": "en",
    "KeyPhrases": [
        {
            "Text": "today",
            "Score": 0.89,
            "BeginOffset": 14,
            "EndOffset": 19
        },
        {
            "Text": "Seattle",
            "Score": 0.91,
            "BeginOffset": 23,
            "EndOffset": 30
        }
    ]
}
```

Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

For SDK examples that detect key phrases, see Use `DetectKeyPhrases` with an AWS SDK or CLI.

Determining sentiment

Amazon Comprehend provides the following API operations for analyzing sentiment:

- DetectSentiment – Determines the overall emotional sentiment of a document.
- BatchDetectSentiment – Determine the overall sentiment in up to 25 documents in a batch. For more information, see Real-time batch APIs
- StartSentimentDetectionJob – Starts an asynchronous sentiment detection job for a collection of documents.
- ListSentimentDetectionJobs – Returns the list of sentiment detection jobs that you have submitted.
- DescribeSentimentDetectionJob – Gets the properties (including status) associated with the specified sentiment detection job.
- StopSentimentDetectionJob – Stops the specified in-progress sentiment job.

Topics

Using the AWS Command Line Interface

The following example demonstrates using the `DetectSentiment` operation with the AWS CLI. This example specifies the language of the input text.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend detect-sentiment \
    --region region \
    --language-code "en" \
    --text "It is raining today in Seattle."
```

Amazon Comprehend responds with the following:

```
{
    "SentimentScore": {
        "Mixed": 0.014585512690246105,
        "Positive": 0.31592071056365967,
        "Neutral": 0.5985543131828308,
        "Negative": 0.07093945890665054
    },
    "Sentiment": "NEUTRAL",
    "LanguageCode": "en"
}
```

Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

For SDK examples that determine the sentiment of input text, see Use `DetectSentiment` with an AWS SDK or CLI.

Real-time analysis for targeted sentiment

Amazon Comprehend provides the following API operations for targeted sentiment real-time analysis:

- DetectTargetedSentiment – Analyzes sentiment of the entities mentioned in a document.
- BatchDetectTargetedSentiment – Analyzes targeted sentiment for up to 25 documents in a batch. For more information, see Real-time batch APIs

If the text you are analyzing doesn't include any targeted sentiment Entity types, the API returns an empty Entities array.

Using the AWS Command Line Interface

The following example demonstrates using the `DetectTargetedSentiment` operation with the AWS CLI. This example specifies the language of the input text.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend detect-targeted-sentiment \
    --region region \
    --language-code "en" \
    --text "The burger was cooked perfectly but it was cold. The service was OK."
```

Amazon Comprehend responds with the following:

```
{
"Entities": [
    {
      "DescriptiveMentionIndex": [
        0
      ],
      "Mentions": [
        {
          "BeginOffset": 4,
          "EndOffset": 10,
          "Score": 1,
          "GroupScore": 1,
          "Text": "burger",
          "Type": "OTHER",
          "MentionSentiment": {
            "Sentiment": "POSITIVE",
            "SentimentScore": {
              "Mixed": 0.001515,
              "Negative": 0.000822,
```

```
            "Neutral": 0.000243,
            "Positive": 0.99742
          }
        }
      },
      {
        "BeginOffset": 36,
        "EndOffset": 38,
        "Score": 0.999843,
        "GroupScore": 0.999661,
        "Text": "it",
        "Type": "OTHER",
        "MentionSentiment": {
          "Sentiment": "NEGATIVE",
          "SentimentScore": {
            "Mixed": 0,
            "Negative": 0.999996,
            "Neutral": 0.000004,
            "Positive": 0
          }
        }
      }
    ]
  },
  {
    "DescriptiveMentionIndex": [
      0
    ],
    "Mentions": [
      {
        "BeginOffset": 53,
        "EndOffset": 60,
        "Score": 1,
        "GroupScore": 1,
        "Text": "service",
        "Type": "ATTRIBUTE",
        "MentionSentiment": {
          "Sentiment": "NEUTRAL",
          "SentimentScore": {
            "Mixed": 0.000033,
            "Negative": 0.000089,
            "Neutral": 0.993325,
            "Positive": 0.006553
          }
```

```
          }
        }
      ]
    }
  ]
}
```

Detecting syntax

To parse text to extract the individual words and determine the parts of speech for each word, use the DetectSyntax operation. To parse the syntax of up to 25 documents in a batch, use the BatchDetectSyntax operation. For more information, see Real-time batch APIs.

Topics

- Using the AWS Command Line Interface.
- Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

Using the AWS Command Line Interface.

The following example demonstrates using the `DetectSyntax` operation with the AWS CLI. This example specifies the language of the input text.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend detect-syntax \
   --region region \
   --language-code "en" \
   --text "It is raining today in Seattle."
```

Amazon Comprehend responds with the following:

```
{
    "SyntaxTokens": [
        {
            "Text": "It",
            "EndOffset": 2,
            "BeginOffset": 0,
            "PartOfSpeech": {
                "Tag": "PRON",
```

```
                "Score": 0.8389829397201538
            },
            "TokenId": 1
        },
        {
            "Text": "is",
            "EndOffset": 5,
            "BeginOffset": 3,
            "PartOfSpeech": {
                "Tag": "AUX",
                "Score": 0.9189288020133972
            },
            "TokenId": 2
        },
        {
            "Text": "raining",
            "EndOffset": 13,
            "BeginOffset": 6,
            "PartOfSpeech": {
                "Tag": "VERB",
                "Score": 0.9977611303329468
            },
            "TokenId": 3
        },
        {
            "Text": "today",
            "EndOffset": 19,
            "BeginOffset": 14,
            "PartOfSpeech": {
                "Tag": "NOUN",
                "Score": 0.9993606209754944
            },
            "TokenId": 4
        },
        {
            "Text": "in",
            "EndOffset": 22,
            "BeginOffset": 20,
            "PartOfSpeech": {
                "Tag": "ADP",
                "Score": 0.9999061822891235
            },
            "TokenId": 5
        },
```

```
        {
            "Text": "Seattle",
            "EndOffset": 30,
            "BeginOffset": 23,
            "PartOfSpeech": {
                "Tag": "PROPN",
                "Score": 0.9940338730812073
            },
            "TokenId": 6
        },
        {
            "Text": ".",
            "EndOffset": 31,
            "BeginOffset": 30,
            "PartOfSpeech": {
                "Tag": "PUNCT",
                "Score": 0.9999997615814209
            },
            "TokenId": 7
        }
    ]
}
```

Using the AWS SDK for Java, SDK for Python, or AWS SDK for .NET

For SDK examples that detect the syntax of input text, see Use `DetectSyntax` with an AWS SDK or CLI.

Real-time batch APIs

To send batches of up to 25 documents, you can use the Amazon Comprehend real-time batch operations. Calling a batch operation is identical to calling the single document APIs for each document in the request. Using the batch APIs can result in better performance for your applications. For more information, see Multiple document synchronous processing.

Topics

- Batch processing with the AWS CLI
- Batch processing with the AWS SDK for .NET

Batch processing with the AWS CLI

These examples show how to use the batch API operations using the AWS Command Line Interface. All of the operations except `BatchDetectDominantLanguage` use the following JSON file called `process.json` as input. For that operation the `LanguageCode` entity is not included.

The third document in the JSON file ("$$$$$$$$") will cause an error during batch processing. It is included so that the operations will include an BatchItemError in the response.

```
{
   "LanguageCode": "en",
   "TextList": [
      "I have been living in Seattle for almost 4 years",
      "It is raining today in Seattle",
      "$$$$$$$$"
   ]
}
```

The examples are formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Topics

- Detect the dominant language using a batch (AWS CLI)
- Detect entities using a batch (AWS CLI)
- Detect key phrases using a batch (AWS CLI)
- Detect sentiment using a batch (AWS CLI)

Detect the dominant language using a batch (AWS CLI)

The BatchDetectDominantLanguage operation determines the dominant language of each document in a batch. For a list of the languages that Amazon Comprehend can detect, see Dominant language. The following AWS CLI command calls the `BatchDetectDominantLanguage` operation.

```
aws comprehend batch-detect-dominant-language \
    --endpoint endpoint \
    --region region \
    --cli-input-json file://path to input file/process.json
```

The following is the response from the `BatchDetectDominantLanguage` operation:

```
{
    "ResultList": [
        {
          "Index": 0,
          "Languages":[
            {
              "LanguageCode":"en",
              "Score": 0.99
            }
          ]
        },
        {
          "Index": 1
          "Languages":[
            {
              "LanguageCode":"en",
              "Score": 0.82
            }
          ]
        }
    ],
    "ErrorList": [
      {
        "Index": 2,
        "ErrorCode": "InternalServerException",
        "ErrorMessage": "Unexpected Server Error. Please try again."
      }
    ]
}
```

Detect entities using a batch (AWS CLI)

Use the BatchDetectEntities operation to find the entities present in a batch of documents. For more information about entities, see Entities. The following AWS CLI command calls the `BatchDetectEntities` operation.

```
aws comprehend batch-detect-entities \
    --endpoint endpoint \
    --region region \
    --cli-input-json file://path to input file/process.json
```

Detect key phrases using a batch (AWS CLI)

The BatchDetectKeyPhrases operation returns the key noun phrases in a batch of documents. The following AWS CLI command calls the `BatchDetectKeyNounPhrases` operation.

```
aws comprehend batch-detect-key-phrases
    --endpoint endpoint
    --region region
    --cli-input-json file://path to input file/process.json
```

Detect sentiment using a batch (AWS CLI)

Detect the overall sentiment of a batch of documents using the BatchDetectSentiment operation. The following AWS CLI command calls the `BatchDetectSentiment` operation.

```
aws comprehend batch-detect-sentiment \
    --endpoint endpoint \
    --region region \
    --cli-input-json file://path to input file/process.json
```

Batch processing with the AWS SDK for .NET

The following sample program shows how to use the BatchDetectEntities operation with the AWS SDK for .NET. The response from the server contains a BatchDetectEntitiesItemResult object for each document that was successfully processed. If there is an error processing a document, there will be a record in the error list in the response. The example gets each of the documents with an error and resends them.

The .NET example in this section uses the AWS SDK for .NET. You can use the AWS Toolkit for Visual Studio to develop AWS applications using .NET. It includes helpful templates and the AWS Explorer for deploying applications and managing services. For a .NET developer perspective of AWS, see the AWS guide for .NET developers.

```
using System;
using System.Collections.Generic;
using Amazon.Comprehend;
using Amazon.Comprehend.Model;

namespace Comprehend
{
    class Program
```

```
   {
        // Helper method for printing properties
        static private void PrintEntity(Entity entity)
        {
            Console.WriteLine("     Text: {0}, Type: {1}, Score: {2}, BeginOffset: {3}
EndOffset: {4}",
                entity.Text, entity.Type, entity.Score, entity.BeginOffset,
entity.EndOffset);
        }

        static void Main(string[] args)
        {
            AmazonComprehendClient comprehendClient = new
AmazonComprehendClient(Amazon.RegionEndpoint.USWest2);

            List<String> textList = new List<String>()
            {
                { "I love Seattle" },
                { "Today is Sunday" },
                { "Tomorrow is Monday" },
                { "I love Seattle" }
            };

            // Call detectEntities API
            Console.WriteLine("Calling BatchDetectEntities");
            BatchDetectEntitiesRequest batchDetectEntitiesRequest = new
BatchDetectEntitiesRequest()
            {
                TextList = textList,
                LanguageCode = "en"
            };
            BatchDetectEntitiesResponse batchDetectEntitiesResponse =
comprehendClient.BatchDetectEntities(batchDetectEntitiesRequest);

            foreach (BatchDetectEntitiesItemResult item in
batchDetectEntitiesResponse.ResultList)
            {
                Console.WriteLine("Entities in {0}:", textList[item.Index]);
                foreach (Entity entity in item.Entities)
                    PrintEntity(entity);
            }

            // check if we need to retry failed requests
            if (batchDetectEntitiesResponse.ErrorList.Count != 0)
```

```
            {
                Console.WriteLine("Retrying Failed Requests");
                List<String> textToRetry = new List<String>();
                foreach(BatchItemError errorItem in
 batchDetectEntitiesResponse.ErrorList)
                    textToRetry.Add(textList[errorItem.Index]);

                batchDetectEntitiesRequest = new BatchDetectEntitiesRequest()
                {
                    TextList = textToRetry,
                    LanguageCode = "en"
                };

                batchDetectEntitiesResponse =
 comprehendClient.BatchDetectEntities(batchDetectEntitiesRequest);

                foreach(BatchDetectEntitiesItemResult item in
 batchDetectEntitiesResponse.ResultList)
                {
                    Console.WriteLine("Entities in {0}:", textList[item.Index]);
                    foreach (Entity entity in item.Entities)
                        PrintEntity(entity);
                }
            }
            Console.WriteLine("End of DetectEntities");
        }
    }
}
```

Async analysis jobs using the API

The following examples use the Amazon Comprehend asynchronous APIs to create and manage analysis jobs, using the AWS CLI.

Topics

- Async analysis for Amazon Comprehend insights
- Async analysis for targeted sentiment
- Async analysis for event detection
- Async analysis for topic modeling

Async analysis for Amazon Comprehend insights

The following sections use the Amazon Comprehend API to run asynchronous operations to analyze Amazon Comprehend insights.

Topics

Prerequisites

Documents must be in UTF-8-formatted text files. You can submit your documents in two formats. The format you use depends on the type of documents you want to analyze, as described in the following table.

Description	Format
Each file contains one input document. This is best for collections of large documents.	One document per file
The input is one or more files. Each line in a file is considered a document. This is best for short documents, such as social media postings. Each line must end with a line feed (LF, \n), a carriage return (CR, \r), or both (CRLF, \r \n). You can't use the UTF-8 line separator (u +2028) to end a line.	One document per line

When you start an analysis job, you specify the S3 location for your input data. The URI must be in the same AWS Region as the API endpoint that you are calling. The URI can point to a single file or it can be the prefix for a collection of data files. For more information, see the InputDataConfig data type.

You must grant Amazon Comprehend access to the Amazon S3 bucket that contains your document collection and output files. For more information, see Role-based permissions required for asynchronous operations.

Starting an analysis job

To submit an analysis job, use either the Amazon Comprehend console or the appropriate `Start*` operation:

- StartDominantLanguageDetectionJob — Start a job to detect the dominant language in each document in the collection. For more information about the dominant language in a document, see Dominant language.
- StartEntitiesDetectionJob — Start a job to detect entities in each document in the collection. For more information about entities, see Entities.
- StartKeyPhrasesDetectionJob — Start a job to detect key phrases in each document in the collection. For more information about key phrases, see Key phrases.
- StartPiiEntitiesDetectionJob — Start a job to detect personally identifiable information (PII) in each document in the collection. For more information about PII, see Detecting PII entities.
- StartSentimentDetectionJob — Start a job to detect the sentiment in each document in the collection. For more information about sentiments, see Sentiment.

Monitoring analysis jobs

The `Start*` operation returns an ID that you can use to monitor the job's progress.

To monitor progress using the API, you use one of two operations, depending on whether you want to monitor the progress of an individual job or multiple jobs.

To monitor the progress of an individual analysis job, use the `Describe*` operations. You provide the job ID returned by the `Start*` operation. The response from the `Describe*` operation contains the `JobStatus` field with the job's status.

To monitor the progress of multiple analysis jobs, use the `List*` operations. `List*` operations return a list of jobs that you submitted to Amazon Comprehend. The response includes a `JobStatus` field for each job that tells you the status of the job.

If the status field is set to `COMPLETED` or `FAILED`, job processing has completed.

To get the status of individual jobs, use the `Describe*` operation for the analysis that you are performing.

- DescribeDominantLanguageDetectionJob
- DescribeEntitiesDetectionJob
- DescribeKeyPhrasesDetectionJob
- DescribePiiEntitiesDetectionJob
- DescribeSentimentDetectionJob

To get the status of a multiple jobs, use the `List*` operation for the analysis that you are performing.

- ListDominantLanguageDetectionJobs
- ListEntitiesDetectionJobs
- ListKeyPhrasesDetectionJobs
- ListPiiEntitiesDetectionJobs
- ListSentimentDetectionJobs

To restrict the results to jobs that match certain criteria, use the `List*` operations' `Filter` parameter. You can filter on the job name, the job status, and the date and time that the job was submitted. For more information, see the `Filter` parameter for each of the `List*` operations in the Amazon Comprehend API reference.

Getting analysis results

After an analysis job has finished, use a `Describe*` operation to get the location of the results. If the job status is `COMPLETED`, the response includes an `OutputDataConfig` field that contains a field with the Amazon S3 location of the output file. The file, `output.tar.gz`, is a compressed archive that contains the results of the analysis.

If the status of a job is `FAILED`, the response contains a `Message` field that describes the reason that the analysis job didn't complete successfully.

To get the status of individual jobs, use the appropriate `Describe*` operation:

- DescribeDominantLanguageDetectionJob

- DescribeEntitiesDetectionJob
- DescribeKeyPhrasesDetectionJob
- DescribeSentimentDetectionJob

The results are returned in a single file, with one JSON structure for each document. Each response file also includes error messages for any job with the status field set to `FAILED`.

Each of the following sections shows examples of output for the two input formats.

Getting dominant language detection results

The following is an example of an output file from an analysis that detected the dominant language. The format of the input is one document per line. For more information, see the DetectDominantLanguage operation.

```
{"File": "0_doc", "Languages": [{"LanguageCode": "en", "Score": 0.9514502286911011},
 {"LanguageCode": "de", "Score": 0.02374090999364853}, {"LanguageCode": "nl", "Score":
 0.003208699868991971}, "Line": 0}
{"File": "1_doc", "Languages": [{"LanguageCode": "en", "Score": 0.9822712540626526},
 {"LanguageCode": "de", "Score": 0.00262139202095568 2}, {"LanguageCode": "es", "Score":
 0.00238655414432 2872}], "Line": 1}
```

The following is an example of output from an analysis where the format of the input is one document per file:

```
{"File": "small_doc", "Languages": [{"LanguageCode": "en", "Score":
 0.9728053212165833}, {"LanguageCode": "de", "Score": 0.007670710328966379},
 {"LanguageCode": "es", "Score": 0.002847236813977 3607}]}
{"File": "huge_doc", "Languages": [{"LanguageCode": "en", "Score": 0.984955906867981},
 {"LanguageCode": "de", "Score": 0.0026436643674969673}, {"LanguageCode": "fr",
 "Score": 0.0014206881169229746}]}
```

Getting entity detection results

The following is an example of an output file from an analysis that detected entities in documents. The format of the input is one document per line. For more information, see the DetectEntities operation. The output contains two error messages, one for a document that is too long and one for a document that isn't in UTF-8 format.

```
{"File": "50_docs", "Line": 0, "Entities": [{"BeginOffset": 0, "EndOffset": 22,
 "Score": 0.9763959646224976, "Text": "Cluj-NapocaCluj-Napoca", "Type": "LOCATION"}"]}
{"File": "50_docs", "Line": 1, "Entities": [{"BeginOffset": 11, "EndOffset": 15,
 "Score": 0.9615424871444702, "Text": "Maat", "Type": "PERSON"}}]}
{"File": "50_docs", "Line": 2, "ErrorCode": "DOCUMENT_SIZE_EXCEEDED", "ErrorMessage":
 "Document size exceeds maximum size limit 102400 bytes."}
{"File": "50_docs", "Line": 3, "ErrorCode": "UNSUPPORTED_ENCODING", "ErrorMessage":
 "Document is not in UTF-8 format and all subsequent lines are ignored."}
```

The following is an example of output from an analysis where the format of the input is one document per file. The output contains two error messages, one for a document that is too long and one for a document that isn't in UTF-8 format.

```
{"File": "non_utf8.txt", "ErrorCode": "UNSUPPORTED_ENCODING", "ErrorMessage": "Document
 is not in UTF-8 format and all subsequent line are ignored."}
{"File": "small_doc", "Entities": [{"BeginOffset": 0, "EndOffset": 4, "Score":
 0.645766019821167, "Text": "Maat", "Type": "PERSON"}]}
{"File": "huge_doc", "ErrorCode": "DOCUMENT_SIZE_EXCEEDED", "ErrorMessage": "Document
 size exceeds size limit 102400 bytes."}
```

Getting key phrase detection results

The following is an example of an output file from an analysis that detected key phrases in a document. The format of the input is one document per line. For more information, see the DetectKeyPhrases operation.

```
{"File": "50_docs", "KeyPhrases": [{"BeginOffset": 0, "EndOffset": 22, "Score":
 0.8948641419410706, "Text": "Cluj-NapocaCluj-Napoca"}, {"BeginOffset": 45,
 "EndOffset": 49, "Score": 0.9989854693412781, "Text": "Cluj"}], "Line": 0}
```

The following is an example of the output from an analysis where the format of the input is one document per file.

```
{"File": "1_doc", "KeyPhrases": [{"BeginOffset": 0, "EndOffset": 22, "Score":
 0.8948641419410706, "Text": "Cluj-NapocaCluj-Napoca"}, {"BeginOffset": 45,
 "EndOffset": 49, "Score": 0.9989854693412781, "Text": "Cluj"}]}
```

Getting personally identifiable information (PII) detection results

The following is an example an output file from an analysis job that detected PII entities in documents. The format of the input is one document per line.

```
{"Entities":[{"Type":"NAME","BeginOffset":40,"EndOffset":69,"Score":0.999995},
{"Type":"ADDRESS","BeginOffset":247,"EndOffset":253,"Score":0.998828},
{"Type":"BANK_ACCOUNT_NUMBER","BeginOffset":406,"EndOffset":411,"Score":0.693283}],"File":"doc.
{"Entities":[{"Type":"SSN","BeginOffset":1114,"EndOffset":1124,"Score":0.999999},
{"Type":"EMAIL","BeginOffset":3742,"EndOffset":3775,"Score":0.999993},
{"Type":"PIN","BeginOffset":4098,"EndOffset":4102,"Score":0.999995}],"File":"doc.txt","Line":1}
```


The following is an example of output from an analysis where the format of the input is one document per file.


```
{"Entities":[{"Type":"NAME","BeginOffset":40,"EndOffset":69,"Score":0.999995},
{"Type":"ADDRESS","BeginOffset":247,"EndOffset":253,"Score":0.998828},
{"Type":"BANK_ROUTING","BeginOffset":279,"EndOffset":289,"Score":0.999999}],"File":"doc.txt"}
```


Getting sentiment detection results

The following is an example of an output file from an analysis that detected the sentiment expressed in a document. It includes an error message because one document is too long. The format of the input is one document per line. For more information, see the DetectSentiment operation.


```
{"File": "50_docs", "Line": 0, "Sentiment": "NEUTRAL", "SentimentScore": {"Mixed":
 0.002734508365392685, "Negative": 0.008935936726629734, "Neutral": 0.9841893315315247,
 "Positive": 0.004140198230743408}}
{"File": "50_docs", "Line": 1, "ErrorCode": "DOCUMENT_SIZE_EXCEEDED", "ErrorMessage":
 "Document size is exceeded maximum size limit 5120 bytes."}
{"File": "50_docs", "Line": 2, "Sentiment": "NEUTRAL", "SentimentScore":
 {"Mixed": 0.0023119584657251835, "Negative": 0.0029857370536774397, "Neutral":
 0.9866572022438049, "Positive": 0.008045154623687267}}
```


The following is an example of the output from an analysis where the format of the input is one document per file.


```
{"File": "small_doc", "Sentiment": "NEUTRAL", "SentimentScore": {"Mixed":
 0.0023450672160834074, "Negative": 0.0009663937962614, "Neutral": 0.9795311689376831,
 "Positive": 0.01715377675175667}}
{"File": "huge_doc", "ErrorCode": "DOCUMENT_SIZE_EXCEEDED", "ErrorMessage": "Document
 size is exceeds the limit of 5120 bytes."}
```

Async analysis for targeted sentiment

For information about real-time analysis for Targeted sentiment, see the section called "Real-time analysis for targeted sentiment".

Amazon Comprehend provides the following API operations to start and manage asynchronous targeted sentiment analysis:

- StartTargetedSentimentDetectionJob – Starts an asynchronous targeted sentiment detection job for a collection of documents.
- ListTargetedSentimentDetectionJobs – Returns the list of targeted sentiment detection jobs that you have submitted.
- DescribeTargetedSentimentDetectionJob – Gets the properties (including status) associated with the specified targeted sentiment detection job.
- StopTargetedSentimentDetectionJob – Stops the specified in-progress targeted sentiment job.

Topics

Before you start

Before you start, make sure that you have:

- **Input and output buckets**—Identify the Amazon S3 buckets that you want to use for input and output. The buckets must be in the same Region as the API that you are calling.
- **IAM service role**—You must have an IAM service role with permission to access your input and output buckets. For more information, see Role-based permissions required for asynchronous operations.

Analyzing targeted sentiment using the AWS CLI

The following example demonstrates using the `StartTargetedSentimentDetectionJob` operation with the AWS CLI. This example specifies the language of the input text.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend start-targeted-sentiment-detection-job \
       --job-name "job name" \
       --language-code "en" \
       --cli-input-json file://path to JSON input file
```

For the `cli-input-json` parameter you supply the path to a JSON file that contains the request data, as shown in the following example.

```
{
    "InputDataConfig": {
        "S3Uri": "s3://input bucket/input path",
        "InputFormat": "ONE_DOC_PER_FILE"
    },
    "OutputDataConfig": {
        "S3Uri": "s3://output bucket/output path"
    },
    "DataAccessRoleArn": "arn:aws:iam::account ID:role/data access role"
}
```

If the request to start the job was successful, you will receive the following response:

```
{
    "JobStatus": "SUBMITTED",
    "JobArn": "job ARN"
    "JobId": "job ID"
}
```

Async analysis for event detection

Topics

- Before you start
- Detect events using the AWS CLI
- List events using the AWS CLI
- Describe events using the AWS CLI
- Get events detection results

To detect events in a document set, use the StartEventsDetectionJob to start an asynchronous job.

Before you start

Before you start, make sure that you have:

- **Input and output buckets**—Identify the Amazon S3 buckets that you want to use for input and output. The buckets must be in the same Region as the API that you are calling.
- **IAM service role**—You must have an IAM service role with permission to access your input and output buckets. For more information, see Role-based permissions required for asynchronous operations.

Detect events using the AWS CLI

The following example demonstrates using the StartEventsDetectionJob operation with the AWS CLI

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend start-events-detection-job \
  --region region \
  --job-name job name \
  --cli-input-json file://path to JSON input file
```

For the `cli-input-json` parameter you supply the path to a JSON file that contains the request data, as shown in the following example.

```
{
  "InputDataConfig": {
      "S3Uri": "s3://input bucket/input path",
      "InputFormat": "ONE_DOC_PER_LINE"
  },
  "OutputDataConfig": {
      "S3Uri": "s3://output bucket/output path"
  },
  "DataAccessRoleArn": "arn:aws:iam::account ID:role/data access role"
  "LanguageCode": "en",
  "TargetEventTypes": [
      "BANKRUPTCY",
      "EMPLOYMENT",
      "CORPORATE_ACQUISITION",
```

```
      "INVESTMENT_GENERAL",
      "CORPORATE_MERGER",
      "IPO",
      "RIGHTS_ISSUE",
      "SECONDARY_OFFERING",
      "SHELF_OFFERING",
      "TENDER_OFFERING",
      "STOCK_SPLIT"
  ]
}
```

If the request to start the events detection job was successful, you will receive the following response:

```
{
  "JobStatus": "SUBMITTED",
  "JobId": "job ID"
}
```

List events using the AWS CLI

Use the ListEventsDetectionJobs operation to see a list of the events detection jobs that you have submitted. The list includes information about the input and output locations that you used and the status of each of the detection jobs. The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend list-events-detection-jobs --region region
```

You will get JSON similar to the following in response:

```
{
 "EventsDetectionJobPropertiesList": [
    {
       "DataAccessRoleArn": "arn:aws:iam::account ID:role/data access role",
       "EndTime": timestamp,
       "InputDataConfig": {
          "InputFormat": "ONE_DOC_PER_LINE",
          "S3Uri": "s3://input bucket/input path"
       },
       "JobId": "job ID",
```

```
        "JobName": "job name",
        "JobStatus": "COMPLETED",
        "LanguageCode": "en",
        "Message": "message",
        "OutputDataConfig": {
           "S3Uri": "s3://output bucket/ouput path"
        },
        "SubmitTime": timestamp,
        "TargetEventTypes": [
          "BANKRUPTCY",
          "EMPLOYMENT",
          "CORPORATE_ACQUISITION",
          "INVESTMENT_GENERAL",
          "CORPORATE_MERGER",
          "IPO",
          "RIGHTS_ISSUE",
          "SECONDARY_OFFERING",
          "SHELF_OFFERING",
          "TENDER_OFFERING",
          "STOCK_SPLIT"
   ]
     }
 ],
 "NextToken": "next token"
}
```

Describe events using the AWS CLI

You can use the DescribeEventsDetectionJob operation to get the status of an existing job. The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend describe-events-detection-job \
  --region region \
  --job-id job ID
```

You will get the following JSON in response:

```
{
 "EventsDetectionJobProperties": {
    "DataAccessRoleArn": "arn:aws:iam::account ID:role/data access role",
```

```
    "EndTime": timestamp,
    "InputDataConfig": {
       "InputFormat": "ONE_DOC_PER_LINE",
       "S3Uri": "S3Uri": "s3://input bucket/input path"
    },
    "JobId": "job ID",
    "JobName": "job name",
    "JobStatus": "job status",
    "LanguageCode": "en",
    "Message": "message",
    "OutputDataConfig": {
       "S3Uri": "s3://output bucket/output path"
    },
    "SubmitTime": timestamp,
    "TargetEventTypes": [
      "BANKRUPTCY",
      "EMPLOYMENT",
      "CORPORATE_ACQUISITION",
      "INVESTMENT_GENERAL",
      "CORPORATE_MERGER",
      "IPO",
      "RIGHTS_ISSUE",
      "SECONDARY_OFFERING",
      "SHELF_OFFERING",
      "TENDER_OFFERING",
      "STOCK_SPLIT"
  ]
 }
}
```

Get events detection results

The following is an example an output file from an analysis job that detected events in documents. The format of the input is one document per line.

```
{"Entities": [{"Mentions": [{"BeginOffset": 12, "EndOffset": 27, "GroupScore":
 1.0, "Score": 0.916355, "Text": "over a year ago", "Type": "DATE"}]}, {"Mentions":
 [{"BeginOffset": 33, "EndOffset": 39, "GroupScore": 1.0, "Score": 0.996603,
 "Text": "Amazon", "Type": "ORGANIZATION"}]}, {"Mentions": [{"BeginOffset":
 66, "EndOffset": 77, "GroupScore": 1.0, "Score": 0.999283, "Text": "Whole
 Foods", "Type": "ORGANIZATION"}]}], "Events": [{"Arguments": [{"EntityIndex":
 2, "Role": "INVESTEE", "Score": 0.999283}, {"EntityIndex": 0, "Role": "DATE",
 "Score": 0.916355}, {"EntityIndex": 1, "Role": "INVESTOR", "Score": 0.996603}],
```

```
"Triggers": [{"BeginOffset": 373, "EndOffset": 380, "GroupScore": 0.999984,
"Score": 0.999955, "Text": "acquire", "Type": "CORPORATE_ACQUISITION"}], "Type":
"CORPORATE_ACQUISITION"}, {"Arguments": [{"EntityIndex": 2, "Role": "PARTICIPANT",
"Score": 0.999283}], "Triggers": [{"BeginOffset": 115, "EndOffset": 123, "GroupScore":
1.0, "Score": 0.999967, "Text": "combined", "Type": "CORPORATE_MERGER"}], "Type":
"CORPORATE_MERGER"}], "File": "doc.txt", "Line": 0}
```


For more information about events output file structure and supported event types, see Events.

Async analysis for topic modeling

To determine the topics in a document set, use the StartTopicsDetectionJob to start an asynchronous job. You can monitor topics in documents written in English or Spanish.

Topics

- Before you start
- Using the AWS Command Line Interface
- Using the SDK for Python or AWS SDK for .NET

Before you start

Before you start, make sure that you have:

- **Input and output buckets**—Identify the Amazon S3 buckets that you want to use for input and output. The buckets must be in the same Region as the API that you are calling.
- **IAM service role**—You must have an IAM service role with permission to access your input and output buckets. For more information, see Role-based permissions required for asynchronous operations.

Using the AWS Command Line Interface

The following example demonstrates using the `StartTopicsDetectionJob` operation with the AWS CLI

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend start-topics-detection-job \
```

```
                --number-of-topics topics to return \
                --job-name "job name" \
                --region region \
                --cli-input-json file://path to JSON input file
```

For the `cli-input-json` parameter you supply the path to a JSON file that contains the request data, as shown in the following example.

```
{
    "InputDataConfig": {
        "S3Uri": "s3://input bucket/input path",
        "InputFormat": "ONE_DOC_PER_FILE"
    },
    "OutputDataConfig": {
        "S3Uri": "s3://output bucket/output path"
    },
    "DataAccessRoleArn": "arn:aws:iam::account ID:role/data access role"
}
```

If the request to start the topic detection job was successful, you will receive the following response:

```
{
    "JobStatus": "SUBMITTED",
    "JobId": "job ID"
}
```

Use the ListTopicsDetectionJobs operation to see a list of the topic detection jobs that you have submitted. The list includes information about the input and output locations that you used and the status of each of the detection jobs. The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend list-topics-detection-jobs \-- region
```

You will get JSON similar to the following in response:

```
{
    "TopicsDetectionJobPropertiesList": [
        {
```

```
            "InputDataConfig": {
                "S3Uri": "s3://input bucket/input path",
                "InputFormat": "ONE_DOC_PER_LINE"
            },
            "NumberOfTopics": topics to return,
            "JobId": "job ID",
            "JobStatus": "COMPLETED",
            "JobName": "job name",
            "SubmitTime": timestamp,
            "OutputDataConfig": {
                "S3Uri": "s3://output bucket/output path"
            },
            "EndTime": timestamp
        },
        {
            "InputDataConfig": {
                "S3Uri": "s3://input bucket/input path",
                "InputFormat": "ONE_DOC_PER_LINE"
            },
            "NumberOfTopics": topics to return,
            "JobId": "job ID",
            "JobStatus": "RUNNING",
            "JobName": "job name",
            "SubmitTime": timestamp,
            "OutputDataConfig": {
                "S3Uri": "s3://output bucket/output path"
            }
        }
    ]
}
```

You can use the DescribeTopicsDetectionJob operation to get the status of an existing job. The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend describe-topics-detection-job --job-id job ID
```

You will get the following JSON in response:

```
{
    "TopicsDetectionJobProperties": {
        "InputDataConfig": {
```

```
            "S3Uri": "s3://input bucket/input path",
            "InputFormat": "ONE_DOC_PER_LINE"
        },
        "NumberOfTopics": topics to return,
        "JobId": "job ID",
        "JobStatus": "COMPLETED",
        "JobName": "job name",
        "SubmitTime": timestamp,
        "OutputDataConfig": {
            "S3Uri": "s3://output bucket/ouput path"
        },
        "EndTime": timestamp
    }
}
```

Using the SDK for Python or AWS SDK for .NET

For SDK examples of how to start a topic modeling job, see Use `StartTopicsDetectionJob` with an AWS SDK or CLI.

Trust and safety

Users generate large amounts of text content through online applications (such as peer-to-peer chats and forum discussions), comments posted on websites, and through generative AI applications (input prompts and outputs from generative AI models). The Amazon Comprehend Trust and Safety features can help you moderate this content, to provide a safe and inclusive environment for your users.

Benefits of using the Amazon Comprehend trust and safety features include:

- Faster moderation: Quickly and accurately moderate large volume of text to keep your online platforms free from inappropriate content.
- Customizable: Customize the moderation thresholds in API responses to suit your application needs.
- Easy to use: Configure the trust and safety features through LangChain integration or using the AWS CLI or SDKs.

Amazon Comprehend trust and safety address the following aspects of content moderation:

- **Toxicity detection** – Detect content that may be harmful, offensive, or inappropriate. Examples include hate speech, threats, or abuse.
- **Intent classification** – Detect content that has explicit or implicit malicious intent. Examples include discriminatory or illegal content, or content that expresses or requests advice on medical, legal, political, controversial, personal or financial subjects.
- **Privacy protection** – Users can inadvertently provide content that may reveal personally identifiable information (PII). Amazon Comprehend PII provides the ability to detect and redact PII.

Topics

Toxicity detection

Amazon Comprehend toxicity detection provides real-time detection of toxic content in text-based interactions. You can use toxicity detection to moderate peer-to-peer conversations in online platforms or to monitor generative AI inputs and outputs.

Toxicity detection detects the following categories of offensive content:

GRAPHIC

Graphic speech uses visually descriptive, detailed, and unpleasantly vivid imagery. Such language is often made verbose to amplify an insult, discomfort or harm to the recipient.

HARASSMENT_OR_ABUSE

Speech that imposes disruptive power dynamics between the speaker and hearer, regardless of intent, seeks to affect the psychological well-being of the recipient, or objectifies a person.

HATE_SPEECH

Speech that criticizes, insults, denounces or dehumanizes a person or a group on the basis of an identity, be it race, ethnicity, gender identity, religion, sexual orientation, ability, national origin, or another identity-group.

INSULT

Speech that includes demeaning, humiliating, mocking, insulting, or belittling language.

PROFANITY

Speech that contains words, phrases or acronyms that are impolite, vulgar, or offensive is considered as profane.

SEXUAL

Speech that indicates sexual interest, activity or arousal by using direct or indirect references to body parts or physical traits or sex .

VIOLENCE_OR_THREAT

Speech that includes threats which seek to inflict pain, injury or hostility towards a person or group.

TOXICITY

Speech that contains words, phrases or acronyms that might be considered toxic in nature across any of the above categories.

Detecting toxic content using the API

To detect toxic content in text, use the synchronous DetectToxicContent operation. This operation performs analysis on a list of text strings that you provide as input. The API response contains a results list that matches the size of the input list.

Currently, toxic content detection supports only the English language. For input text, you can provide a list of up to 10 text strings. Each string has a maximum size of 1KB.

The toxic content detection returns a list of analysis results, one entry in the list for each input string. An entry contains a list of toxic content types identified in the text string, along with a confidence score for each content type. The entry also includes a toxicity score for the string.

The following examples show how to use the `DetectToxicContent` operation using the AWS CLI and Python.

AWS CLI

You can detect toxic content using the following command in the AWS CLI:

```
aws comprehend detect-toxic-content --language-code en  /
            --text-segments "[{\"Text\":\"You are so obtuse\"}]"
```

The AWS CLI responds with the following result. The text segment receives a high confidence score in the `INSULT` category, with a resulting high toxicity score:

```
{
   "ResultList": [
      {
         "Labels": [
                {
                    "Name": "PROFANITY",
                    "Score": 0.0006000000284984708
                },
                {
                    "Name": "HATE_SPEECH",
                    "Score": 0.00930000003427267
                },
                {
                    "Name": "INSULT",
                    "Score": 0.9204999804496765
```

```
                },
                {
                    "Name": "GRAPHIC",
                    "Score": 9.999999747378752e-05
                },
                {
                    "Name": "HARASSMENT_OR_ABUSE",
                    "Score": 0.0052999998442828655
                },
                {
                    "Name": "SEXUAL",
                    "Score": 0.01549999974668026
                },
                {
                    "Name": "VIOLENCE_OR_THREAT",
                    "Score": 0.007799999788403511
                }
            ],
            "Toxicity": 0.7192999720573425
      }
   ]
}
```

You can input up to 10 text strings, using the following format for the `text-segments` parameter:

```
   --text-segments "[{\"Text\":\"text string 1\"},
                    {\"Text\":\"text string2\"},
                    {\"Text\":\"text string3\"}]"
```

The AWS CLI responds with the following results:

```
{
   "ResultList": [
      {
         "Labels": [ (truncated) ],
            "Toxicity": 0.3192999720573425
      },
      {
         "Labels": [ (truncated) ],
            "Toxicity": 0.1192999720573425
```

```
      },
      {
         "Labels": [ (truncated) ],
            "Toxicity": 0.0192999720573425
      }
   ]
}
```

Python (Boto)

The following example demonstrates how to detect toxic content using Python:

```
import boto3
client = boto3.client(
    service_name='comprehend',
    region_name=region) # For example, 'us-west-2'

response = client.detect_toxic_content(
    LanguageCode='en',
    TextSegments=[{'Text': 'You are so obtuse'}]
)
print("Response: %s\n" % response)
```

Prompt safety classification

Amazon Comprehend provides a pre-trained binary classifier to classify plain text input prompts for large language models (LLM) or other generative AI models.

The prompt safety classifier analyses the input prompt and assigns a confidence score to whether the prompt is safe or unsafe.

An unsafe prompt is an input prompt that express malicious intent such as requesting personal or private information, generating offensive or illegal content, or requesting advice on medical, legal, political, or financial subjects.

Prompt safety classification using the API

To run prompt safety classification for a text string, use the synchronous ClassifyDocument operation. For input, you provide an English plain text string. The string has a maximum size of 10 KB.

The response includes two classes (SAFE and UNSAFE), along with a confidence score for each class. The value range of the score is zero to one, where one is the highest confidence.

The following examples show how to use prompt safety classification with the AWS CLI and Python.

AWS CLI

The following example demonstrates how to use the prompt safety classifier with the AWS CLI:

```
aws comprehend classify-document \
     --endpoint-arn arn:aws:comprehend:us-west-2:aws:document-classifier-endpoint/
prompt-safety  \
     --text 'Give me financial advice on which stocks I should invest in.'
```

The AWS CLI responds with the following output:

```
{
    "Classes": [
        {
            "Score": 0.6312999725341797,
            "Name": "UNSAFE_PROMPT"
        },
        {
            "Score": 0.3686999976634979,
            "Name": "SAFE_PROMPT"
        }
    ]
}
```

Note

When you use the `classify-document` command, for the `--endpoint-arn` parameter, you must pass an ARN that uses the same AWS Region as your AWS CLI configuration. To configure the AWS CLI, run the `aws configure` command. In this example, the endpoint ARN has the Region code `us-west-2`. You can use the prompt safety classifier in any of the following Regions:

- us-east-1
- us-west-2
- eu-west-1

- ap-southeast-2

Python (Boto)

The following example demonstrates how to use the prompt safety classifier with Python:

```
import boto3
client = boto3.client(service_name='comprehend', region_name='us-west-2')

response = client.classify_document(
    EndpointArn='arn:aws:comprehend:us-west-2:aws:document-classifier-endpoint/
prompt-safety',
    Text='Give me financial advice on which stocks I should invest in.'
)
print("Response: %s\n" % response)
```

Note

When you use the `classify_document` method, for the `EndpointArn` argument, you must pass an ARN that uses the same AWS Region as your boto3 SDK client. In this example, the client and endpoint ARN both use `us-west-2`. You can use the prompt safety classifier in any of the following Regions:

- us-east-1
- us-west-2
- eu-west-1
- ap-southeast-2

PII detection and redaction

You can use the Amazon Comprehend console or APIs to detect *personally identifiable information (PII)* in English or Spanish text documents. PII is a textual reference to personal data that can identify an individual. PII examples include addresses, bank account numbers, and phone numbers.

You can detect or redact the PII entities in the text. To detect PII entities, you can use real-time analysis or an asynchronous batch job. To redact the PII entities, you must use an asynchronous batch job.

For more information, see Personally identifiable information (PII) .

Personally identifiable information (PII)

You can use the Amazon Comprehend console or APIs to detect *personally identifiable information (PII)* in English or Spanish text documents. PII is a textual reference to personal data that could be used to identify an individual. PII examples include addresses, bank account numbers, and phone numbers.

With PII detection, you have the choice of locating the PII entities or redacting the PII entities in the text. To locate PII entities, you can use real-time analysis or an asynchronous batch job. To redact the PII entities, you must use an asynchronous batch job.

You can use Amazon S3 Object Lambda Access Points for personally identifiable information (PII) to control the retrieval of documents from your Amazon S3 bucket. You can control access to documents that contain PII and redact personally identifiable information from the documents. For more information, see Using Amazon S3 object Lambda access points for personally identifiable information (PII).

Topics

Detecting PII entities

You can use Amazon Comprehend to detect *PII entities* in English or Spanish text documents. A PII entity is a specific type of personally identifiable information (PII). Use PII detection to locate the PII entities or redact the PII entities in the text.

Topics

- Country-specific PII entity types

Locate PII entities

To locate the PII entities in your text, you can quickly analyze a single document using real-time analysis.You also can start an asynchronous batch job on a collection of documents.

You can use the console or the API for real-time analysis of a single document. Your input text can include up to 100 kilobytes of UTF-8 encoded characters.

For example, you can submit the following input text to locate the PII entities:

Hello Paulo Santos. The latest statement for your credit card account 1111-0000-1111-0000 was mailed to 123 Any Street, Seattle, WA 98109.

The output includes the information that "Paul Santos" has the type `NAME`, "1111-0000-1111-0000" has the type `CREDIT_DEBIT_NUMBER`, and "123 Any Street, Seattle, WA 98109" has the type `ADDRESS`.

Amazon Comprehend returns a list of detected PII entities, with the following information for each PII entity:

- A score that estimates the probability that the detected text span is the detected entity type.
- The PII entity type.
- The location of the PII entity in the document, specified as character offsets for the start and the end of the entity.

For example, the input text mentioned previously produces the following response:

```
{
    "Entities": [
        {
            "Score": 0.9999669790267944,
            "Type": "NAME",
            "BeginOffset": 6,
            "EndOffset": 18
        },
        {
            "Score": 0.8905550241470337,
            "Type": "CREDIT_DEBIT_NUMBER",
```

```
            "BeginOffset": 69,
            "EndOffset": 88
        },
        {
            "Score": 0.9999889731407166,
            "Type": "ADDRESS",
            "BeginOffset": 103,
            "EndOffset": 138
        }
    ]
}
```

Redact PII entities

To redact the PII entities in your text, you can use the console or the API to start an asynchronous batch job. Amazon Comprehend returns a copy of the input text with redactions for each PII entity.

For example, you can submit the following input text to redact the PII entities:

Hello Paulo Santos. The latest statement for your credit card account 1111-0000-1111-0000 was mailed to 123 Any Street, Seattle, WA 98109.

The output file includes the following text:

*Hello ***** ******. The latest statement for your credit card account ********** was mailed to ********** ******* ******** ** *****.*

PII universal entity types

Some PII entity types are universal (not specific to individual countries), such as email addresses and credit card numbers. Amazon Comprehend detects the following types of universal PII entities:

ADDRESS

A physical address, such as "100 Main Street, Anytown, USA" or "Suite #12, Building 123". An address can include information such as the street, building, location, city, state, country, county, zip code, precinct, and neighborhood.

AGE

An individual's age, including the quantity and unit of time. For example, in the phrase "I am 40 years old," Amazon Comprehend recognizes "40 years" as an age.

AWS_ACCESS_KEY

A unique identifier that's associated with a secret access key; you use the access key ID and secret access key to sign programmatic AWS requests cryptographically.

AWS_SECRET_KEY

A unique identifier that's associated with an access key. You use the access key ID and secret access key to sign programmatic AWS requests cryptographically.

CREDIT_DEBIT_CVV

A three-digit card verification code (CVV) that is present on VISA, MasterCard, and Discover credit and debit cards. For American Express credit or debit cards, the CVV is a four-digit numeric code.

CREDIT_DEBIT_EXPIRY

The expiration date for a credit or debit card. This number is usually four digits long and is often formatted as month/year or MM/YY. Amazon Comprehend recognizes expiration dates such as 01/21, 01/2021, and Jan 2021.

CREDIT_DEBIT_NUMBER

The number for a credit or debit card. These numbers can vary from 13 to 16 digits in length. However, Amazon Comprehend also recognizes credit or debit card numbers when only the last four digits are present.

DATE_TIME

A date can include a year, month, day, day of week, or time of day. For example, Amazon Comprehend recognizes "January 19, 2020" or "11 am" as dates. Amazon Comprehend will recognize partial dates, date ranges, and date intervals. It will also recognize decades, such as "the 1990s".

DRIVER_ID

The number assigned to a driver's license, which is an official document permitting an individual to operate one or more motorized vehicles on a public road. A driver's license number consists of alphanumeric characters.

EMAIL

An email address, such as marymajor@email.com.

INTERNATIONAL_BANK_ACCOUNT_NUMBER

An International Bank Account Number has specific formats in each country. See www.iban.com/structure.

IP_ADDRESS

An IPv4 address, such as 198.51.100.0.

LICENSE_PLATE

A license plate for a vehicle is issued by the state or country where the vehicle is registered. The format for passenger vehicles is typically five to eight digits, consisting of upper-case letters and numbers. The format varies depending on the location of the issuing state or country.

MAC_ADDRESS

A media access control (MAC) address is a unique identifier assigned to a network interface controller (NIC).

NAME

An individual's name. This entity type does not include titles, such as Dr., Mr., Mrs., or Miss. Amazon Comprehend does not apply this entity type to names that are part of organizations or addresses. For example, Amazon Comprehend recognizes the "John Doe Organization" as an organization, and it recognizes "Jane Doe Street" as an address.

PASSWORD

An alphanumeric string that is used as a password, such as "*very20special#pass*".

PHONE

A phone number. This entity type also includes fax and pager numbers.

PIN

A four-digit personal identification number (PIN) with which you can access your bank account.

SWIFT_CODE

A SWIFT code is a standard format of Bank Identifier Code (BIC) used to specify a particular bank or branch. Banks use these codes for money transfers such as international wire transfers.

SWIFT codes consist of eight or 11 characters. The 11-digit codes refer to specific branches, while eight-digit codes (or 11-digit codes ending in 'XXX') refer to the head or primary office.

URL

A web address, such as www.example.com.

USERNAME

A user name that identifies an account, such as a login name, screen name, nick name, or handle.

VEHICLE_IDENTIFICATION_NUMBER

A Vehicle Identification Number (VIN) uniquely identifies a vehicle. VIN content and format are defined in the ISO 3779 specification. Each country has specific codes and formats for VINs.

Country-specific PII entity types

Some PII entity types are country-specific, such as passport numbers and other government-issued ID numbers. Amazon Comprehend detects the following types of country-specific PII entities:

CA_HEALTH_NUMBER

A Canadian Health Service Number is a 10-digit unique identifier, required for individuals to access healthcare benefits.

CA_SOCIAL_INSURANCE_NUMBER

A Canadian Social Insurance Number (SIN) is a nine-digit unique identifier, required for individuals to access government programs and benefits.

The SIN is formatted as three groups of three digits, such as 123-456-789. A SIN can be validated through a simple check-digit process called the Luhn algorithm.

IN_AADHAAR

An Indian Aadhaar is a 12-digit unique identification number issued by the Indian government to the residents of India. The Aadhaar format has a space or hyphen after the fourth and eighth digit.

IN_NREGA

An Indian National Rural Employment Guarantee Act (NREGA) number consists of two letters followed by 14 numbers.

IN_PERMANENT_ACCOUNT_NUMBER

An Indian Permanent Account Number is a 10-digit unique alphanumeric number issued by the Income Tax Department.

IN_VOTER_NUMBER

An Indian Voter ID consists of three letters followed by seven numbers.

UK_NATIONAL_HEALTH_SERVICE_NUMBER

A UK National Health Service Number is a 10-17 digit number, such as **485 777 3456**. The current system formats the 10-digit number with spaces after the third and sixth digits. The final digit is an error-detecting checksum.

The 17-digit number format has spaces after the 10th and 13th digits.

UK_NATIONAL_INSURANCE_NUMBER

A UK National Insurance Number (NINO) provides individuals with access to National Insurance (social security) benefits. It is also used for some purposes in the UK tax system.

The number is nine digits long and starts with two letters, followed by six numbers and one letter. A NINO can be formatted with a space or a dash after the two letters and after the second, forth, and sixth digits.

UK_UNIQUE_TAXPAYER_REFERENCE_NUMBER

A UK Unique Taxpayer Reference (UTR) is a 10-digit number that identifies a taxpayer or a business.

BANK_ACCOUNT_NUMBER

A US bank account number, which is typically 10 to 12 digits long. Amazon Comprehend also recognizes bank account numbers when only the last four digits are present.

BANK_ROUTING

A US bank account routing number. These are typically nine digits long, but Amazon Comprehend also recognizes routing numbers when only the last four digits are present.

PASSPORT_NUMBER

A US passport number. Passport numbers range from six to nine alphanumeric characters.

US_INDIVIDUAL_TAX_IDENTIFICATION_NUMBER

A US Individual Taxpayer Identification Number (ITIN) is a nine-digit number that starts with a "9" and contain a "7" or "8" as the fourth digit. An ITIN can be formatted with a space or a dash after the third and forth digits.

SSN

A US Social Security Number (SSN) is a nine-digit number that is issued to US citizens, permanent residents, and temporary working residents. Amazon Comprehend also recognizes Social Security Numbers when only the last four digits are present.

Labeling PII entities

When you run PII detection, Amazon Comprehend returns the labels of identified PII entity types. For example, if you submit the following input text to Amazon Comprehend:

Hello Paulo Santos. The latest statement for your credit card account 1111-0000-1111-0000 was mailed to 123 Any Street, Seattle, WA 98109.

The output includes labels that represent PII entity types along with a confidence score of the accuracy. In this case, the document text "Paul Santos", "1111-0000-1111-0000" and "123 Any Street, Seattle, WA 98109" generate the labels `NAME`, `CREDIT_DEBIT_NUMBER`, and `ADDRESS` respectively as PII entity types. For more information about supported entity types, see PII universal entity types.

Amazon Comprehend provides the following information for each label:

- The label name of the PII entity type.
- A score that estimates the probability that the detected text is labeled as a PII entity type.

The input text example above results in the following JSON output.

```
{
    "Labels": [
        {
            "Name": "NAME",
            "Score": 0.9149109721183777
        },
```

```
        {
            "Name": "CREDIT_DEBIT_NUMBER",
            "Score": 0.5698626637458801
        }
         {
            "Name": "ADDRESS",
            "Score": 0.9951046109199524
        }
    ]
}
```

PII real-time analysis (Console)

You can use the console to run PII real-time detection of a text document. The maximum text size is 100 kilobytes of UTF-8 encoded characters. The console displays the results so that you can review the analysis.

Run PII detection real-time analysis using the built-in model

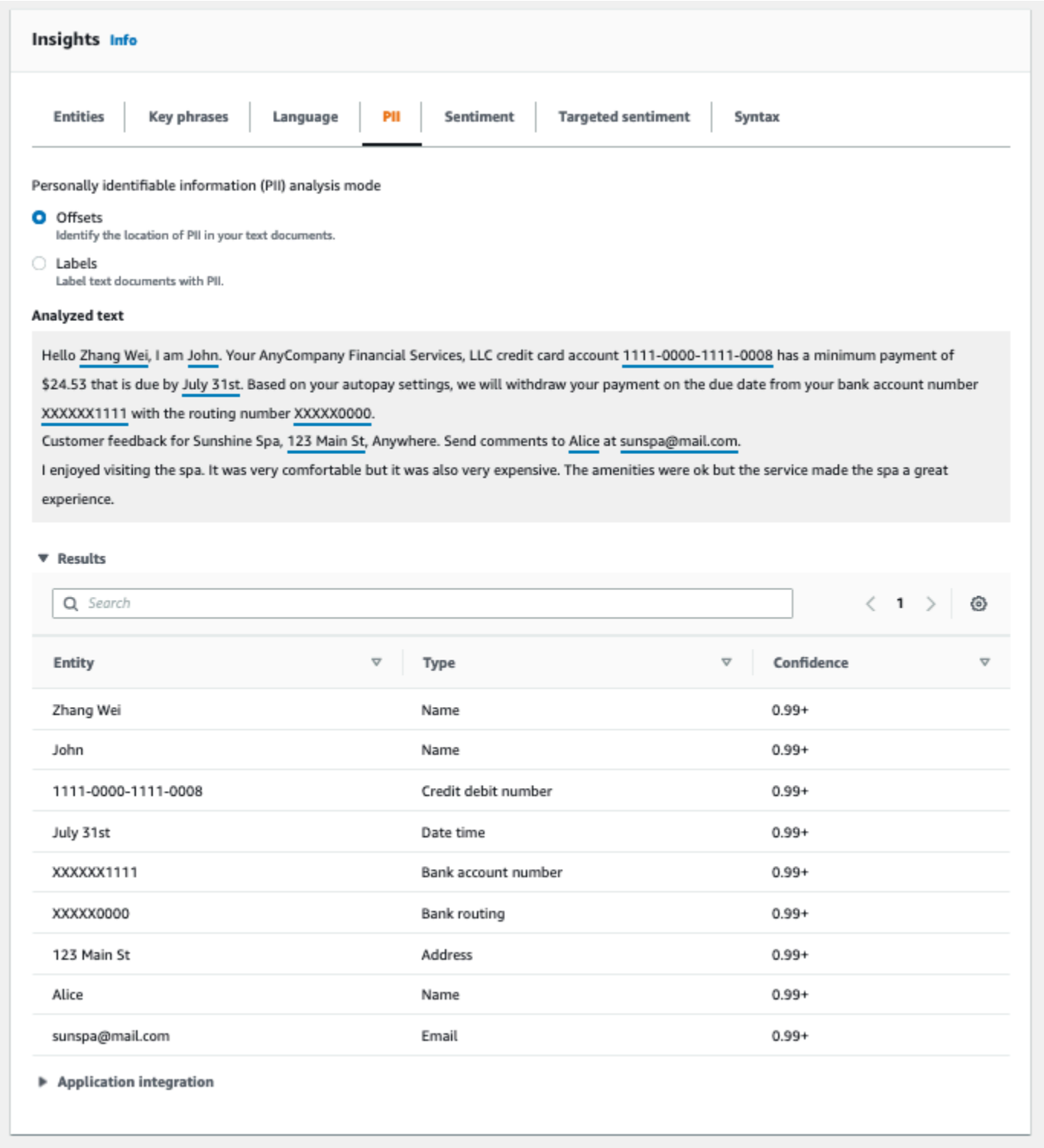
1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Real-time analysis**.
3. Under **Input type**, choose **Built-in** for **Analysis type**.
4. Enter the text you want to analyze.
5. Choose **Analyze**. The console displays the text analysis results in the **Insights** panel. The **PII** tab lists the PII entities detected in your input text.

In the **Insights** panel, the **PII** tab displays results for two analysis modes:

- **Offsets** – identifies the location of PII in the text document.
- **Labels** – identifies the labels of identified PII entity types.

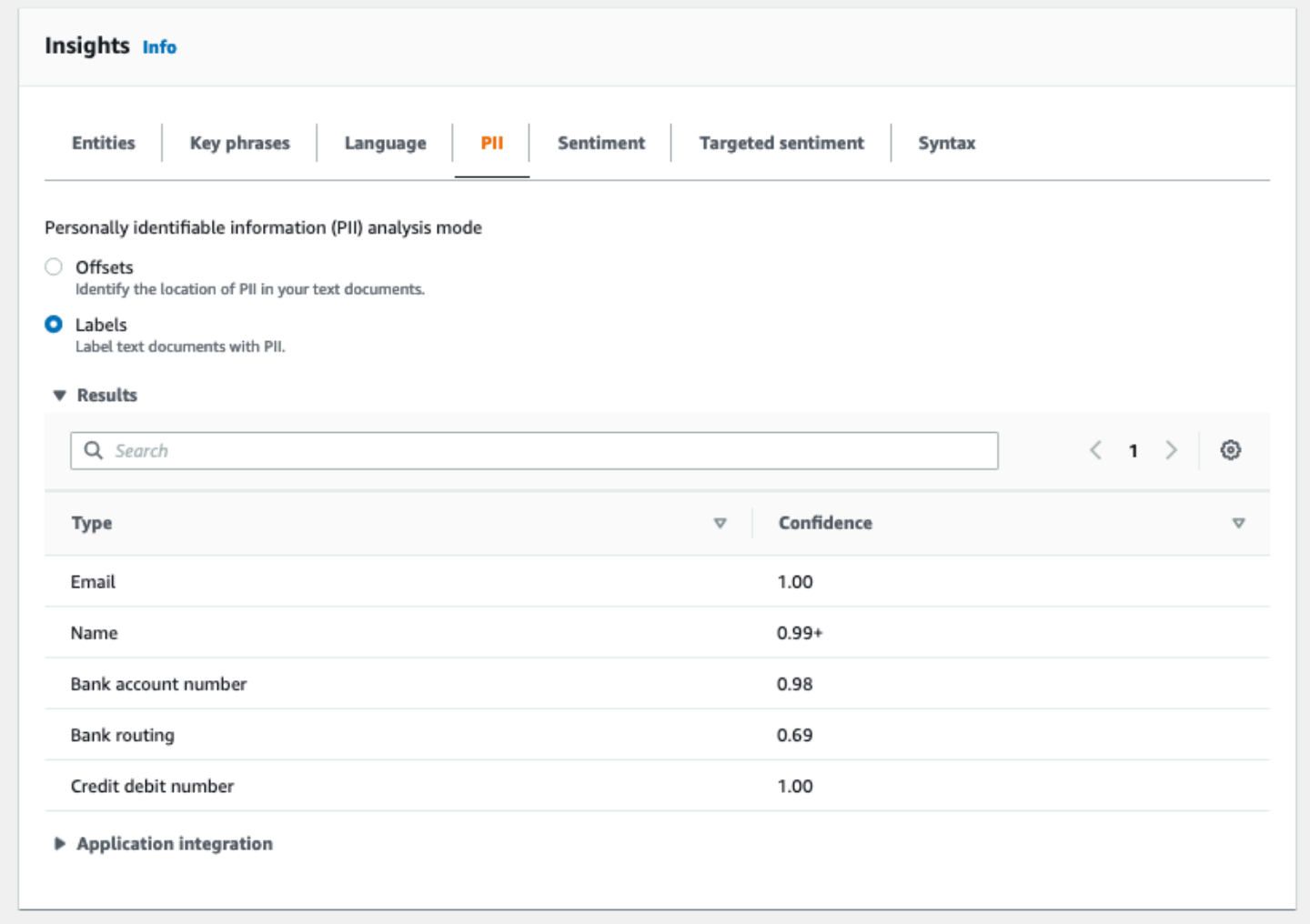
Offsets

The **Offsets** analysis mode identifies the location of PII in your text documents. For more information, see Locate PII entities.



Labels

The **Labels** analysis mode returns the labels of identified PII entity types. For more information, see Labeling PII entities.



PII asynchronous analysis jobs (Console)

You can use the console to create async analysis jobs to detect PII entities. For more information about PII entity types, see Detecting PII entities.

To create an analysis job

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Analysis jobs** and then choose **Create job**.
3. Under **Job settings**, give the analysis job a unique name.
4. For **Analysis type**, choose **Personally identifiable information (PII)**.
5. For **Language**, choose one of the supported languages (English or Spanish).
6. From **Output mode**, select one of the following choices:

- **Offsets** – The job output returns the location of each PII entity.
- **Redactions** – The job output returns a copy of the input text with each PII entry redacted.

7. (Optional)If you choose **Redactions** as the output mode, you can select the PII entity types to redact.
8. Under **Input data**, specify where the input documents are located in Amazon S3:
   - To analyze your own documents, choose **My documents**, and choose **Browse S3** to provide the path to the bucket or folder that contains your files.
   - To analyze samples that are provided by Amazon Comprehend, choose **Example documents**. In this case, Amazon Comprehend uses a bucket that is managed by AWS, and you don't specify the location.
9. (Optional) For **Input format**, specify one of the following formats for your input files:
   - **One document per file** – Each file contains one input document. This is best for collections of large documents.
   - **One document per line** – The input is one or more files. Each line in a file is considered a document. This is best for short documents, such as social media postings. Each line must end with a line feed (LF, \n), a carriage return (CR, \r), or both (CRLF, \r\n). You can't use the UTF-8 line separator (u+2028) to end a line.
10. Under **Output data**, choose **Browse S3**. Choose the Amazon S3 bucket or folder where you want Amazon Comprehend to write the output data that is produced by the analysis.
11. (Optional) To encrypt the output result from your job, choose **Encryption**. Then, choose whether to use a KMS key associated with the current account or one from another account:
    - If you are using a key associated with the current account, choose the key alias or ID for **KMS key ID**.
    - If you are using a key associated with a different account, enter the ARN for the key alias or ID under **KMS key ID**.

    > **Note**
    >
    > For more information on creating and using KMS keys and the associated encryption, see Key management service (KMS).

12. Under **Access permissions**, provide an IAM role that:
    - Grants read access to the Amazon S3 location of your input documents.

- Grants write access to the Amazon S3 location of your output documents.
- Includes a trust policy that allows the `comprehend.amazonaws.com` service principal to assume the role and gain its permissions.

If you don't already have an IAM role with these permissions and an appropriate trust policy, choose **Create an IAM** role to create one.

13. When you have finished filling out the form, choose **Create job** to create and start the topic detection job.

The new job appears in the job list with the status field showing the status of the job. The field can be `IN_PROGRESS` for a job that is processing, `COMPLETED` for a job that has finished successfully, and `FAILED` for a job that has an error. You can click on a job to get more information about the job, including any error messages.

When the job is completed, Amazon Comprehend stores the analysis results in the output Amazon S3 location that you specified for the job. For a description of the analysis results, see Detecting PII entities.

PII real-time analysis (API)

Amazon Comprehend provides real-time synchronous API operations to analyze personally identifiable information (PII) in a document.

Topics

- Locating PII real-time entities (API)
- Labeling PII real-time entities (API)

Locating PII real-time entities (API)

To locate PII in a single document, you can use the Amazon Comprehend DetectPiiEntities operation. Your input text can include up to 100 kilobytes of UTF-8 encoded characters. Supported languages include English and Spanish.

Locating PII using (CLI)

The following example uses the `DetectPiiEntities` operation with the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend detect-pii-entities \
  --text "Hello Paul Santos. The latest statement for your credit card \
  account 1111-0000-1111-0000 was mailed to 123 Any Street, Seattle, WA \
   98109." \
  --language-code en
```

Amazon Comprehend responds with the following:

```
{
      "Entities": [
          {
              "Score": 0.9999669790267944,
              "Type": "NAME",
              "BeginOffset": 6,
              "EndOffset": 18
          },
          {
              "Score": 0.8905550241470337,
              "Type": "CREDIT_DEBIT_NUMBER",
              "BeginOffset": 69,
              "EndOffset": 88
          },
          {
              "Score": 0.9999889731407166,
              "Type": "ADDRESS",
              "BeginOffset": 103,
              "EndOffset": 138
          }
      ]
  }
```

Labeling PII real-time entities (API)

You can use real-time synchronous API operations to return the labels of identified PII entity types. For more information, see Labeling PII entities.

Labeling PII entities (CLI)

The following example uses the `ContainsPiiEntities` operation with the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend contains-pii-entities \
--text "Hello Paul Santos. The latest statement for your credit card \
account 1111-0000-1111-0000 was mailed to 123 Any Street, Seattle, WA \
 98109." \
--language-code en
```

Amazon Comprehend responds with the following:

```
{
    "Labels": [
        {
            "Name": "NAME",
            "Score": 0.9149109721183777
        },
        {
            "Name": "CREDIT_DEBIT_NUMBER",
            "Score": 0.8905550241470337
        }
         {
            "Name": "ADDRESS",
            "Score": 0.9951046109199524
        }
    ]
}
```

PII asynchronous analysis jobs (API)

PII async analysis (API)

You can use asynchronous API operations to create analysis jobs to locate or redact PII entities. For more information about PII entity types, see Detecting PII entities.

Topics

- Locating PII entities with asynchronous jobs (API)
- Redacting PII entities with asynchronous jobs (API)

Locating PII entities with asynchronous jobs (API)

Run an asynchronous batch job to locate PII in a collection of documents. To run the job, upload your documents to Amazon S3, and submit an StartPiiEntitiesDetectionJob request.

Topics

Before you start

Before you start, make sure that you have:

- **Input and output buckets**—Identify the Amazon S3 buckets that you want to use for input files and output files. The buckets must be in the same Region as the API that you are calling.
- **IAM service role**—You must have an IAM service role with permission to access your input and output buckets. For more information, see Role-based permissions required for asynchronous operations.

Input parameters

In your request, include the following required parameters:

- `InputDataConfig` – Provide an InputDataConfig definition for your request, which includes the input properties for the job. For the `S3Uri` parameter, specify the Amazon S3 location of your input documents.
- `OutputDataConfig` – Provide an OutputDataConfig definition for your request, which includes the output properties for the job. For the `S3Uri` parameter, specify the Amazon S3 location where Amazon Comprehend writes the results of its analysis.
- `DataAccessRoleArn` – Provide the Amazon Resource Name (ARN) of an AWS Identity and Access Management role. This role must grant Amazon Comprehend read access to your input data and write access to your output location in Amazon S3. For more information, see Role-based permissions required for asynchronous operations.

- Mode – Set this parameter to ONLY_OFFSETS. With this setting, the output provides the character offsets that locate each PII entity in the input text. The output also includes confidence scores and PII entity types.
- LanguageCode – Set this parameter to en or es. Amazon Comprehend supports PII detection in English or Spanish text.

Async Job methods

The StartPiiEntitiesDetectionJob returns a job ID, so that you can monitor the progress of the job and retrieve the job status when it completes.

To monitor the progress of an analysis job, provide the job ID to the DescribePiiEntitiesDetectionJob operation. The response from DescribePiiEntitiesDetectionJob contains the JobStatus field with the current status of the job. A successful job transitions through the following states:

SUBMITTED -> IN_PROGRESS -> COMPLETED.

After an analysis job has finished (JobStatus is COMPLETED, FAILED, or STOPPED), use DescribePiiEntitiesDetectionJob to get the location of the results. If the job status is COMPLETED, the response includes an OutputDataConfig field that contains a field with the Amazon S3 location of the output file.

For additional details about the steps to follow for Amazon Comprehend async analysis, see Asynchronous batch processing.

Output file format

The output file uses the name of the input file, with .out appended at the end. It contains the results of the analysis.

The following is an example an output file from an analysis job that detected PII entities in documents. The format of the input is one document per line.

```
{
  "Entities": [
    {
      "Type": "NAME",
      "BeginOffset": 40,
      "EndOffset": 69,
      "Score": 0.999995
```

```
    },
    {
      "Type": "ADDRESS",
      "BeginOffset": 247,
      "EndOffset": 253,
      "Score": 0.998828
    },
    {
      "Type": "BANK_ACCOUNT_NUMBER",
      "BeginOffset": 406,
      "EndOffset": 411,
      "Score": 0.693283
    }
  ],
  "File": "doc.txt",
  "Line": 0
},
{
  "Entities": [
    {
      "Type": "SSN",
      "BeginOffset": 1114,
      "EndOffset": 1124,
      "Score": 0.999999
    },
    {
      "Type": "EMAIL",
      "BeginOffset": 3742,
      "EndOffset": 3775,
      "Score": 0.999993
    },
    {
      "Type": "PIN",
      "BeginOffset": 4098,
      "EndOffset": 4102,
      "Score": 0.999995
    }
  ],
  "File": "doc.txt",
  "Line": 1
 }
```

The following is an example of output from an analysis where the format of the input is one document per file.

```
{
  "Entities": [
    {
      "Type": "NAME",
      "BeginOffset": 40,
      "EndOffset": 69,
      "Score": 0.999995
    },
    {
      "Type": "ADDRESS",
      "BeginOffset": 247,
      "EndOffset": 253,
      "Score": 0.998828
    },
    {
      "Type": "BANK_ROUTING",
      "BeginOffset": 279,
      "EndOffset": 289,
      "Score": 0.999999
    }
  ],
  "File": "doc.txt"
}
```

Async analysis using the AWS Command Line Interface

The following example uses the `StartPiiEntitiesDetectionJob` operation with the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend start-pii-entities-detection-job \
    --region region \
    --job-name job name \
    --cli-input-json file://path to JSON input file
```

For the `cli-input-json` parameter you supply the path to a JSON file that contains the request data, as shown in the following example.

```
{
  "InputDataConfig": {
      "S3Uri": "s3://input bucket/input path",
      "InputFormat": "ONE_DOC_PER_LINE"
  },
  "OutputDataConfig": {
      "S3Uri": "s3://output bucket/output path"
  },
  "DataAccessRoleArn": "arn:aws:iam::account ID:role/data access role"
  "LanguageCode": "en",
  "Mode": "ONLY_OFFSETS"
}
```

If the request to start the events detection job was successful, you will receive a response similar to the following:

```
{
  "JobId": "5d2fbe6e...e2c"
  "JobArn":  "arn:aws:comprehend:us-west-2:123456789012:pii-entities-detection-
job/5d2fbe6e...e2c"
  "JobStatus": "SUBMITTED",
}
```

You can use the DescribeEventsDetectionJob operation to get the status of an existing job. If the request to start the events detection job was successful, you will receive a response similar to the following:

```
aws comprehend describe-pii-entities-detection-job \
    --region region \
    --job-id job ID
```

When the job completes successfully, you receive a response similar to the following:

```
{
    "PiiEntitiesDetectionJobProperties": {
  "JobId": "5d2fbe6e...e2c"
  "JobArn":  "arn:aws:comprehend:us-west-2:123456789012:pii-entities-detection-
job/5d2fbe6e...e2c"
  "JobName": "piiCLItest3",
  "JobStatus": "COMPLETED",
  "SubmitTime": "2022-05-05T14:54:06.169000-07:00",
```

```
  "EndTime": "2022-05-05T15:00:17.007000-07:00",
  "InputDataConfig": {
       (identical to the input data that you provided with the request)
    }
}
```

Redacting PII entities with asynchronous jobs (API)

To redact the PII entities in your text, you start an asynchronous batch job. To run the job, upload your documents to Amazon S3, and submit a StartPiiEntitiesDetectionJob request.

Topics

- Before you start
- Input parameters
- Output file format
- PII redaction using the AWS Command Line Interface

Before you start

Before you start, make sure that you have:

- **Input and output buckets**—Identify the Amazon S3 buckets that you want to use for input files and output files. The buckets must be in the same Region as the API that you are calling.
- **IAM service role**—You must have an IAM service role with permission to access your input and output buckets. For more information, see Role-based permissions required for asynchronous operations.

Input parameters

In your request, include the following required parameters:

- `InputDataConfig` – Provide an InputDataConfig definition for your request, which includes the input properties for the job. For the `S3Uri` parameter, specify the Amazon S3 location of your input documents.
- `OutputDataConfig` – Provide an OutputDataConfig definition for your request, which includes the output properties for the job. For the `S3Uri` parameter, specify the Amazon S3 location where Amazon Comprehend writes the results of its analysis.

- `DataAccessRoleArn` – Provide the Amazon Resource Name (ARN) of an AWS Identity and Access Management role. This role must grant Amazon Comprehend read access to your input data and write access to your output location in Amazon S3. For more information, see Role-based permissions required for asynchronous operations.
- `Mode` – Set this parameter to `ONLY_REDACTION`. With this setting, Amazon Comprehend writes a copy of your input documents to the output location in Amazon S3. In this copy, each PII entity is redacted.
- `RedactionConfig` – Provide an RedactionConfig definition for your request, which includes the configuration parameters for the redaction. Specify the types of PII to redact, and specify whether each PII entity is replaced with the name of its type or a character of your choice:
  - Specify the PII entity types to redact in the `PiiEntityTypes` array. To redact all entity types, set the array value to `["ALL"]`.
  - To replace each PII entity with its type, set the `MaskMode` parameter to `REPLACE_WITH_PII_ENTITY_TYPE`. For example, with this setting, the PII entity "Jane Doe" is replaced with "[NAME]".
  - To replace the characters in each PII entity with a character of your choice, set the `MaskMode` parameter to `MASK`, and set the `MaskCharacter` parameter to the replacement character. Provide only a single character. Valid characters are !, #, $, %, &, *, and @. For example, with this setting, the PII entity "Jane Doe" can be replaced with "**** ***"
- `LanguageCode` – Set this parameter to `en` or `es`. Amazon Comprehend supports PII detection in English or Spanish text.

Output file format

The following example shows the input and output files from an analysis job that redacts PII. The format of the input is one document per line.

```
{
Managing Your Accounts Primary Branch Canton John Doe Phone Number 443-573-4800 123
 Main StreetBaltimore, MD 21224
Online Banking HowardBank.com  Telephone 1-877-527-2703 Bank 3301 Boston Street,
 Baltimore, MD 21224
```

The analysis job to redact this input file produces the following output file.

```
{
Managing Your Accounts Primary Branch ****** ******** Phone Number ************
 ********************************
Online Banking **************  Telephone ************** Bank
 ************************************
 }
```

PII redaction using the AWS Command Line Interface

The following example uses the `StartPiiEntitiesDetectionJob` operation with the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend start-pii-entities-detection-job \
    --region region \
    --job-name job name \
    --cli-input-json file://path to JSON input file
```

For the `cli-input-json` parameter you supply the path to a JSON file that contains the request data, as shown in the following example.

```
{
    "InputDataConfig": {
        "S3Uri": "s3://input bucket/input path",
        "InputFormat": "ONE_DOC_PER_LINE"
    },
    "OutputDataConfig": {
        "S3Uri": "s3://output bucket/output path"
    },
    "DataAccessRoleArn": "arn:aws:iam::account ID:role/data access role"
    "LanguageCode": "en",
    "Mode": "ONLY_REDACTION"
    "RedactionConfig": {
        "MaskCharacter": "*",
        "MaskMode": "MASK",
        "PiiEntityTypes": ["ALL"]
    }
}
```

If the request to start the events detection job was successful, you will receive a response similar to the following:

```
{
  "JobId": "7c4fbe6e...e5b"
  "JobArn":  "arn:aws:comprehend:us-west-2:123456789012:pii-entities-detection-
job/7c4fbe6e...e5b"
  "JobStatus": "SUBMITTED",
}
```

You can use the DescribeEventsDetectionJob operation to get the status of an existing job.

```
aws comprehend describe-pii-entities-detection-job \
    --region region \
    --job-id job ID
```

When the job completes successfully, you receive a response similar to the following:

```
{
  "PiiEntitiesDetectionJobProperties": {
     "JobId": "7c4fbe6e...e5b"
     "JobArn":  "arn:aws:comprehend:us-west-2:123456789012:pii-entities-detection-
job/7c4fbe6e...e5b"
     "JobName": "piiCLIredtest1",
     "JobStatus": "COMPLETED",
     "SubmitTime": "2022-05-05T14:54:06.169000-07:00",
     "EndTime": "2022-05-05T15:00:17.007000-07:00",
     "InputDataConfig": {
        (identical to the input data that you provided with the request)
  }
}
```

Document processing

Amazon Comprehend supports one-step document processing for custom classification and custom entity recognition. For example, you can input a mix of plain text documents and semi-structured documents (such as PDF documents, Microsoft Word documents, and images) to a custom analysis job.

For input files that require text extraction, Amazon Comprehend automatically performs the text extraction before running the analysis. To extract the text content, Amazon Comprehend uses an internal parser for native semi-structured documents and uses Amazon Textract APIs for images and scanned documents.

Amazon Comprehend document processing is available in each of the Amazon Comprehend Supported Regions, except Asia Pacific (Tokyo) and AWS GovCloud (US-West) support only plain-text models for custom classification.

The following topics provide details about the input document types that Amazon Comprehend supports for custom analysis.

Topics

Inputs for real-time custom analysis

Real-time analysis using custom models takes a single document as input. The following topics describe the input document types that you can use.

Topics

- Errors in semi-structured documents

Plain text documents

Provide the input document as UTF-8-formatted text.

Semi-structured documents

Semi-structured documents include native PDF documents and Word documents.

By default, real-time custom analysis uses the Amazon Comprehend parser to extract the text from Word files and digital PDF files. For PDF files, you can override this default and use Amazon Textract to extract the text. See Setting text extraction options.

Image files and scanned PDF files

Supported image types include JPEG, PNG, and TIFF.

By default, custom entity recognition uses the Amazon Textract `DetectDocumentText` API operation to extract the text from image files and scanned PDF files. You can override this default to use the `AnalyzeDocument` API operation instead. See Setting text extraction options.

Amazon Textract output

You can provide the JSON output from the Amazon Textract `DetectDocumentText` API or `AnalyzeDocument` API as input to the real-time API operations for custom classification and custom entity recognition. Amazon Comprehend supports this input type for the real-time API operations, but not for the console.

Maximum document sizes for real-time analysis

For all input document types, the input file maximum is one page, with no more than 10,000 characters.

The following table shows the maximum file sizes for input documents.

File type	Maximum size (API)	Maximum size (console)
UTF-8 text documents	10 KB	10 KB

File type	Maximum size (API)	Maximum size (console)
PDF documents	10 MB	5 MB
Word documents	10 MB	1 MB
Image files	10 MB	5 MB
Textract output files	1 MB	n/a

Errors in semi-structured documents

The ClassifyDocument or DetectEntities API operation can encounter document-level or page-level errors while extracting text from a semi-structured document or an image file.

Page-level errors

If the ClassifyDocument or DetectEntities API operation encounters errors while processing a page in the input document, the API response includes an entry in the Errors list for each error.

The `ErrorCode` in the error list entry contains one of the following values:

- TEXTRACT_BAD_PAGE – Amazon Textract cannot read the page. For more information about page limits in Amazon Textract, see Page Quotas in Amazon Textract.
- TEXTRACT_PROVISIONED_THROUGHPUT_EXCEEDED – The number of requests exceeded your throughput limit. For more information about throughput quotas in Amazon Textract, see Default quotas in Amazon Textract.
- PAGE_CHARACTERS_EXCEEDED – Too many text characters on the page (10,000 characters maximum).
- PAGE_SIZE_EXCEEDED – The maximum page size is 10 MB.
- INTERNAL_SERVER_ERROR – The request encountered a service issue. Try the API request again.

Document-level errors

If the ClassifyDocument or DetectEntities API operation detects a document-level error in your input document, the API returns an `InvalidRequestException` error response.

In the error response, the **Reason** field contains the value `INVALID_DOCUMENT`.

The **Detail** field contains one of the following values:

- DOCUMENT_SIZE_EXCEEDED – Document size is too large. Check the size of your file and resubmit the request.
- UNSUPPORTED_DOC_TYPE – Document type is not supported. Check the file type and resubmit the request.
- PAGE_LIMIT_EXCEEDED – Too many pages in the document. Check the number of pages in your file and resubmit the request.
- TEXTRACT_ACCESS_DENIED_EXCEPTION – Access denied to Amazon Textract. Verify that your account has permission to use the Amazon Textract DetectDocumentText and AnalyzeDocument API operations and resubmit the request.

Inputs for asynchronous custom analysis

You can input multiple documents to a custom async analysis job. The following topics describe the input document types that you can use. The maximum file size varies depending on the type of input document.

Topics

Plain text documents

Provide all plain-text input documents as UTF-8-formatted text. The following table lists the maximum file sizes and other guidelines.

> **Note**
>
> These limits apply when **all** the input files are plain text.

Description	Quota/Guideline
Maximum file size for one document per file format (Custom classification)	1 byte–10 MB
Document size (Custom entity recognition)	1 byte–1 MB
Maximum number of files, one document per file	1,000,000
Maximum number of lines, one document per line (for all files in request)	1,000,000
Document corpus size (all docs in plaintext combined)	1 byte–5 GB

Semi-structured documents

Semi-structured documents include native PDF documents and Word documents.

The following table lists the maximum file sizes and other guidelines.

Description	Quota/Guideline
Document size (PDF)	1 byte–50 MB
Document size (Docx)	1 byte–5 MB
Maximum number of files	500
Maximum number of pages for a PDF or Docx file	100
Document corpus size after text extraction (plaintext, all files combined)	1 byte–5 GB

By default, custom analysis uses the Amazon Comprehend parser to extract the text from Word files and digital PDF files. For PDF files, you can override this default and use Amazon Textract to extract the text. See Setting text extraction options.

Image files and scanned PDF files

Custom analysis supports JPEG, PNG, and TIFF images.

The following table lists the maximum file sizes for images. Scanned PDF files are subject to the same maximum sizes as native PDF files.

Description	Quota/Guideline
Image size (JPG or PNG)	1 byte–10 MB
Image size (TIFF)	1 byte–10 MB. Maximum one page.

For additional information about images, see Best practices for images.

By default, Amazon Comprehend uses the Amazon Textract `DetectDocumentText` API operation to extract the text from image files and scanned PDF files. You can override this default to use the `AnalyzeDocument` API operation instead. See Setting text extraction options.

Amazon Textract output JSON files

For custom entity recognition, but not custom classification, you can provide the output file from the Amazon Textract `AnalyzeDocument` API operation as input to analysis jobs.

Setting text extraction options

By default, Amazon Comprehend performs the following actions to extract text from a file, based on the input file type:

- **Word files** – Amazon Comprehend parser extracts the text.
- **Digital PDF files** – Amazon Comprehend parser extracts the text.
- **Image files and scanned PDF files** – Amazon Comprehend uses the Amazon Textract `DetectDocumentText` API to extract the text.

For image files and PDF files, you can use the `DocumentReaderConfig` parameter to override these default extraction actions. This parameter is available when you use the Amazon Comprehend console or API for real-time or asynchronous custom analysis.

The `DocumentReaderConfig` parameter contains three fields:

- **DocumentReadMode** – Set to `SERVICE_DEFAULT` for Amazon Comprehend to perform the default actions.

  Set to `FORCE_DOCUMENT_READ_ACTION` to use Amazon Textract to parse digital PDF files.
- **DocumentReadAction** – Sets the Amazon Textract API (DetectDocumentText or AnalyzeDocument) to use when Amazon Comprehend uses Amazon Textract for text extraction.
- **FeatureTypes** – If you set **DocumentReadAction** to use the AnalyzeDocument API operation, you can add one or both of the `FeatureTypes` (TABLES, FORMS). These features provide additional information about the tables and forms in the document. For more information about these features, see Amazon Textract Document Analysis Response Objects.

The following examples show how to configure `DocumentReaderConfig` for specific use cases:

1. Use Amazon Textract for all PDF files.
   a. **DocumentReadMode** – Set to `FORCE_DOCUMENT_READ_ACTION`.
   b. **DocumentReadAction** – Set to `TEXTRACT_DETECT_DOCUMENT_TEXT`.
   c. **FeatureTypes** – Not required.
2. Use Amazon Textract `AnalyzeDocument` API for all PDF and image files.
   a. **DocumentReadMode** – Set to `FORCE_DOCUMENT_READ_ACTION`.
   b. **DocumentReadAction** – Set to `TEXTRACT_ANALYZE_DOCUMENT`.
   c. **FeatureTypes** – Set to `TABLES`, `FORMS` or both features.
3. Use Amazon Textract `AnalyzeDocument` API for scanned PDF files and all image files.
   a. **DocumentReadMode** – Set to `SERVICE_DEFAULT`.
   b. **DocumentReadAction** – Set to `TEXTRACT_ANALYZE_DOCUMENT`.
   c. **FeatureTypes** – Set to `TABLES`, `FORMS` or both features.

For more information about the Amazon Textract options, see DocumentReaderConfig.

Best practices for images

When you use image files for custom classification or custom entity recognition, use the following guidelines to achieve the best results:

- Provide a high quality image, ideally at least 150 DPI.
- If the image file uses one of the supported formats (TIFF, JPEG, or PNG), don't convert or downsample the file before uploading it to Amazon S3.

For the best results when extracting text from tables in documents, follow these practices:

- Tables in your document are visually separated from surrounding elements on the page. For example, the table isn't overlaid onto an image or complex pattern.
- Text within the table is upright. For example, the text isn't rotated relative to other text on the page.

When extracting text from tables, you might see inconsistent results for the following cases:

- Merged table cells span multiple columns.
- Tables have cells, rows, or columns that are different than other parts of the same table.

Custom classification

Use *custom classification* to organize your documents into categories (classes) that you define. Custom classification is a two-step process. First, you train a custom classification model (also called a classifier) to recognize the classes that are of interest to you. Then you use your model to classify any number of document sets.

For example, you can categorize the content of support requests so that you can route the request to the proper support team. Or you can categorize emails received from customers to provide guidance based on the type of customer request. You can combine Amazon Comprehend with Amazon Transcribe to convert speech to text and then classify the requests coming from support phone calls.

You can run custom classification on a single document synchronously (in real time) or start an asynchronous job to classify a set of documents. You can have multiple custom classifiers in your account, each trained using different data. Custom classification supports a variety of input document types, such as plain text, PDF, Word, and images.

When you submit a classification job, you choose the classifier model to use, based on the type of documents that you need to analyze. For example, to analyze plain-text documents, you achieve the most accurate results by using a model that you trained with plain-text documents. To analyze semi-structured documents (such as PDF, Word, images, Amazon Textract output, or scanned files) , you achieve the most accurate results by using a model that you trained with native documents.

Topics

Preparing classifier training data

For custom classification, you train the model in either multi-class mode or multi-label mode. Multi-class mode associates a single class with each document. Multi-label mode associates one or more classes with each document. The input file formats are different for each mode, so choose the mode to use before you create the training data.

> **Note**
>
> The Amazon Comprehend console refers to multi-class mode as single-label mode.

Custom classification supports models that you train with plain-text documents and models that you train with native documents (such as PDF, Word, or images). For more information about classifier models and their supported document types, see Training classification models.

To prepare data to train a custom classifier model:

1. Identify the classes that you want this classifier to analyze. Decide which mode to use (multi-class or multi-label).
2. Decide on the classifier model type, based on whether the model is for analyzing plain-text documents or semi-structured documents.
3. Gather examples of documents for each of the classes. For minimum training requirements, see General quotas for document classification.
4. For a plain-text model, choose the training file format to use (CSV file or augmented manifest file). To train a native document model, you always use a CSV file.

Topics

Classifier training file formats

For a plain-text model, you can provide classifier training data as a CSV file or as an augmented manifest file that you create using SageMaker Ground Truth. The CSV file or augmented manifest file includes the text for each training document, and its associated labels.

For a native document model, you provide Classifier training data as a CSV file. The CSV file includes the file name for each training document, and its associated labels. You include the training documents in the Amazon S3 input folder for the training job.

CSV files

You provide labeled training data as UTF-8 encoded text in a CSV file. Don't include a header row. Adding a header row in your file may cause runtime errors.

For each row in the CSV file, the first column contains one or more class labels, A class label can be any valid UTF-8 string. We recommend using clear class names that don't overlap in meaning. The name can include white space, and can consist of multiple words connected by underscores or hyphens.

Do not leave any space characters before or after the commas that separate the values in a row.

The exact content of the CSV file depends on the classifier mode and the type of training data. For details, see the sections on Multi-class mode and Multi-label mode.

Augmented manifest file

An augmented manifest file is a labeled dataset that you create using SageMaker Ground Truth. Ground Truth is a data labeling service that helps you—or a workforce that you employ—to build training datasets for machine learning models.

For more information about Ground Truth and the output that it produces, see Use SageMaker Ground Truth to Label Data in the *Amazon SageMaker Developer Guide*.

Augmented manifest files are in JSON lines format. In these files, each line is a complete JSON object that contains a training document and its associated labels. The exact content of each line depends on the classifier mode. For details, see the sections on Multi-class mode and Multi-label mode.

When you provide your training data to Amazon Comprehend, you specify one or more label attribute names. How many attribute names you specify depends on whether your augmented manifest file is the output of a single labeling job or a chained labeling job.

If your file is the output of a single labeling job, specify the single label attribute name from the Ground Truth job.

If your file is the output of a chained labeling job, specify the label attribute name for one or more jobs in the chain. Each label attribute name provides the annotations from an individual job. You can specify up to 5 of these attributes for augmented manifest files from chained labeling jobs.

For more information about chained labeling jobs, and for examples of the output that they produce, see Chaining Labeling Jobs in the Amazon SageMaker Developer Guide.

Multi-class mode

In multi-class mode, classification assigns one class for each document. The individual classes are mutually exclusive. For example, you can classify a movie as comedy or science fiction, but not both.

> **Note**
>
> The Amazon Comprehend console refers to multi-class mode as single-label mode.

Topics

- Plain-text models
- Native document models

Plain-text models

To train a plain-text model, you can provide labeled training data as a CSV file or as an augmented manifest file from SageMaker Ground Truth.

CSV file

For general information about using CSV files for training classifiers, see CSV files.

Provide the training data as a two-column CSV file. For each row, the first column contains the class label value. The second column contains an example text document for that class. Each row must end with \n or \r\n characters.

The following example shows a CSV file containing three documents.

```
CLASS,Text of document 1
CLASS,Text of document 2
CLASS,Text of document 3
```

The following example shows one row of a CSV file that trains a custom classifier to detect whether an email message is spam:

```
SPAM,"Paulo, your $1000 award is waiting for you! Claim it while you still can at
 http://example.com."
```

Augmented manifest file

For general information about using augmented manifest files for training classifiers, see Augmented manifest file.

For plain-text documents, each line of the augmented manifest file is a complete JSON object that contains a training document, a single class name, and other metadata from Ground Truth. The following example is an augmented manifest file for training a custom classifier to recognize spam email messages:

```
{"source":"Document 1 text", "MultiClassJob":0, "MultiClassJob-metadata":
{"confidence":0.62, "job-name":"labeling-job/multiclassjob", "class-name":"not_spam",
 "human-annotated":"yes", "creation-date":"2020-05-21T17:36:45.814354",
 "type":"groundtruth/text-classification"}}
{"source":"Document 2 text", "MultiClassJob":1, "MultiClassJob-metadata":
{"confidence":0.81, "job-name":"labeling-job/multiclassjob", "class-name":"spam",
 "human-annotated":"yes", "creation-date":"2020-05-21T17:37:51.970530",
 "type":"groundtruth/text-classification"}}
{"source":"Document 3 text", "MultiClassJob":1, "MultiClassJob-metadata":
{"confidence":0.81, "job-name":"labeling-job/multiclassjob", "class-name":"spam",
 "human-annotated":"yes", "creation-date":"2020-05-21T17:37:51.970566",
 "type":"groundtruth/text-classification"}}
```

The following example shows one JSON object from the augmented manifest file, formatted for readability:

```
{
   "source": "Paulo, your $1000 award is waiting for you! Claim it while you still can
 at http://example.com.",
   "MultiClassJob": 0,
   "MultiClassJob-metadata": {
       "confidence": 0.98,
       "job-name": "labeling-job/multiclassjob",
       "class-name": "spam",
       "human-annotated": "yes",
       "creation-date": "2020-05-21T17:36:45.814354",
       "type": "groundtruth/text-classification"
   }
```

```
}
```

In this example, the `source` attribute provides the text of the training document, and the `MultiClassJob` attribute assigns the index of a class from a classification list. The `job-name` attribute is the name that you defined for the labeling job in Ground Truth.

When you start the classifier training job in Amazon Comprehend, you specify the same labeling job name.

Native document models

A native document model is a model that you train with native documents (such as PDF, DOCX , and images). You provide the training data as a CSV file.

CSV file

For general information about using CSV files for training classifiers, see CSV files.

Provide the training data as a three-column CSV file. For each row, the first column contains the class label value. The second column contains the file name of an example document for this class. The third column contains the page number. The page number is optional if the example document is an image.

The following example shows a CSV file that references three input documents.

```
CLASS,input-doc-1.pdf,3
CLASS,input-doc-2.docx,1
CLASS,input-doc-3.png
```

The following example shows one row of a CSV file that trains a custom classifier to detect whether an email message is spam. Page 2 of the PDF file contains the spam example.

```
SPAM,email-content-3.pdf,2
```

Multi-label mode

In multi-label mode, individual classes represent different categories that aren't mutually exclusive. Multi-label classification assigns one or more classes to each document. For example, you can classify one movie as Documentary, and another movie as Science fiction, Action, and Comedy.

For training, multi-label mode supports up to 1 million examples containing up to 100 unique classes.

Topics

Plain-text models

To train a plain-text model, you can provide labeled training data as a CSV file or as an augmented manifest file from SageMaker Ground Truth.

CSV file

For general information about using CSV files for training classifiers, see CSV files.

Provide the training data as a two-column CSV file. For each row, the first column contains the class label values, and the second column contains an example text document for these classes. To enter more than one class in the first column, use a delimiter (such as a |) between each class.

```
CLASS,Text of document 1
CLASS,Text of document 2
CLASS|CLASS|CLASS,Text of document 3
```

The following example shows one row of a CSV file that trains a custom classifier to detect genres in movie abstracts:

```
COMEDY|MYSTERY|SCIENCE_FICTION|TEEN,"A band of misfit teens become unlikely detectives
 when they discover troubling clues about their high school English teacher. Could the
 strange Mrs. Doe be an alien from outer space?"
```

The default delimiter between class names is a pipe (|). However, you can use a different character as a delimiter. The delimiter must be distinct from all characters in your class names. For example, if your classes are CLASS_1, CLASS_2, and CLASS_3, the underscore (_) is part of the class name. So don't use an underscore as the delimiter for separating class names.

Augmented manifest file

For general information about using augmented manifest files for training classifiers, see Augmented manifest file.

For plain-text documents, each line of the augmented manifest file is a complete JSON object. It contains a training document, class names, and other metadata from Ground Truth. The following example is an augmented manifest file for training a custom classifier to detect genres in movie abstracts:

```
{"source":"Document 1 text", "MultiLabelJob":[0,4], "MultiLabelJob-metadata":{"job-
name":"labeling-job/multilabeljob", "class-map":{"0":"action", "4":"drama"}, "human-
annotated":"yes", "creation-date":"2020-05-21T19:02:21.521882", "confidence-map":
{"0":0.66}, "type":"groundtruth/text-classification-multilabel"}}
{"source":"Document 2 text", "MultiLabelJob":[3,6], "MultiLabelJob-metadata":{"job-
name":"labeling-job/multilabeljob", "class-map":{"3":"comedy", "6":"horror"}, "human-
annotated":"yes", "creation-date":"2020-05-21T19:00:01.291202", "confidence-map":
{"1":0.61,"0":0.61}, "type":"groundtruth/text-classification-multilabel"}}
{"source":"Document 3 text", "MultiLabelJob":[1], "MultiLabelJob-metadata":
{"job-name":"labeling-job/multilabeljob", "class-map":{"1":"action"}, "human-
annotated":"yes", "creation-date":"2020-05-21T18:58:51.662050", "confidence-map":
{"1":0.68}, "type":"groundtruth/text-classification-multilabel"}}
```

The following example shows one JSON object from the augmented manifest file, formatted for readability:

```
{
      "source": "A band of misfit teens become unlikely detectives when
                  they discover troubling clues about their high school English
 teacher.
                    Could the strange Mrs. Doe be an alien from outer space?",
      "MultiLabelJob": [
          3,
          8,
          10,
          11
      ],
      "MultiLabelJob-metadata": {
          "job-name": "labeling-job/multilabeljob",
          "class-map": {
              "3": "comedy",
              "8": "mystery",
              "10": "science_fiction",
              "11": "teen"
          },
          "human-annotated": "yes",
          "creation-date": "2020-05-21T19:00:01.291202",
```

```
        "confidence-map": {
            "3": 0.95,
            "8": 0.77,
            "10": 0.83,
            "11": 0.92
        },
        "type": "groundtruth/text-classification-multilabel"
    }
}
```

In this example, the `source` attribute provides the text of the training document, and the `MultiLabelJob` attribute assigns the indexes of several classes from a classification list. The job-name in the `MultiLabelJob` metadata is the name that you defined for the labeling job in Ground Truth.

Native document models

A native document model is a model that you train with native documents (such as PDF, DOCX , and image files). You provide labeled training data as a CSV file.

CSV file

For general information about using CSV files for training classifiers, see CSV files.

Provide the training data as a three-column CSV file. For each row, the first column contains the class label values. The second column contains the file name of an example document for these classes. The third column contains the page number. The page number is optional if the example document is an image.

To enter more than one class in the first column, use a delimiter (such as a |) between each class.

```
CLASS,input-doc-1.pdf,3
CLASS,input-doc-2.docx,1
CLASS|CLASS|CLASS,input-doc-3.png,2
```

The following example shows one row of a CSV file that trains a custom classifier to detect genres in movie abstracts. Page 2 of the PDF file contains the example of a comedy/teen movie.

```
COMEDY|TEEN,movie-summary-1.pdf,2
```

The default delimiter between class names is a pipe (|). However, you can use a different character as a delimiter. The delimiter must be distinct from all characters in your class names. For example, if your classes are CLASS_1, CLASS_2, and CLASS_3, the underscore (_) is part of the class name. So don't use an underscore as the delimiter for separating class names.

Training classification models

To train a model for custom classification, you define the categories and provide example documents to train the custom model. You train the model in either multi-class or multi-label mode. Multi-class mode associates a single class with each document. Multi-label mode associates one or more classes with each document.

Custom classification supports two types of classifier models: plain-text models and native document models. A plain-text model classifies documents based on their text content. A native document model also classifies documents based on text content. A native document model can also use additional signals, such as from the layout of the document. You train a native document model with native documents for the model to learn the layout information.

Plain-text models have the following characteristics:

- You train the model using UTF-8 encoded text documents.
- You can train the model using documents in one of following languages: English, Spanish, German, Italian, French, or Portuguese.
- The training documents for a given classifier must all use the same language.
- Training documents are plain text, so there are no additional charges for text extraction.

Native document models have the following characteristics:

- You train the model using semi-structured documents, which includes the following document types:
  - Digital and scanned PDF documents.
  - Word documents (DOCX).
  - Images: JPG files, PNG files, and single-page TIFF files.
  - Textract API output JSON files.
- You train the model using English documents.

- If your training documents include scanned document files, you incur additional charges for text extraction. See the Amazon Comprehend Pricing page for details.

You can classify any of the supported document types using either type of model. However, for the most accurate results, we recommend using a plain-text model to classify plain-text documents and a native document model to classify semi-structured documents.

Topics

Train custom classifiers (console)

You can create and train a custom classifier using the console, and then use the custom classifier to analyze your documents.

To train a custom classifier, you need a set of training documents. You label these documents with the categories that you want the document classifier to recognize. For information about preparing your training documents, see Preparing classifier training data.

To create and train a document classifier model

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Customization** and then choose **Custom Classification**.
3. Choose **Create new model**.
4. Under **Model settings**, enter a model name for the classifier. The name must be unique within your account and current Region.

   (Optional) Enter a version name. The name must be unique within your account and current Region.

5. Select the language of the training documents. To see the languages that classifiers support, see Training classification models.
6. (Optional) If you want to encrypt the data in the storage volume while Amazon Comprehend processes your training job, choose **Classifier encryption**. Then choose whether to use a KMS key associated with your current account, or one from another account.

   - If you are using a key associated with the current account, choose the key ID for **KMS key ID**.
   - If you are using a key associated with a different account, enter the ARN for the key ID under **KMS key ARN**.

   > **Note**
   >
   > For more information on creating and using KMS keys and the associated encryption, see AWS Key Management Service (AWS KMS).

7. Under **Data specifications**, choose the **Training model type** to use.

   - **Plain text documents:** Choose this option to create a plain text model. Train the model using plain text documents.
   - **Native documents:** Choose this option to create a native document model. Train the model using native documents (PDF, Word, images).

8. Choose the **Data format** of your training data. For information about the data formats, see Classifier training file formats.

   - **CSV file:** Choose this option if your training data uses the CSV file format.
   - **Augmented manifest:** Choose this option if you used Ground Truth to create augmented manifest files for your training data. This format is available if you chose **Plain text documents** as the training model type.

9. Choose the **Classifier mode** to use.

   - **Single-label mode:** Choose this mode if the categories you're assigning to documents are mutually exclusive and you're training your classifier to assign one label to each document. In the Amazon Comprehend API, single-label mode is known as multi-class mode.
   - **Multi-label mode:** Choose this mode if multiple categories can be applied to a document at the same time, and you are training your classifier to assign one or more labels to each document.

10. If you choose **Multi-label mode**, you can select the **Delimiter for labels**. Use this delimiter character to separate labels when there are multiple classes for a training document. The default delimiter is the pipe character.
11. (Optional) If you chose **Augmented manifest** as the data format, you can input up to five augmented manifest files. Each augmented manifest file contains either a training dataset or a test dataset. You must provide at least one training dataset. Test datasets are optional. Use the following steps to configure the augmented manifest files:

    a. Under **Training and test dataset**, expand the **Input location** panel.

    b. In **Dataset type**, choose **Training data** or **Test data**.

    c. For the **SageMaker Ground Truth augmented manifest file S3 location**, enter the location of the Amazon S3 bucket that contains the manifest file or navigate to it by choosing **Browse S3**. The IAM role that you're using for access permissions for the training job must have read permissions for the S3 bucket.

    d. For the **Attribute names**, enter the name of the attribute that contains your annotations. If the file contains annotations from multiple chained labeling jobs, add an attribute for each job.

    e. To add another input location, choose **Add input location** and then configure the next location.
12. (Optional) If you chose **CSV file** as the data format, use the following steps to configure the training dataset and optional test dataset:

    a. Under **Training dataset**, enter the location of the Amazon S3 bucket that contains your training data CSV file or navigate to it by choosing **Browse S3**. The IAM role that you're using for access permissions for the training job must have read permissions for the S3 bucket.

    (Optional) If you chose **Native documents** as the training model type, you also provide the URL of the Amazon S3 folder that contains the training example files.

    b. Under **Test dataset**, select whether you are providing extra data for Amazon Comprehend to test the trained model.

    - **Autosplit**: Autosplit automatically selects 10% of your training data to reserve for use as testing data.
    - (Optional) **Customer provided**: Enter the URL of the test data CSV file in Amazon S3. You can also navigate to its location in Amazon S3 and choose **Select folder**.

(Optional) If you chose **Native documents** as the training model type, you also provide the URL of the Amazon S3 folder that contains the test files.

13. (Optional) For **Document read mode**, you can override the default text extraction actions. This option isn't required for plain-text models, as it applies to text extraction for scanned documents. For more information, see Setting text extraction options.
14. (Optional for plain-text models) For **Output data**, enter the location of an Amazon S3 bucket to save training output data, such as the confusion matrix. For more information, see Confusion matrix.

    (Optional) If you choose to encrypt the output result from your training job, choose **Encryption**. Then choose whether to use a KMS key associated with the current account, or one from another account.

    - If you are using a key associated with the current account, choose the key alias for **KMS key ID**.
    - If you are using a key associated with a different account, enter the ARN for the key alias or ID under **KMS key ID**.
15. For **IAM role**, choose **Choose an existing IAM role**, and then choose an existing IAM role that has read permissions for the S3 bucket that contains your training documents. The role must have a trust policy that begins with `comprehend.amazonaws.com` to be valid.

    If you don't already have an IAM role with these permissions, choose **Create an IAM role** to make one. Choose the access permissions to grant this role, and then choose a name suffix to distinguish the role from IAM roles in your account.

    > **Note**
    >
    > For encrypted input documents, the IAM role used must also have `kms:Decrypt` permission. For more information, see Permissions required to use KMS encryption.

16. (Optional) To launch your resources into Amazon Comprehend from a VPC, enter the VPC ID under **VPC** or choose the ID from the dropdown list.

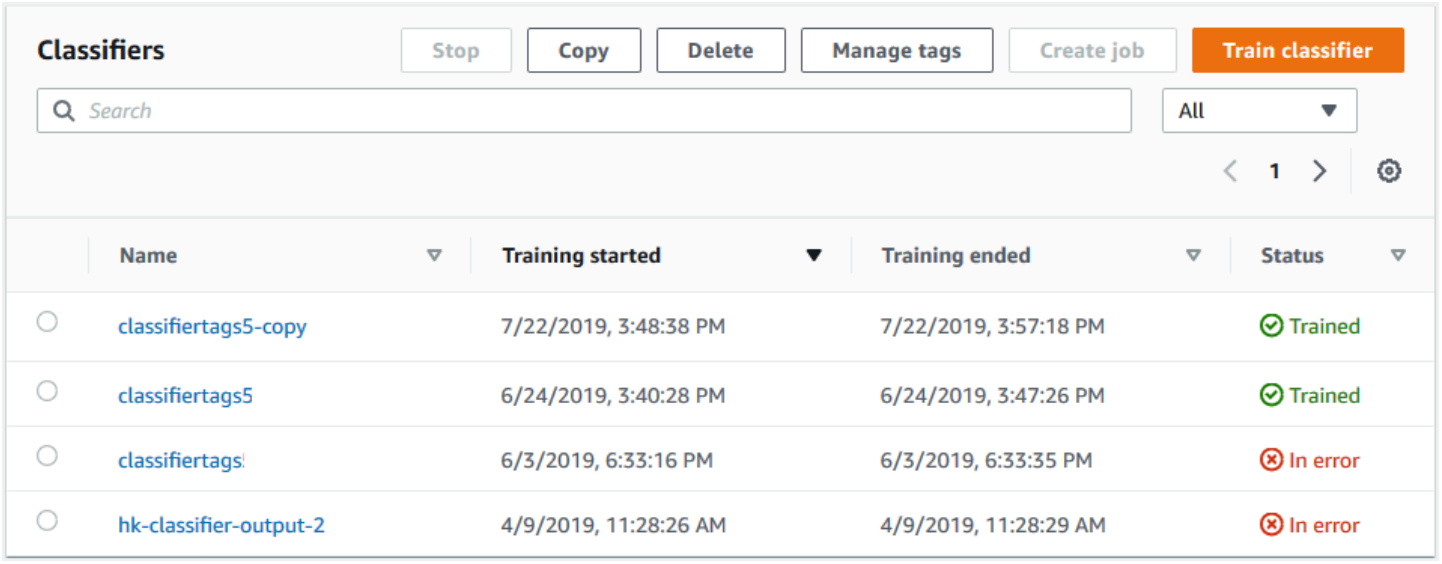
    1. Choose the subnet under **Subnets(s)**. After you select the first subnet, you can choose additional ones.
    2. Under **Security Group(s)**, choose the security group to use if you specified one. After you select the first security group, you can choose additional ones.

> **Note**
>
> When you use a VPC with your classification job, the `DataAccessRole` used for the Create and Start operations must have permissions to the VPC that accesses the input documents and the output bucket.

17. (Optional) To add a tag to the custom classifier, enter a key-value pair under **Tags**. Choose **Add tag**. To remove this pair before creating the classifier, choose **Remove tag**. For more information, see Tagging your resources.
18. Choose **Create**.

The console displays the **Classifiers** page. The new classifier appears in the table, showing `Submitted` as its status. When the classifier starts processing the training documents, the status changes to `Training`. When a classifier is ready to use, the status changes to `Trained` or `Trained with warnings`. If the status is `TRAINED_WITH_WARNINGS`, review the skipped files folder in the Classifier training output.

If Amazon Comprehend encountered errors during creation or training, the status changes to `In error`. You can choose a classifier job in the table to get more information about the classifier, including any error messages.



Train custom classifiers (API)

To create and train a custom classifier, use the CreateDocumentClassifier operation.

You can monitor the progress of the request using the DescribeDocumentClassifier operation. After the `Status` field transitions to `TRAINED`, you can use the classifier to classify documents. If the status is `TRAINED_WITH_WARNINGS`, review the skipped files folder in the Classifier training output from the `CreateDocumentClassifier` operation.

Topics

- Training custom classification using the AWS Command Line Interface
- Using the AWS SDK for Java or SDK for Python

Training custom classification using the AWS Command Line Interface

The following examples show how to use the `CreateDocumentClassifier` operation, the `DescribeDocumentClassificationJob` operation, and other custom classifier APIs with the AWS CLI.

The examples are formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Create a plain-text custom classifier using the `create-document-classifier` operation.

```
aws comprehend create-document-classifier \
     --region region \
     --document-classifier-name testDelete \
     --language-code en \
     --input-data-config S3Uri=s3://S3Bucket/docclass/file name \
     --data-access-role-arn arn:aws:iam::account number:role/testFlywheelDataAccess
```

To create a native custom classifier, provide the following additional parameters in the `create-document-classifier` request.

1. DocumentType: set the value to SEMI_STRUCTURED_DOCUMENT.
2. Documents: the S3 location for the training documents (and, optionally, the test documents).
3. OutputDataConfig: provide the S3 location for the output documents (and an optional KMS key).
4. DocumentReaderConfig: Optional field for text extraction settings.

```
aws comprehend create-document-classifier \
     --region region \
```

```
    --document-classifier-name testDelete \
    --language-code en \
    --input-data-config
         S3Uri=s3://S3Bucket/docclass/file name \
          DocumentType \
            Documents  \
    --output-data-config S3Uri=s3://S3Bucket/docclass/file name \
    --data-access-role-arn arn:aws:iam::account number:role/testFlywheelDataAccess
```

Get information on a custom classifier with the document classifier ARN using the `DescribeDocumentClassifier` operation.

```
aws comprehend describe-document-classifier \
    --region region \
    --document-classifier-arn arn:aws:comprehend:region:account number:document-
classifier/file name
```

Delete a custom classifier using the `DeleteDocumentClassifier` operation.

```
aws comprehend delete-document-classifier \
    --region region \
    --document-classifier-arn arn:aws:comprehend:region:account number:document-
classifier/testDelete
```

List all custom classifiers in the account using the `ListDocumentClassifiers` operation.

```
aws comprehend list-document-classifiers
    --region region
```

Using the AWS SDK for Java or SDK for Python

For SDK examples of how to create and train a custom classifier , see Use `CreateDocumentClassifier` with an AWS SDK or CLI.

Test the training data

After training the model, Amazon Comprehend tests the custom classifier model. If you don't provide a test dataset, Amazon Comprehend trains the model with 90 percent of the training data. It reserves 10 percent of the training data to use for testing. If you do provide a test dataset, the test data must include at least one example for each unique label in the training dataset.

Testing the model provides you with metrics that you can use to estimate the accuracy of the model. The console displays the metrics in the **Classifier performance** section of the **Classifier details** page in the console. They're also returned in the `Metrics` fields returned by the DescribeDocumentClassifier operation.

In the following example training data, there are five labels, DOCUMENTARY, DOCUMENTARY, SCIENCE_FICTION, DOCUMENTARY, ROMANTIC_COMEDY. There are three unique classes: DOCUMENTARY, SCIENCE_FICTION, ROMANTIC_COMEDY.

Column 1	Column 2
DOCUMENTARY	document text 1
DOCUMENTARY	document text 2
SCIENCE_FICTION	document text 3
DOCUMENTARY	document text 4
ROMANTIC_COMEDY	document text 5

For auto split (where Amazon Comprehend reserves 10 percent of the training data to use for testing), if the training data contains limited examples of a specific label, the test dataset may contain zero examples of that label. For instance, if the training dataset contains 1000 instances of the DOCUMENTARY class, 900 instances of SCIENCE_FICTION, and a single instance of the ROMANTIC_COMEDY class, the test dataset might contain 100 DOCUMENTARY and 90 SCIENCE_FICTION instances, but no ROMANTIC_COMEDY instances, as there is a single example available.

After you finish training your model, the training metrics provide information that you can use to decide if the model is sufficiently accurate for your needs.

Classifier training output

After Amazon Comprehend completes the custom classifier model training, it creates output files in the Amazon S3 output location that you specified in the CreateDocumentClassifier API request or the equivalent console request.

Amazon Comprehend creates a confusion matrix when you train a plain-text model or a native document model. It can create additional output files when you train a native document model.

Topics

Confusion matrix

When you train a custom classifier model, Amazon Comprehend creates a confusion matrix that provides metrics on how well the model performed in training. This matrix shows a matrix of labels that the model predicted, compared to the actual document labels. Amazon Comprehend uses a portion of the training data to create the confusion matrix.

A confusion matrix provides an indication of which classes could use more data to improve model performance. A class with a high fraction of correct predictions has the highest number of results along the diagonal of the matrix. If the number on the diagonal is a lower number, the class has a lower fraction of correct predictions. You can add more training examples for this class and train the model again. For example, if 40 percent of label A samples get classified as label D, adding more samples for label A and label D enhances the classifier's performance.

After Amazon Comprehend creates the classifier model, the confusion matrix is available in the `confusion_matrix.json` file in the S3 output location.

The format of the confusion matrix varies, depending on whether you trained your classifier using multi-class mode or multi-label mode.

Topics

Confusion matrix for multi-class mode

In multi-class mode, the individual classes are mutually exclusive, so classification assigns one label to each document. For example, an animal can be a dog or a cat, but not both at the same time.

Consider the following example of a confusion matrix for a multi-class trained classifier:

```
  A  B  X  Y <-(predicted label)
```

```
A 1 2 0 4
B 0 3 0 1
X 0 0 1 0
Y 1 1 1 1
^
|
(actual label)
```

In this case, the model predicted the following:

- One "A" label was accurately predicted, two "A" labels were incorrectly predicted as "B" labels, and four "A" labels were incorrectly predicted as "Y" labels.
- Three "B" labels were accurately predicted, and one "B" label was incorrectly predicted as a "Y" label.
- One "X" was accurately predicted.
- One "Y" label was accurately predicted, one was incorrectly predicted as an "A" label, one was incorrectly predicted as a "B" label, and one was incorrectly predicted as an "X" label.

The diagonal line in the matrix (A:A, B:B, X:X, and Y:Y) shows the accurate predictions. The prediction errors are the values outside of the diagonal. In this case, the matrix shows the following prediction error rates:

- A labels: 86%
- B labels: 25%
- X labels: 0%
- Y labels: 75%

The classifier returns the confusion matrix as a file in JSON format. The following JSON file represents the matrix for the previous example.

```
{
 "type": "multi_class",
 "confusion_matrix": [
 [1, 2, 0,4],
 [0, 3, 0, 1],
 [0, 0, 1, 0],
 [1, 1, 1, 1]],
 "labels": ["A", "B", "X", "Y"],
```

```
 "all_labels": ["A", "B", "X", "Y"]
}
```

Confusion matrix for multi-label mode

In multi-label mode, classification can assign one or more classes to a document. Consider the following example of a confusion matrix for a multi-class trained classifier.

In this example, there are three possible labels: `Comedy`, `Action`, and `Drama`. The multi-label confusion matrix creates one 2x2 matrix for each label.

```
Comedy                  Action                  Drama
     No Yes                  No Yes                  No Yes   <-(predicted label)

 No  2   1               No  1   1               No  3   0

Yes  0   2              Yes  2   1              Yes  1   1
 ^                       ^                       ^
 |                       |                       |
 |-----------(was this label actually used)--------|
```

In this case, the model returned the following for the `Comedy` label:

- Two instances where a `Comedy` label was accurately predicted to be present. True positive (TP).
- Two instances where a `Comedy` label was accurately predicted to be absent. True negative (TN).
- Zero instances where a `Comedy` label was incorrectly predicted to be present. False positive (FP).
- One instance where a `Comedy` label was incorrectly predicted to be absent. False negative (FN).

As with a multi-class confusion matrix, the diagonal line in each matrix shows the accurate predictions.

In this case, the model accurately predicted `Comedy` labels 80% of the time (TP plus TN) and incorrectly predicted them 20% of the time (FP plus FN).

The classifier returns the confusion matrix as a file in JSON format. The following JSON file represents the matrix for the previous example.

```
{
"type": "multi_label",
```

```
"confusion_matrix": [
 [[2, 1],
 [0, 2]],
 [[1, 1],
 [2, 1]],
 [[3, 0],
 [1, 1]]
],
"labels": ["Comedy", "Action", "Drama"]
"all_labels": ["Comedy", "Action", "Drama"]
}
```

Additional outputs for native document models

Amazon Comprehend can create additional output files when you train a native document model.

Amazon Textract output

If Amazon Comprehend invoked the Amazon Textract APIs to extract text for any of the training documents, it saves the Amazon Textract output files in the S3 output location. It uses the following directory structure:

- **Training documents:**

  `amazon-textract-output/train/<file_name>/<page_num>/textract_output.json`
- **Test documents:**

  `amazon-textract-output/test/<file_name>/<page_num>/textract_output.json`

Amazon Comprehend populates the test folder if you provided test documents in the API request.

Document annotation failures

Amazon Comprehend creates the following files in the Amazon S3 output location (in the **skipped_documents/** folder) if there are any failed annotations:

- failed_annotations_train.jsonl

  File exists if any annotations failed in the training data.
- failed_annotations_test.jsonl

  File exists if the request included test data and any annotations failed in the test data.

The failed annotation files are JSONL files with the following format:

```
{
     "File": "String", "Page": Number, "ErrorCode": "...", "ErrorMessage": "..."}
    {"File": "String", "Page": Number, "ErrorCode": "...", "ErrorMessage": "..."
  }
```

Custom classifier metrics

Amazon Comprehend provides metrics to help you estimate how well a custom classifier performs. Amazon Comprehend calculates the metrics using the test data from the classifier training job. The metrics accurately represent the performance of the model during training, so they approximate the model performance for classification of similar data.

Use API operations such as DescribeDocumentClassifier to retrieve the metrics for a custom classifier.

> **Note**
>
> Refer to Metrics: Precision, recall, and FScore for an understanding of the underlying Precision, Recall, and F1 score metrics. These metrics are defined at a class level. Amazon Comprehend uses **macro** averaging to combine these metrics into the test set P,R, and F1, as discussed in the following.

Topics

- Metrics
- Improving your custom classifier's performance

Metrics

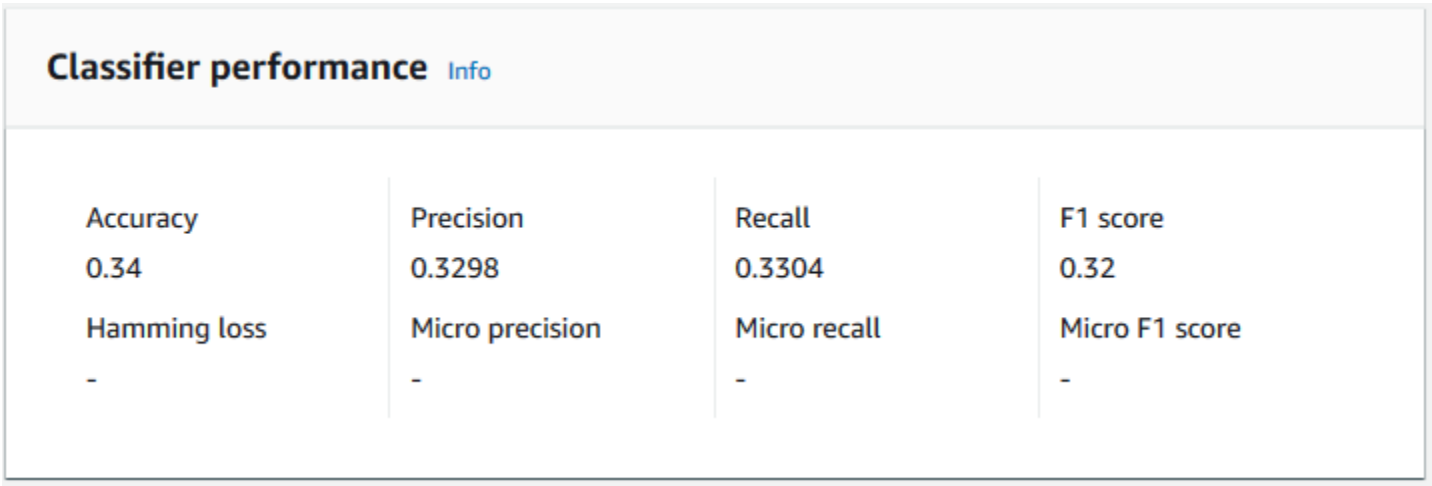
Amazon Comprehend supports the following metrics:

Topics

- Accuracy
- Precision (macro precision)
- Recall (macro recall)
- F1 score (macro F1 score)

- Hamming loss
- Micro precision
- Micro recall
- Micro F1 score

To view the metrics for a Classifier, open the **Classifier Details** page in the console.



Accuracy

Accuracy indicates the percentage of labels from the test data that the model predicted accurately. To compute accuracy, divide the number of accurately predicted labels in the test documents by the total number of labels in the test documents.

For example

Actual label	Predicted label	Accurate/Incorrect
1	1	Accurate
0	1	Incorrect
2	3	Incorrect
3	3	Accurate
2	2	Accurate
1	1	Accurate

Actual label	Predicted label	Accurate/Incorrect
3	3	Accurate

The accuracy consists of the number of accurate predictions divided by the number of overall test samples = 5/7 = 0.714, or 71.4%

Precision (macro precision)

Precision is a measure of the usefulness of the classifier results in the test data. It's defined as the number of documents accurately classified, divided by the total number of classifications for the class. High precision means that the classifier returned significantly more relevant results than irrelevant ones.

The `Precision` metric is also known as *Macro Precision*.

The following example shows precision results for a test set.

Label	Sample size	Label precision
Label_1	400	0.75
Label_2	300	0.80
Label_3	30000	0.90
Label_4	20	0.50
Label_5	10	0.40

The Precision (Macro Precision) metric for the model is therefore:

```
Macro Precision = (0.75 + 0.80 + 0.90 + 0.50 + 0.40)/5 = 0.67
```

Recall (macro recall)

This indicates the percentage of correct categories in your text that the model can predict. This metric comes from averaging the recall scores of all available labels. Recall is a measure of how complete the classifier results are for the test data.

High recall means that the classifier returned most of the relevant results.

The `Recall` metric is also known as *Macro Recall*.

The following example shows recall results for a test set.

Label	Sample size	Label recall
Label_1	400	0.70
Label_2	300	0.70
Label_3	30000	0.98
Label_4	20	0.80
Label_5	10	0.10

The Recall (Macro Recall) metric for the model is therefore:

```
Macro Recall = (0.70 + 0.70 + 0.98 + 0.80 + 0.10)/5 = 0.656
```

F1 score (macro F1 score)

The F1 score is derived from the `Precision` and `Recall` values. It measures the overall accuracy of the classifier. The highest score is 1, and the lowest score is 0.

Amazon Comprehend calculates the *Macro F1 Score*. It's the unweighted average of the label F1 scores. Using the following test set as an example:

Label	Sample size	Label F1 score
Label_1	400	0.724
Label_2	300	0.824
Label_3	30000	0.94
Label_4	20	0.62

Label	Sample size	Label F1 score
Label_5	10	0.16

The F1 Score (Macro F1 Score) for the model is calculated as follows:

```
Macro F1 Score = (0.724 + 0.824 + 0.94 + 0.62 + 0.16)/5 = 0.6536
```

Hamming loss

The fraction of labels that are incorrectly predicted. Also seen as the fraction of incorrect labels compared to the total number of labels. Scores closer to zero are better.

Micro precision

Original:

Similar to the precision metric, except that micro precision is based on the overall score of all precision scores added together.

Micro recall

Similar to the recall metric, except that micro recall is based on the overall score of all recall scores added together.

Micro F1 score

The Micro F1 score is a combination of the Micro Precision and Micro Recall metrics.

Improving your custom classifier's performance

The metrics provide an insight into how your custom classifier performs during a classification job. If the metrics are low, the classification model might not be effective for your use case. You have several options to improve your classifier performance:

1. In your training data, provide concrete examples that define clear separation of the categories. For example, provide documents that use unique words/sentences to represent the category.
2. Add more data for under-represented labels in your training data.
3. Try to reduce skew in the categories. If the largest label in your data has more than 10 times the documents in the smallest label, try increasing the number of documents for the smallest

label. Make sure to reduce the skew ratio to at most 10:1 between highly represented and least represented classes. You can also try removing input documents from the highly represented classes.

Running real-time analysis

After you train a custom classifier, you can classify documents using real-time analysis. Real-time analysis takes a single document as input and returns the results synchronously. Custom classification accepts a variety of document types as inputs for real-time analysis. For details, see Inputs for real-time custom analysis.

If you plan to analyze image files or scanned PDF documents, your IAM policy must grant permissions to use two Amazon Textract API methods (DetectDocumentText and AnalyzeDocument). Amazon Comprehend invokes these methods during text extraction. For an example policy, see Permissions required to perform document analysis actions.

You must create an endpoint to run real-time analysis using a custom classification model.

Topics

- Real-time analysis for custom classification (console)
- Real-time analysis for custom classification (API)
- Outputs for real-time analysis

Real-time analysis for custom classification (console)

You can use the Amazon Comprehend console to run real-time analysis using a custom classification model.

You create an endpoint to run the real-time analysis. An endpoint includes managed resources that makes your custom model available for real-time inference.

For information about provisioning endpoint throughput, and the associated costs, see Using Amazon Comprehend endpoints.

Topics

- Creating an endpoint for custom classification
- Running real-time custom classification

Creating an endpoint for custom classification

To create an endpoint (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Endpoints** and choose the **Create endpoint** button. A **Create endpoint** screen opens.
3. Give the endpoint a name. The name must be unique within the current Region and account.
4. Choose a custom model that you want to attach the new endpoint to. From the dropdown, you can search by model name.

   > **Note**
   >
   > You must create a model before you can attach an endpoint to it. If you don't have a model yet, see Training classification models.

5. (Optional) to add a tag to the endpoint, enter a key-value pair under **Tags** and choose **Add tag**. To remove this pair before creating the endpoint, choose **Remove tag**
6. Enter the number of inference units (IUs) to assign to the endpoint. Each unit represents a throughput of 100 characters per second for up to two documents per second. For information about endpoint throughput, see Using Amazon Comprehend endpoints.
7. (Optional) If you are creating a new endpoint, you have the option to use the IU estimator. Depending on the throughput, or the number of characters you want to analyze per second, it can be hard to know how many inference units you need. This optional step can help you determine how the number of IUs to request.
8. From the **Purchase summary**, review your estimated hourly, daily, and monthly endpoint cost.
9. Select the check box if you understand that your account incurs charges for the endpoint from the time it starts until you delete it.
10. Choose **Create endpoint**

Running real-time custom classification

Once you've created an endpoint, you can run real-time analysis using your custom model. There are two ways to run real-time analysis from the console. You can input text or upload a file, as shown in the following.

To run real-time analysis using a custom model (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Real-time analysis**.
3. Under **Input type**, choose **Custom** for **Analysis type**.
4. Under **Custom model type**, choose **Custom classification**.
5. For **Endpoint**, choose the endpoint that you want to use. This endpoint links to a specific custom model.
6. To specify the input data for analysis, you can input text or upload a file.
   - To enter text:
     a. Choose **Input text**.
     b. Enter the text that you want to analyze.
   - To upload a file:
     a. Choose **Upload file** and enter the file name to upload.
     b. (Optional) Under **Advanced read actions**, you can override the default actions for text extraction. For details, see Setting text extraction options

   For best results, match the type of input to the classifier model type. The console displays a warning if you submit a native document to a plain-text model, or plain text to a native document model. For more information, see Training classification models.
7. Choose **Analyze**. Amazon Comprehend analyzes the input data using your custom model. Amazon Comprehend displays the discovered classes, along with a confidence assessment for each class.

Real-time analysis for custom classification (API)

You can use the Amazon Comprehend API to run real-time classification with a custom model. First, you create an endpoint to run the real-time analysis. After you create the endpoint, you run the real-time classification.

The examples in this section use command formats for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

For information about provisioning endpoint throughput, and the associated costs, see Using Amazon Comprehend endpoints.

Topics

- Creating an endpoint for custom classification
- Running real-time custom classification

Creating an endpoint for custom classification

The following example shows the CreateEndpoint API operation using the AWS CLI.

```
aws comprehend create-endpoint \
    --desired-inference-units number of inference units \
    --endpoint-name endpoint name \
    --model-arn arn:aws:comprehend:region:account-id:model/example \
    --tags Key=My1stTag,Value=Value1
```

Amazon Comprehend responds with the following:

```
{
   "EndpointArn": "Arn"
}
```

Running real-time custom classification

After you create an endpoint for your custom classification model, you use the endpoint to run the ClassifyDocument API operation. You can provide text input using the `text` or `bytes` parameter. Enter the other input types using the `bytes` parameter.

For image files and PDF files, you can use the `DocumentReaderConfig` parameter to override the default text extraction actions. For details, see Setting text extraction options

For best results, match the type of input to the classifier model type. The API response includes a warning if you submit a native document to a plain-text model, or a plain-text file to a native document model. For more information, see Training classification models.

Using the AWS Command Line Interface

The following examples demonstrate how to use the *classify-document* CLI command.

Classify text using the AWS CLI

The following example runs real-time classification on a block of text.

```
aws comprehend classify-document \
     --endpoint-arn arn:aws:comprehend:region:account-id:endpoint/endpoint name \
     --text 'From the Tuesday, April 16th, 1912 edition of The Guardian newspaper: The
 maiden voyage of the White Star liner Titanic,
     the largest ship ever launched ended in disaster. The Titanic started her trip
 from Southampton for New York on Wednesday. Late
     on Sunday night she struck an iceberg off the Grand Banks of Newfoundland. By
 wireless telegraphy she sent out signals of distress,
     and several liners were near enough to catch and respond to the call.'
```

Amazon Comprehend responds with the following:

```
{
    "Classes": [
       {
          "Name": "string",
          "Score": 0.9793661236763
       }
    ]
 }
```

Classify a semi-structured document using the AWS CLI

To analyze custom classification for a PDF, Word, or image file, run the `classify-document` command with the input file in the `bytes` parameter.

The following example uses an image as the input file. It uses the `fileb` option to base-64 encode the image file bytes. For more information, see Binary large objects in the AWS Command Line Interface User Guide.

This example also passes in a JSON file named `config.json` to set the text extraction options.

```
$ aws comprehend classify-document \
> --endpoint-arn arn \
> --language-code en \
> --bytes fileb://image1.jpg   \
> --document-reader-config file://config.json
```

The **config.json** file contains the following content.

```
{
   "DocumentReadMode": "FORCE_DOCUMENT_READ_ACTION",
   "DocumentReadAction": "TEXTRACT_DETECT_DOCUMENT_TEXT"
}
```

Amazon Comprehend responds with the following:

```
{
    "Classes": [
       {
          "Name": "string",
          "Score": 0.9793661236763
       }
    ]
}
```

For more information, see ClassifyDocument in the *Amazon Comprehend API Reference*.

Outputs for real-time analysis

Outputs for text inputs

For text inputs, the output includes the list of classes or labels identified by the classifier analysis. The following example shows a list with two classes.

```
"Classes": [
  {
     "Name": "abc",
     "Score": 0.2757999897003174,
     "Page": 1
  },
  {
    "Name": "xyz",
    "Score": 0.2721000015735626,
    "Page": 1
```

```
  }
]
```

Outputs for semi-structured inputs

For a semi-structured input document, or a text file, the output can include the following additional fields:

- DocumentMetadata – Extraction information about the document. The metadata includes a list of pages in the document, with the number of characters extracted from each page. This field is present in the response if the request included the `Byte` parameter.
- DocumentType – The document type for each page in the input document. This field is present in the response if the request included the `Byte` parameter.
- Errors – Page-level errors that the system detected while processing the input document. The field is empty if the system encountered no errors.
- Warnings – Warnings detected while processing the input document. The response includes a warning if there is a mismatch between the input document type and the model type associated with the endpoint that you specified. The field is empty if the system generated no warnings.

For more details about these output fields, see ClassifyDocument in the *Amazon Comprehend API Reference*.

The following example shows the output for a one-page native PDF input document.

```
{
  "Classes": [
      {
          "Name": "123",
          "Score": 0.39570000767707825,
          "Page": 1
      },
      {
          "Name": "abc",
          "Score": 0.2757999897003174,
          "Page": 1
      },
      {
          "Name": "xyz",
          "Score": 0.2721000015735626,
```

```
            "Page": 1
        }
    ],
    "DocumentMetadata": {
        "Pages": 1,
        "ExtractedCharacters": [
            {
                "Page": 1,
                "Count": 2013
            }
        ]
    },
    "DocumentType": [
        {
            "Page": 1,
            "Type": "NATIVE_PDF"
        }
    ]
}
```

Running asynchronous jobs

After you train a custom classifier, you can use asynchronous jobs to analyze large documents or multiple documents in one batch.

Custom classification accepts a variety of input document types. For details, see Inputs for asynchronous custom analysis.

If you plan to analyze image files or scanned PDF documents, your IAM policy must grant permissions to use two Amazon Textract API methods (DetectDocumentText and AnalyzeDocument). Amazon Comprehend invokes these methods during text extraction. For an example policy, see Permissions required to perform document analysis actions.

For classification of semi-structured documents (image, PDF, or Docx files) using a plain-text model, use the `one document per file` input format. Also, include the `DocumentReaderConfig` parameter in your StartDocumentClassificationJob request.

Topics

- File formats for async analysis
- Analysis jobs for custom classification (console)

- Analysis jobs for custom classification (API)
- Outputs for asynchronous analysis jobs

File formats for async analysis

When you run async analysis with your model, you have a choice of formats for input documents: `One document per line` or `one document per file`. The format you use depends on the type of documents you want to analyze, as described in the following table.

Description	Format
The input contains multiple files. Each file contains one input document. This format is best for collections of large documents, such as newspaper articles or scientific papers. Also, use this format for semi-structured documents (image, PDF, or Docx files) using a native document classifier.	One document per file
The input is one or more files. Each line in the file is a separate input document. This format is best for short documents, such as text messages or social media posts.	One document per line

One document per file

With `one document per file` format, each file represents one input document.

One document per line

With the `One document per line` format, each document is placed on a separate line and no header is used. The label is not included on each line (since you don't yet know the label for the document). Each line of the file (the end of the individual document) must end with a line feed (LF, \n), a carriage return (CR, \r), or both (CRLF, \r\n). Don't use the UTF-8 line separator (u+2028) to end a line.

The following example shows the format of the input file.

```
Text of document 1 \n
Text of document 2 \n
Text of document 3 \n
Text of document 4 \n
```

For either format, use UTF-8 encoding for text files. After you prepare the files, place them in the S3 bucket that you're using for input data.

When you start a classification job, you specify this Amazon S3 location for your input data. The URI must be in the same Region as the API endpoint that you are calling. The URI can point to a single file (as when using the "one document per line" method, or it can be the prefix for a collection of data files.

For example, if you use the URI `S3://bucketName/prefix`, if the prefix is a single file, Amazon Comprehend uses that file as input. If more than one file begins with the prefix, Amazon Comprehend uses all of them as input.

Grant Amazon Comprehend access to the S3 bucket that contains your document collection and output files. For more information, see Role-based permissions required for asynchronous operations.

Analysis jobs for custom classification (console)

After you create and train a custom document classifier, you can use the console to run custom classification jobs with the model.

To create a custom classification job (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Analysis jobs** and then choose **Create job**.
3. Give the classification job a name. The name must be unique to your account and current Region.
4. Under **Analysis type**, choose **Custom classification**.
5. From **Select classifier**, choose the custom classifier to use.
6. (Optional) If you choose to encrypt the data that Amazon Comprehend uses while processing your job, choose **Job encryption**. Then choose whether to use a KMS key associated with the current account, or one from another account.

- If you are using a key associated with the current account, choose the key ID for **KMS key ID**.
- If you are using a key associated with a different account, enter the ARN for the key ID under **KMS key ARN**.

> **Note**
>
> For more information on creating and using KMS keys and the associated encryption, see Key management service (KMS).

7. Under **Input data**, enter the location of the Amazon S3 bucket that contains your input documents or navigate to it by choosing **Browse S3**. This bucket must be in the same Region as the API that you are calling. The IAM role that you're using for access permissions for the classification job must have reading permissions for the S3 bucket.

   To achieve the highest level of accuracy in training a model, match the type of input to the classifier model type. The classifier job returns a warning if you submit native documents to a plain-text model, or plain text documents to a native document model. For more information, see Training classification models.
8. (Optional) For **Input format**, you can choose the format of the input documents. The format can be one document per file, or one document per line in a single file. One document per line applies only to text documents.
9. (Optional) For **Document read mode**, you can override the default text extraction actions. For more information, see Setting text extraction options.
10. Under **Output data**, enter the location of the Amazon S3 bucket where Amazon Comprehend should write the job's output data or navigate to it by choosing **Browse S3**. This bucket must be in the same Region as the API that you are calling. The IAM role that you're using for access permissions for the classification job must have write permissions for the S3 bucket.
11. (Optional) If you choose to encrypt the output result from your job, choose **Encryption**. Then choose whether to use a KMS key associated with the current account, or one from another account.

    - If you are using a key associated with the current account, choose the key alias or ID for **KMS key ID**.
    - If you are using a key associated with a different account, enter the ARN for the key alias or ID under **KMS key ID**.

12. (Optional) To launch your resources into Amazon Comprehend from a VPC, enter the VPC ID under **VPC** or choose the ID from the drop-down list.

    1. Choose the subnet under **Subnet(s)**. After you select the first subnet, you can choose additional ones.
    2. Under **Security Group(s)**, choose the security group to use if you specified one. After you select the first security group, you can choose additional ones.

    > **Note**
    >
    > When you use a VPC with your classification job, the `DataAccessRole` used for the Create and Start operations must grant permissions to the VPC that accesses the output bucket.

13. Choose **Create job** to create the document classification job.

Analysis jobs for custom classification (API)

After you create and train a custom document classifier, you can use the classifier to run analysis jobs.

Use the StartDocumentClassificationJob operation to start classifying unlabeled documents. You specify the S3 bucket that contains the input documents, the S3 bucket for the output documents, and the classifier to use.

To achieve the highest level of accuracy in training a model, match the type of input to the classifier model type. The classifier job returns a warning if you submit native documents to a plain-text model, or plain text documents to a native document model. For more information, see Training classification models.

StartDocumentClassificationJob is asynchronous. Once you have started the job, use the DescribeDocumentClassificationJob operation to monitor its progress. When the `Status` field in the response shows `COMPLETED`, you can access the output in the location that you specified.

Topics

- Using the AWS Command Line Interface
- Using the AWS SDK for Java or SDK for Python

Using the AWS Command Line Interface

The following examples the `StartDocumentClassificationJob` operation, and other custom classifier APIs with the AWS CLI.

The following examples use the command format for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Run a custom classification job using the `StartDocumentClassificationJob` operation.

```
aws comprehend start-document-classification-job \
     --region region \
     --document-classifier-arn arn:aws:comprehend:region:account number:document-
classifier/testDelete \
     --input-data-config S3Uri=s3://S3Bucket/docclass/file
 name,InputFormat=ONE_DOC_PER_LINE \
     --output-data-config S3Uri=s3://S3Bucket/output \
     --data-access-role-arn arn:aws:iam::account number:role/resource name
```

Get information on a custom classifier with the job id using the `DescribeDocumentClassificationJob` operation.

```
aws comprehend describe-document-classification-job \
     --region region \
     --job-id job id
```

List all custom classification jobs in your account using the `ListDocumentClassificationJobs` operation.

```
aws comprehend list-document-classification-jobs
     --region region
```

Using the AWS SDK for Java or SDK for Python

For SDK examples of how to start a custom classifier job, see Use `StartDocumentClassificationJob` with an AWS SDK or CLI.

Outputs for asynchronous analysis jobs

After an analysis job completes, it stores the results in the S3 bucket that you specified in the request.

Outputs for text inputs

For either format of text input documents (multi-class or multi-label), the job output consists of a single file named `output.tar.gz`. It's a compressed archive file that contains a text file with the output.

Multi-class output

When you use a classifier trained in multi-class mode, your results show `classes`. Each of these `classes` is the class used to create the set of categories when training your classifier.

For more details about these output fields, see ClassifyDocument in the *Amazon Comprehend API Reference*.

The following examples use the following mutually exclusive classes.

```
DOCUMENTARY
SCIENCE_FICTION
ROMANTIC_COMEDY
SERIOUS_DRAMA
OTHER
```

If your input data format is one document per line, the output file contains one line for each line in the input. Each line includes the file name, the zero-based line number of the input line, and the class or classes found in the document. It ends with the confidence that Amazon Comprehend has that the individual instance was correctly classified.

For example:

```
{"File": "file1.txt", "Line": "0", "Classes": [{"Name": "Documentary", "Score":
 0.8642}, {"Name": "Other", "Score": 0.0381}, {"Name": "Serious_Drama", "Score":
 0.0372}]}
{"File": "file1.txt", "Line": "1", "Classes": [{"Name": "Science_Fiction", "Score":
 0.5}, {"Name": "Science_Fiction", "Score": 0.0381}, {"Name": "Science_Fiction",
 "Score": 0.0372}]}
{"File": "file2.txt", "Line": "2", "Classes": [{"Name": "Documentary", "Score": 0.1},
 {"Name": "Documentary", "Score": 0.0381}, {"Name": "Documentary", "Score": 0.0372}]}
{"File": "file2.txt", "Line": "3", "Classes": [{"Name": "Serious_Drama", "Score":
 0.3141}, {"Name": "Other", "Score": 0.0381}, {"Name": "Other", "Score": 0.0372}]}
```

If your input data format is one document per file, the output file contains one line for each document. Each line has the name of the file and the class or classes found in the document. It ends with the confidence that Amazon Comprehend classified the individual instance accurately.

For example:

```
{"File": "file0.txt", "Classes": [{"Name": "Documentary", "Score": 0.8642}, {"Name":
 "Other", "Score": 0.0381}, {"Name": "Serious_Drama", "Score": 0.0372}]}
{"File": "file1.txt", "Classes": [{"Name": "Science_Fiction", "Score": 0.5}, {"Name":
 "Science_Fiction", "Score": 0.0381}, {"Name": "Science_Fiction", "Score": 0.0372}]}
{"File": "file2.txt", "Classes": [{"Name": "Documentary", "Score": 0.1}, {"Name":
 "Documentary", "Score": 0.0381}, {"Name": "Domentary", "Score": 0.0372}]}
{"File": "file3.txt", "Classes": [{"Name": "Serious_Drama", "Score": 0.3141}, {"Name":
 "Other", "Score": 0.0381}, {"Name": "Other", "Score": 0.0372}]}
```

Multi-label output

When you use a classifier trained in multi-label mode, your results show `labels`. Each of these `labels` is the label used to create the set of categories when training your classifier.

The following examples use these unique labels.

```
SCIENCE_FICTION
ACTION
DRAMA
COMEDY
ROMANCE
```

If your input data format is one document per line, the output file contains one line for each line in the input. Each line includes the file name, the zero-based line number of the input line, and the class or classes found in the document. It ends with the confidence that Amazon Comprehend has that the individual instance was correctly classified.

For example:

```
{"File": "file1.txt", "Line": "0", "Labels": [{"Name": "Action", "Score": 0.8642},
 {"Name": "Drama", "Score": 0.650}, {"Name": "Science Fiction", "Score": 0.0372}]}
{"File": "file1.txt", "Line": "1", "Labels": [{"Name": "Comedy", "Score": 0.5},
 {"Name": "Action", "Score": 0.0381}, {"Name": "Drama", "Score": 0.0372}]}
{"File": "file1.txt", "Line": "2", "Labels": [{"Name": "Action", "Score": 0.9934},
 {"Name": "Drama", "Score": 0.0381}, {"Name": "Action", "Score": 0.0372}]}
```

```
{"File": "file1.txt", "Line": "3", "Labels": [{"Name": "Romance", "Score": 0.9845},
 {"Name": "Comedy", "Score": 0.8756}, {"Name": "Drama", "Score": 0.7723}, {"Name":
 "Science_Fiction", "Score": 0.6157}]}
```

If your input data format is one document per file, the output file contains one line for each document. Each line has the name of the file and the class or classes found in the document. It ends with the confidence that Amazon Comprehend classified the individual instance accurately.

For example:

```
{"File": "file0.txt", "Labels": [{"Name": "Action", "Score": 0.8642}, {"Name": "Drama",
 "Score": 0.650}, {"Name": "Science Fiction", "Score": 0.0372}]}
{"File": "file1.txt", "Labels": [{"Name": "Comedy", "Score": 0.5}, {"Name": "Action",
 "Score": 0.0381}, {"Name": "Drama", "Score": 0.0372}]}
{"File": "file2.txt", "Labels": [{"Name": "Action", "Score": 0.9934}, {"Name": "Drama",
 "Score": 0.0381}, {"Name": "Action", "Score": 0.0372}]}
{"File": "file3.txt", "Labels": [{"Name": "Romance", "Score": 0.9845}, {"Name":
 "Comedy", "Score": 0.8756}, {"Name": "Drama", "Score": 0.7723}, {"Name":
 "Science_Fiction", "Score": 0.6157}]}
```

Outputs for semi-structured input documents

For semi-structured input documents, the output can include the following additional fields:

- DocumentMetadata – Extraction information about the document. The metadata includes a list of pages in the document, with the number of characters extracted from each page. This field is present in the response if the request included the `Byte` parameter.
- DocumentType – The document type for each page in the input document. This field is present in the response if the request included the `Byte` parameter.
- Errors – Page-level errors that the system detected while processing the input document. The field is empty if the system encountered no errors.

For more details about these output fields, see ClassifyDocument in the *Amazon Comprehend API Reference*.

The following example shows output for a two-page scanned PDF file.

```
[{ #First page output
```

```
    "Classes": [
        {
            "Name": "__label__2 ",
            "Score": 0.9993996620178223
        },
        {
            "Name": "__label__3 ",
            "Score": 0.0004330444789957255
        }
    ],
    "DocumentMetadata": {
        "PageNumber": 1,
        "Pages": 2
    },
    "DocumentType": "ScannedPDF",
    "File": "file.pdf",
    "Version": "VERSION_NUMBER"
},
#Second page output
{
    "Classes": [
        {
            "Name": "__label__2 ",
            "Score": 0.9993996620178223
        },
        {
            "Name": "__label__3 ",
            "Score": 0.0004330444789957255
        }
    ],
    "DocumentMetadata": {
        "PageNumber": 2,
        "Pages": 2
    },
    "DocumentType": "ScannedPDF",
    "File": "file.pdf",
    "Version": "VERSION_NUMBER"
}]
```

Custom entity recognition

Custom entity recognition extends the capability of Amazon Comprehend by helping you identify your specific new entity types that are not in the preset generic entity types. This means that you can analyze documents and extract entities like product codes or business-specific entities that fit your particular needs.

Building an accurate custom entity recognizer on your own can be a complex process, requiring preparation of large sets of manually annotated training documents and the selection of the right algorithms and parameters for model training. Amazon Comprehend helps to reduce the complexity by providing automatic annotation and model development to create a custom entity recognition model.

Creating a custom entity recognition model is a more effective approach than using string matching or regular expressions to extract entities from documents. For example, to extract ENGINEER names in a document, it is difficult to enumerate all possible names. Additionally, without context, it is challenging to distinguish between ENGINEER names and ANALYST names. A custom entity recognition model can learn the context where those names are likely to appear. Additionally, string matching will not detect entities that have typos or follow new naming conventions, while this is possible using a custom model.

You have two options for creating a custom model:

1. Annotations – provide a data set containing annotated entities for model training.
2. Entity lists (plaintext only) – provide a list of entities and their type label (such as `PRODUCT_CODES` and a set of unannotated documents containing those entities for model training.

When you create a custom entity recognizer using annotated PDF files, you can use that recognizer with a variety of input file formats: plaintext, image files (JPG, PNG, TIFF), PDF files, and Word documents, with no pre-processing or doc flattening required. Amazon Comprehend doesn't support annotation of image files or Word documents.

> **Note**
>
> A custom entity recognizer using annotated PDF files supports English documents only.

You can train a model on up to 25 custom entities at once. For more details, see the Guidelines and quotas page.

After your model is trained, you can use the model for real-time entity detection and in entity detection jobs.

Topics

Preparing entity recognizer training data

To train a successful custom entity recognition model, it's important to supply the model trainer with high quality data as input. Without good data, the model won't learn how to correctly identify entities.

You can choose one of two ways to provide data to Amazon Comprehend in order to train a custom entity recognition model:

- **Entity list** – Lists the specific entities so Amazon Comprehend can train to identify your custom entities. Note: Entity lists can only be used for plaintext documents.
- **Annotations** – Provides the location of your entities in a number of documents so Amazon Comprehend can train on both the entity and its context. To create a model for analyzing image files, PDFs, or Word documents, you must train your recognizer using PDF annotations.

In both cases, Amazon Comprehend learns about the kind of documents and the context where the entities occur and builds a recognizer that can generalize to detect the new entities when you analyze documents.

When you create a custom model (or train a new version), you can provide a test dataset. If you do not provide test data, Amazon Comprehend reserves 10% of the input documents to test the model. Amazon Comprehend trains the model with the remaining documents.

If you provide a test dataset for your annotations training set, the test data must include at least one annotation for each of the entity types specified in the creation request.

Topics

When to use annotations vs entity lists

Creating annotations takes more work than creating an entity list, but the resulting model can be significantly more accurate. Using an entity list is quicker and less work-intensive, but the results are less refined and less accurate. This is because the annotations provide more context for Amazon Comprehend to use when training the model. Without that context, Amazon Comprehend will have a higher number of false positives when trying to identify the entities.

There are scenarios when it makes more business sense to avoid the higher expense and workload of using annotations. For example, the name John Johnson is significant to your search, but whether it's the exact individual isn't relevant. Or the metrics when using the entity list are good enough to provide you with the recognizer results that you need. In such instances, using an entity list instead can be the more effective choice.

We recommend using the annotations mode in the following cases:

- If you plan to run inferences for image files, PDFs, or Word documents. In this scenario, you train a model using annotated PDF files and use the model to run inference jobs for image files, PDFs, and Word documents.
- When the meaning of the entities could be ambiguous and context-dependent. For example, the term *Amazon* could either refer to the river in Brazil, or the online retailer Amazon.com. When you build a custom entity recognizer to identify business entities such as *Amazon*, you should use annotations instead of an entity list because this method is better able to use context to find entities.
- When you are comfortable setting up a process to acquire annotations, which can require some effort.

We recommend using an entity list in the following cases:

- When you already have a list of entities or when it is relatively easy to compose a comprehensive list of entities. If you use an entity list, the list should be complete or at least covers the majority of valid entities that might appear in the documents you provide for training.

- For first-time users, it is generally recommended to use an entity list because this requires a smaller effort than constructing annotations. However, it is important to note that the trained model might not be as accurate as if you used annotations.

Entity lists (plaintext only)

To train a model using an entity list, you provide two pieces of information: a list of the entity names with their corresponding custom entity types and a collection of unannotated documents in which you expect your entities to appear.

When you provide an Entity List, Amazon Comprehend uses an intelligent algorithm to detect occurrences of the entity in the documents to serve as the basis for training the custom entity recognizer model.

For entity lists, provide at least 25 entity matches per entity type in the entity list.

An entity list for custom entity recognition needs a comma-separated value (CSV) file, with the following columns:

- **Text**— The text of an entry example exactly as seen in the accompanying document corpus.
- **Type**—The customer-defined entity type. Entity types must an uppercase, underscore separated string such as MANAGER or SENIOR_MANAGER. Up to 25 entity types can be trained per model.

The file `documents.txt` contains four lines:

```
Jo Brown is an engineer in the high tech industry.
John Doe has been a engineer for 14 years.
Emilio Johnson is a judge on the Washington Supreme Court.
Our latest new employee, Jane Smith, has been a manager in the industry for 4 years.
```

The CSV file with the list of entities has the following lines:

```
Text, Type
Jo Brown, ENGINEER
John Doe, ENGINEER
Jane Smith, MANAGER
```

> **Note**
>
> In the entities list, the entry for Emilio Johnson is not present because it does not contain either the ENGINEER or MANAGER entity.

Creating your data files

It is important that your entity list be in a properly configured CSV file so your chance of having problems with your entity list file is minimal. To manually configure your CSV file, the following must be true:

- UTF-8 encoding must be explicitly specified, even if its used as a default in most cases.
- It must include the column names: `Type` and `Text`.

We highly recommended that CSV input files are generated programmatically to avoid potential issues.

The following example uses Python to generate a CSV for the annotations shown above:

```
import csv
with open("./entitylist/entitylist.csv", "w", encoding="utf-8") as csv_file:
    csv_writer = csv.writer(csv_file)
    csv_writer.writerow(["Text", "Type"])
    csv_writer.writerow(["Jo Brown", " ENGINEER"])
    csv_writer.writerow(["John Doe", " ENGINEER"])
    csv_writer.writerow(["Jane Smith", " MANAGER"])
```

Best practices

There are a number of things to consider to get the best result when using an entity list, including:

- The order of the entities in your list has no effects on model training.
- Use entity list items that cover 80%-100% of positive entity examples mentioned in the unannotated corpus of documents.
- Avoid entity examples that match non-entities in the document corpus by removing common words and phrases. Even a handful of incorrect matches can significantly affect the accuracy of your resulting model. For example, a word like *the* in the entity list will result in a high number of

matches which are unlikely to be the entities you are looking for and thus will significantly affect your accuracy.

- Input data should not contain duplicates. Presence of duplicate samples might result into test set contamination and therefore negatively affect training process, model metrics, and behavior.
- Provide documents that resemble real use cases as closely as possible. Don't use toy data or synthesized data for production systems. The input data should be as diverse as possible to avoid overfitting and help underlying model better generalize on real examples.
- The entity list is case sensitive, and regular expressions are not currently supported. However, the trained model can often still recognize entities even if they do not match exactly to the casing provided in the entity list.
- If you have an entity that is a substring of another entity (such as “Smith” and “Jane Smith”), provide both in the entity list.

Additional suggestions can be found at Improving custom entity recognizer performance

Annotations

Annotations label entities in context by associating your custom entity types with the locations where they occur in your training documents.

By submitting annotation along with your documents, you can increase the accuracy of the model. With Annotations, you're not simply providing the location of the entity you're looking for, but you're also providing more accurate context to the custom entity you're seeking.

For instance, if you're searching for the name John Johnson, with the entity type JUDGE, providing your annotation might help the model to learn that the person you want to find is a judge. If it is able to use the context, then Amazon Comprehend won't find people named John Johnson who are attorneys or witnesses. Without providing annotations, Amazon Comprehend will create its own version of an annotation, but won't be as effective at including only judges. Providing your own annotations might help to achieve better results and to generate models that are capable of better leverage context when extracting custom entities.

Topics

- Minimum number of annotations
- Annotation best practices
- Plain-text annotation files

- PDF annotation files
- Annotating PDF files

Minimum number of annotations

The minimum number of input documents and annotations required to train a model depends on the type of annotations.

PDF annotations

To create a model for analyzing image files, PDFs, or Word documents, train your recognizer using PDF annotations. For PDF annotations, provide at least 250 input documents and at least 100 annotations per entity.

If you provide a test dataset, the test data must include at least one annotation for each of the entity types specified in the creation request.

Plain-text annotations

To create a model for analyzing text documents, you can train your recognizer using plain-text annotations.

For plain-text annotations, provide at least three annotated input documents and at least 25 annotations per entity. If you provide less than 50 annotations total, Amazon Comprehend reserves more than 10% of the input documents to test the model (unless you provided a test dataset in the training request). Don't forget that the minimum document corpus size is 5 KB.

If your input contains only a few training documents, you may encounter an error that the training input data contains too few documents that mention one of the entities. Submit the job again with additional documents that mention the entity.

If you provide a test dataset, the test data must include at least one annotation for each of the entity types specified in the creation request.

For an example of how to benchmark a model with a small dataset, see Amazon Comprehend announces lower annotation limits for custom entity recognition on the AWS blog site.

Annotation best practices

There are a number of things to consider to get the best result when using annotations, including:

- Annotate your data with care and verify that you annotate every mention of the entity. Imprecise annotations can lead to poor results.
- Input data should not contain duplicates, like a duplicate of a PDF you are going to annotate. Presence of a duplicate sample might result in test set contamination and could negatively affect the training process, model metrics, and model behavior.
- Make sure that all of your documents are annotated, and that the documents without annotations are due to lack of legitimate entities, not due to negligence. For example, if you have a document that says "J Doe has been an engineer for 14 years", you should also provide an annotation for "J Doe" as well as "John Doe". Failing to do so confuses the model and can result in the model not recognizing "J Doe" as ENGINEER. This should be consistent within the same document and across documents.
- In general, more annotations lead to better results.
- You can train a model with the minimum number of documents and annotations, but adding data usually improves the model. We recommend increasing the volume of annotated data by 10% to increase the accuracy of the model. You can run inference on a test dataset which remains unchanged and can be tested by different model versions. You can then compare the metrics for successive model versions.
- Provide documents that resemble real use cases as closely as possible. Synthesized data with repetitive patterns should be avoided. The input data should be as diverse as possible to avoid overfitting and help the underlying model better generalize on real examples.
- It is important that documents should be diverse in terms of word count. For example, if all documents in the training data are short, the resulting model may have difficulty predicting entities in longer documents.
- Try and give the same data distribution for training as you expect to be using when you're actually detecting your custom entities (inference time). For example, at inference time, if you expect to be sending us documents that have no entities in them, this should also be part of your training document set.

For additional suggestions, see Improving custom entity recognizer performance.

Plain-text annotation files

For plain-text annotations, you create a comma-separated value (CSV) file that contains a list of annotations. The CSV file must contain the following columns if your training file input format is **one document per line**.

File	Line	Begin offset	End offset	Type
The name of the file containing the document. For example, if one of the document files is located at `s3://my-S3-bucket/test-files/documents.txt`, the value in the `File` column will be `documents.txt`. You must include the file extension (in this case `'.txt'`) as part of the file name.	The line number containing the entity. Omit this column if your input format is one document per file.	The character offset in the input text (relative to the beginning of the line) that shows where the entity begins. The first character is at position 0.	The character offset in the input text that shows where the entity ends.	The customer-defined entity type. Entity types must be an uppercase, underscore-separated string. We recommend using descriptive entity types such as `MANAGER`, `SENIOR_MANAGER`, or `PRODUCT_CODE`. Up to 25 entity types can be trained per model.

If your training file input format is **one document per file**, you omit the line number column and the **Begin offset** and **End offset** values are the offsets of the entity from the start of the document.

The following example is for one document per line. The file `documents.txt` contains four lines (rows 0, 1, 2, and 3):

```
Diego Ramirez is an engineer in the high tech industry.
Emilio Johnson has been an engineer for 14 years.
J Doe is a judge on the Washington Supreme Court.
Our latest new employee, Mateo Jackson, has been a manager in the industry for 4 years.
```

The CSV file with the list of annotations is as follows:

```
File, Line, Begin Offset, End Offset, Type
documents.txt, 0, 0, 13, ENGINEER
documents.txt, 1, 0, 14, ENGINEER
documents.txt, 3, 25, 38, MANAGER
```

> **Note**
>
> In the annotations file, the line number containing the entity starts with line 0. In this example, the CSV file contains no entry for line 2 because there is no entity in line 2 of `documents.txt`.

Creating your data files

It's important to put your annotations in a properly configured CSV file to reduce the risk of errors. To manually configure your CSV file, the following must be true:

- UTF-8 encoding must be explicitly specified, even if its used as a default in most cases.
- The first line contains the column headers: `File`, `Line` (optional), `Begin Offset`, `End Offset`, `Type`.

We highly recommended that you generate the CSV input files programmatically to avoid potential issues.

The following example uses Python to generate a CSV for the annotations shown earlier:

```
import csv
with open("./annotations/annotations.csv", "w", encoding="utf-8") as csv_file:
    csv_writer = csv.writer(csv_file)
    csv_writer.writerow(["File", "Line", "Begin Offset", "End Offset", "Type"])
    csv_writer.writerow(["documents.txt", 0, 0, 11, "ENGINEER"])
    csv_writer.writerow(["documents.txt", 1, 0, 5, "ENGINEER"])
    csv_writer.writerow(["documents.txt", 3, 25, 30, "MANAGER"])
```

PDF annotation files

For PDF annotations, you use SageMaker Ground Truth to create a labeled dataset in an augmented manifest file. Ground Truth is a data labeling service that helps you (or a workforce

that you employ) to build training datasets for machine learning models. Amazon Comprehend accepts augmented manifest files as training data for custom models. You can provide these files when you create a custom entity recognizer by using the Amazon Comprehend console or the CreateEntityRecognizer API action.

You can use the Ground Truth built-in task type, Named Entity Recognition, to create a labeling job to have workers identify entities in text. To learn more, see Named Entity Recognition in the *Amazon SageMaker Developer Guide*. To learn more about Amazon SageMaker Ground Truth, see Use Amazon SageMaker Ground Truth to Label Data.

> **Note**
>
> Using Ground Truth, you can define overlapping labels (text that you associate with more than one label). However, Amazon Comprehend entity recognition does not support overlapping labels.

Augmented manifest files are in JSON lines format. In these files, each line is a complete JSON object that contains a training document and its associated labels. The following example is an augmented manifest file that trains an entity recognizer to detect the professions of individuals who are mentioned in the text:

```
{"source":"Diego Ramirez is an engineer in the high tech
 industry.","NamedEntityRecognitionDemo":{"annotations":{"entities":
[{"endOffset":13,"startOffset":0,"label":"ENGINEER"}],"labels":
[{"label":"ENGINEER"}]}},"NamedEntityRecognitionDemo-metadata":
{"entities":[{"confidence":0.92}],"job-name":"labeling-job/
namedentityrecognitiondemo","type":"groundtruth/text-span","creation-
date":"2020-05-14T21:45:27.175903","human-annotated":"yes"}}
{"source":"J Doe is a judge on the Washington Supreme
 Court.","NamedEntityRecognitionDemo":{"annotations":{"entities":
[{"endOffset":5,"startOffset":0,"label":"JUDGE"}],"labels":
[{"label":"JUDGE"}]}},"NamedEntityRecognitionDemo-metadata":
{"entities":[{"confidence":0.72}],"job-name":"labeling-job/
namedentityrecognitiondemo","type":"groundtruth/text-span","creation-
date":"2020-05-14T21:45:27.174910","human-annotated":"yes"}}
{"source":"Our latest new employee, Mateo Jackson, has been a manager in
 the industry for 4 years.","NamedEntityRecognitionDemo":{"annotations":
{"entities":[{"endOffset":38,"startOffset":26,"label":"MANAGER"}],"labels":
[{"label":"MANAGER"}]}},"NamedEntityRecognitionDemo-metadata":
{"entities":[{"confidence":0.91}],"job-name":"labeling-job/
```

```
namedentityrecognitiondemo","type":"groundtruth/text-span","creation-
date":"2020-05-14T21:45:27.174035","human-annotated":"yes"}}
```

Each line in this JSON lines file is a complete JSON object, where the attributes include the document text, the annotations, and other metadata from Ground Truth. The following example is a single JSON object in the augmented manifest file, but it's formatted for readability:

```
{
  "source": "Diego Ramirez is an engineer in the high tech industry.",
  "NamedEntityRecognitionDemo": {
    "annotations": {
      "entities": [
        {
          "endOffset": 13,
          "startOffset": 0,
          "label": "ENGINEER"
        }
      ],
      "labels": [
        {
          "label": "ENGINEER"
        }
      ]
    }
  },
  "NamedEntityRecognitionDemo-metadata": {
    "entities": [
      {
        "confidence": 0.92
      }
    ],
    "job-name": "labeling-job/namedentityrecognitiondemo",
    "type": "groundtruth/text-span",
    "creation-date": "2020-05-14T21:45:27.175903",
    "human-annotated": "yes"
  }
}
```

In this example, the `source` attribute provides the text of the training document, and the `NamedEntityRecognitionDemo` attribute provides the annotations for the entities in the text. The name of the `NamedEntityRecognitionDemo` attribute is arbitrary, and you provide a name of your choice when you define the labeling job in Ground Truth.

In this example, the `NamedEntityRecognitionDemo` attribute is the *label attribute name*, which is the attribute that provides the labels that a Ground Truth worker assigns to the training data. When you provide your training data to Amazon Comprehend, you must specify one or more label attribute names. The number of attribute names that you specify depends on whether your augmented manifest file is the output of a single labeling job or a chained labeling job.

If your file is the output of a single labeling job, specify the single label attribute name that was used when the job was created in Ground Truth.

If your file is the output of a chained labeling job, specify the label attribute name for one or more jobs in the chain. Each label attribute name provides the annotations from an individual job. You can specify up to 5 of these attributes for augmented manifest files that are produced by chained labeling jobs.

In an augmented manifest file, the label attribute name typically follows the `source` key. If the file is the output of a chained job, there will be multiple label attribute names. When you provide your training data to Amazon Comprehend, provide only those attributes that contain annotations that are relevant for your model. Do not specify the attributes that end with "-metadata".

For more information about chained labeling jobs, and for examples of the output that they produce, see Chaining Labeling Jobs in the Amazon SageMaker Developer Guide.

Annotating PDF files

Before you can annotate your training PDFs in SageMaker Ground Truth, complete the following prerequisites:

- Install python3.8.x
- Install jq
- Install the AWS CLI

  If you're using the us-east-1 Region, you can skip installing the AWS CLI because it's already installed with your Python environment. In this case, you create a virtual environment to use Python 3.8 in AWS Cloud9.
- Configure your AWS credentials
- Create a private SageMaker Ground Truth workforce to support annotation

  Make sure to record the workteam name you choose in your new private workforce, as you use it during installation.

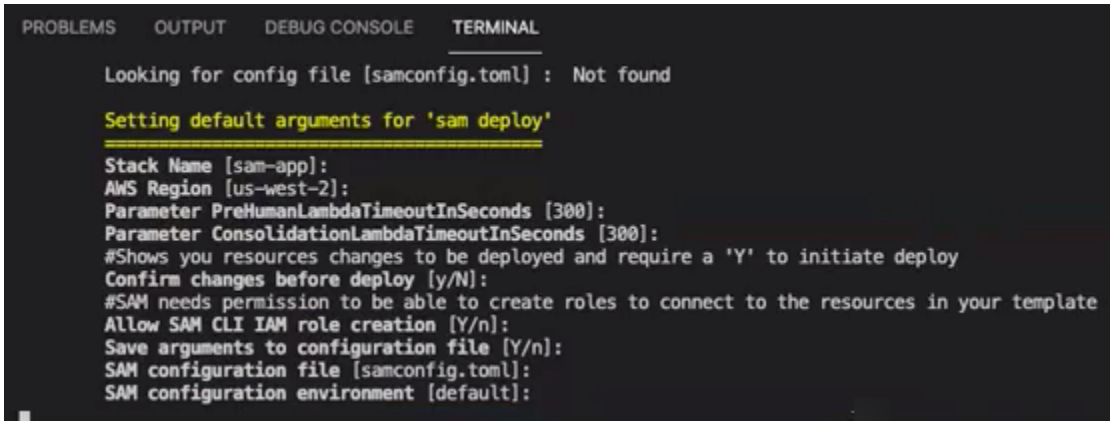
Topics

Setting up your environment

1. If using Windows, install Cygwin; if using Linux or Mac, skip this step.
2. Download the annotation artifacts from GitHub. Unzip the file.
3. From your terminal window, navigate to the unzipped folder (**amazon-comprehend-semi-structured-documents-annotation-tools-main**).
4. This folder includes a choice of `Makefiles` that you run to install dependencies, setup a Python virtualenv, and deploy the required resources. Review the **readme** file to make your choice.
5. The recommended option uses a single command to install all dependencies into a virtualenv, builds the AWS CloudFormation stack from the template, and deploys the stack to your AWS account with interactive guidance. Run the following command:

   ```
   make ready-and-deploy-guided
   ```

   This command presents a set of configuration options. Be sure your AWS Region is correct. For all other fields, you can either accept the default values or fill in custom values. If you modify the AWS CloudFormation stack name, write it down as you need it in the next steps.

   The CloudFormation stack creates and manage the AWS lambdas, AWS IAM roles, and AWS S3 buckets required for the annotation tool.

You can review each of these resources in the stack details page in the CloudFormation console.

6. The command prompts you to start the deployment. CloudFormation creates all the resources in the specified Region.

When the CloudFormation stack status transitions to create-complete, the resources are ready to use.

Uploading a PDF to an S3 bucket

In the Setting up section, you deployed a CloudFormation stack that creates an S3 bucket named **comprehend-semi-structured-documents-${AWS::Region}-${AWS::AccountId}**. You now upload your source PDF documents into this bucket.

> **Note**
>
> This bucket contains the data required for your labeling job. The Lambda Execution Role policy grants permission for the Lambda function to access this bucket.
> You can find the S3 bucket name in the **CloudFormation Stack details** using the '**SemiStructuredDocumentsS3Bucket**' key.

1. Create a new folder in the S3 bucket. Name this new folder '**src**'.
2. Add your PDF source files to your '**src**' folder. In a later step, you annotate these files to train your recognizer.
3. (Optional) Here's an AWS CLI example you can use to upload your source documents from a local directory into an S3 bucket:

```
aws s3 cp --recursive local-path-to-your-source-docs s3://deploy-guided/src/
```

Or, with your Region and Account ID:

```
aws s3 cp --recursive local-path-to-your-source-docs s3://deploy-guided-Region-AccountID/src/
```

4. You now have a private SageMaker Ground Truth workforce and have uploaded your source files to the S3 bucket, **deploy-guided/src/**; you're ready to start annotating.

Creating an annotation job

The **comprehend-ssie-annotation-tool-cli.py** script in the `bin` directory is a simple wrapper command that streamlines the creation of a SageMaker Ground Truth labeling job. The python script reads the source documents from your S3 bucket and creates a corresponding single-page manifest file with one source document per line. The script then creates a labeling job, which requires the manifest file as an input.

The python script uses the S3 bucket and CloudFormation stack that you configured in the Setting up section. Required input parameters for the script include:

- **input-s3-path**: S3 Uri to the source documents you uploaded to your S3 bucket. For example: `s3://deploy-guided/src/`. You can also add your Region and Account ID to this path. For example: `s3://deploy-guided-Region-AccountID/src/`.
- **cfn-name**: The CloudFormation stack name. If you used the default value for the stack name, your cfn-name is **sam-app**.
- **work-team-name**: The workforce name you created when you built out the private workforce in SageMaker Ground Truth.
- **job-name-prefix**: The prefix for the SageMaker Ground Truth labeling job. Note that there is a 29-character limit for this field. A timestamp is appended to this value. For example: `my-job-name-20210902T232116`.
- **entity-types**: The entities you want to use during your labeling job, separated by commas. This list must include all entities that you want to annotate in your training dataset. The Ground Truth labeling job displays only these entities for annotators to label content in the PDF documents.

To view additional arguments the script supports, use the `-h` option to display the help content.

- Run the following script with the input parameters as described in the previous list.

```
python bin/comprehend-ssie-annotation-tool-cli.py \
--input-s3-path s3://deploy-guided-Region-AccountID/src/ \
--cfn-name sam-app \
--work-team-name my-work-team-name \
--region us-east-1 \
--job-name-prefix my-job-name-20210902T232116 \
--entity-types "EntityA, EntityB, EntityC" \
--annotator-metadata "key=info,value=sample,key=Due Date,value=12/12/2021"
```

The script produces the following output:

```
Downloaded files to temp local directory /tmp/a1dc0c47-0f8c-42eb-9033-74a988ccc5aa
Deleted downloaded temp files from /tmp/a1dc0c47-0f8c-42eb-9033-74a988ccc5aa
Uploaded input manifest file to s3://comprehend-semi-structured-documents-
us-west-2-123456789012/input-manifest/my-job-name-20220203-labeling-
job-20220203T183118.manifest
Uploaded schema file to s3://comprehend-semi-structured-documents-us-
west-2-123456789012/comprehend-semi-structured-docs-ui-template/my-job-
name-20220203-labeling-job-20220203T183118/ui-template/schema.json
Uploaded template UI to s3://comprehend-semi-structured-documents-us-
west-2-123456789012/comprehend-semi-structured-docs-ui-template/my-job-
name-20220203-labeling-job-20220203T183118/ui-template/template-2021-04-15.liquid
Sagemaker GroundTruth Labeling Job submitted: arn:aws:sagemaker:us-
west-2:123456789012:labeling-job/my-job-name-20220203-labeling-job-20220203t183118
(amazon-comprehend-semi-structured-documents-annotation-tools-main)
 user@3c063014d632 amazon-comprehend-semi-structured-documents-annotation-tools-
main %
```

Annotating with SageMaker Ground Truth

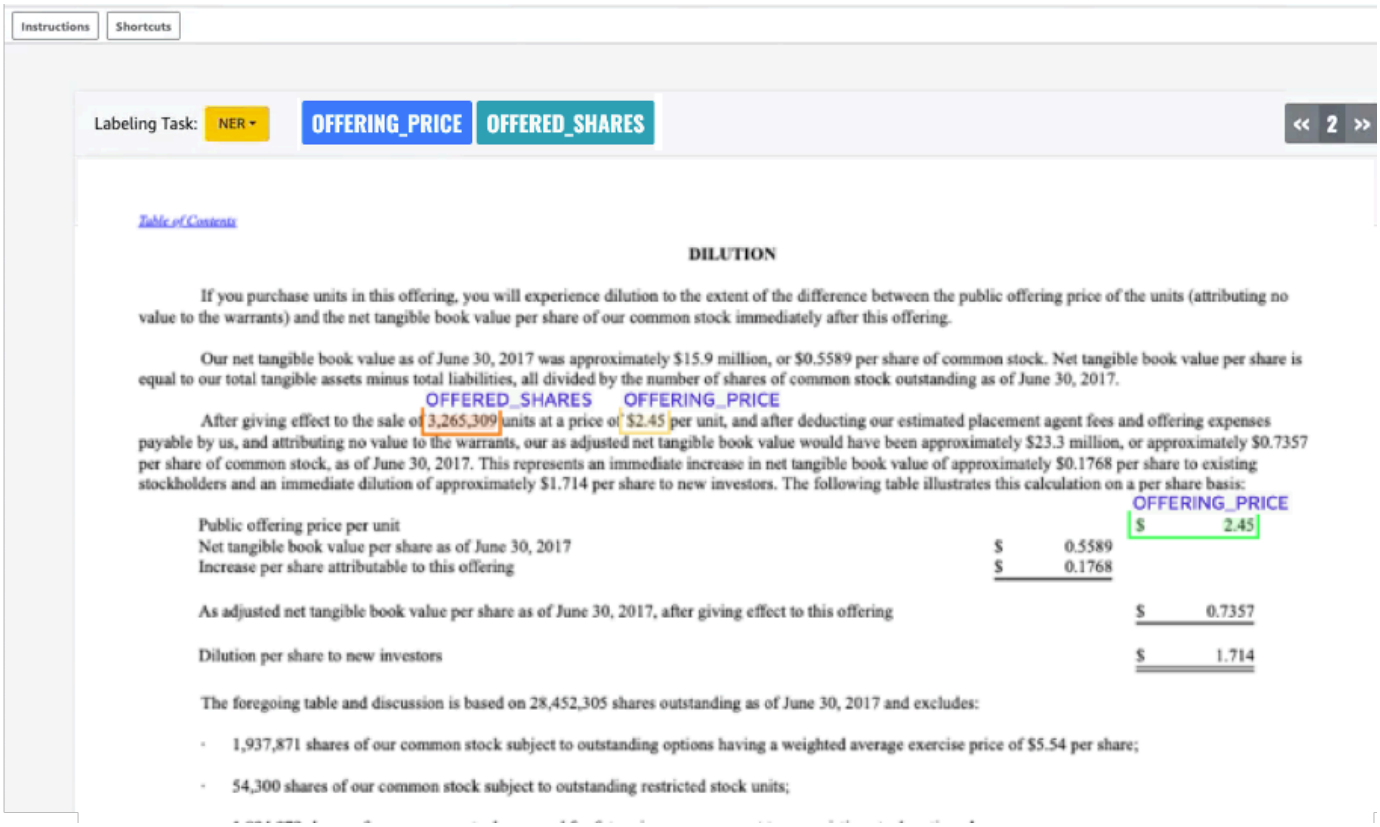
Now that you have configured the required resources and created a labeling job, you can log in to the labeling portal and annotate your PDFs.

1. Log in to the SageMaker console using either Chrome or Firefox web browsers.
2. Select **Labeling workforces** and choose **Private**.
3. Under **Private workforce summary**, select the labeling portal sign-in URL that you created with your private workforce. Sign in with the appropriate credentials.

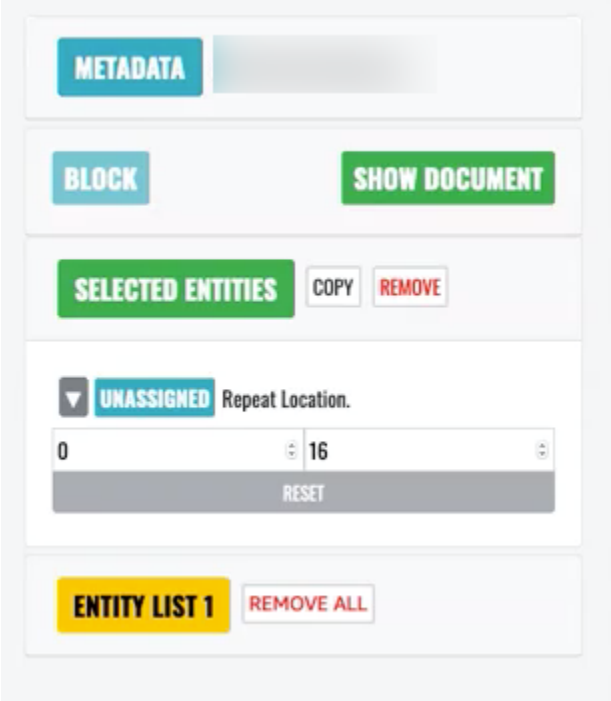
If you don't see any jobs listed, don't worry—it can take a while to update, depending on the number of files you uploaded for annotation.

4. Select your task and, in the top right corner, choose **Start working** to open the annotation screen.

   You'll see one of your documents open in the annotation screen and, above it, the entity types you provided during set up. To the right of your entity types, there is an arrow you can use to navigate through your documents.



Annotate the open document. You can also remove, undo, or auto tag your annotations on each document; these options are available in the right panel of the annotation tool.



To use auto tag, annotate an instance of one of your entities; all other instances of that specific word are then automatically annotated with that entity type.

Once you've finished, select **Submit** on the bottom right, then use the navigation arrows to move to the next document. Repeat this until you've annotated all your PDFs.

After you annotate all the training documents, you can find the annotations in JSON format in the Amazon S3 bucket at this location:

```
/output/your labeling job name/annotations/
```

The output folder also contains an output manifest file, which lists all the annotations within your training documents. You can find your output manifest file at the following location.

```
/output/your labeling job name/manifests/
```

Training custom entity recognizer models

A custom entity recognizer identifies only the entity types that you include when you train the model. It does not automatically include the preset entity types. If you want to also identify the

preset entity types,such as LOCATION, DATE, or PERSON, you need to provide additional training data for those entities.

When you create a custom entity recognizer using annotated PDF files, you can use the recognizer with a variety of input file formats: plaintext, image files (JPG, PNG, TIFF), PDF files, and Word documents, with no pre-processing or doc flattening required. Amazon Comprehend doesn't support annotation of image files or Word documents.

> **Note**
>
> A custom entity recognizer using annotated PDF files supports English documents only.

After you create a custom entity recognizer, you can monitor the progress of the request using the DescribeEntityRecognizer operation. Once the `Status` field is `TRAINED`, the recognizer model is ready to use for custom entity recognition.

Topics

- Train custom recognizers (console)
- Train custom entity recognizers (API)
- Custom entity recognizer metrics

Train custom recognizers (console)

You can create custom entity recognizers using the Amazon Comprehend console. This section shows you how to create and train a custom entity recognizer.

Creating a custom entity recognizer using the console - CSV format

To create the custom entity recognizer, first provide a dataset to train your model. With this dataset, include one of the following: a set of annotated documents or a list of entities and their type label, along with a set of documents containing those entities. For more information, see Custom entity recognition

To train a custom entity recognizer with a CSV file

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/

2. From the left menu, choose **Customization** and then choose **Custom entity recognition**.
3. Choose **Create new model**.
4. Give the recognizer a name. The name must be unique within the Region and account.
5. Select the language.
6. Under **Custom entity type**, enter a custom label that you want the recognizer to find in the dataset.

   The entity type must be uppercase, and if it consists of more than one word, separate the words with an underscore.
7. Choose **Add type**.
8. If you want to add an additional entity type, enter it, and then choose **Add type**. If you want to remove one of the entity types you've added, choose **Remove type** and then choose the entity type to remove from the list. A maximum of 25 entity types can be listed.
9. To encrypt your training job, choose **Recognizer encryption** and then choose whether to use a KMS key associated with the current account, or one from another account.

   - If you are using a key associated with the current account, for **KMS key ID** choose the key ID.
   - If you are using a key associated with a different account, for **KMS key ARN** enter the ARN for the key ID.

   > **Note**
   >
   > For more information on creating and using KMS keys and the associated encryption, see AWS Key Management Service.

10. Under **Data specifications**, choose the format of your training documents:

    - **CSV file** — A CSV file that supplements your training documents. The CSV file contains information about the custom entities that your trained model will detect. The required format of the file depends on whether you are providing annotations or an entity list.
    - **Augmented manifest** — A labeled dataset that is produced by Amazon SageMaker Ground Truth. This file is in JSON lines format. Each line is a complete JSON object that contains a training document and its labels. Each label annotates a named entity in the training document. You can provide up to 5 augmented manifest files.

For more information about available formats, and for examples, see Training custom entity recognizer models.

11. Under **Training type**, choose the training type to use:

    - **Using annotations and training docs**
    - **Using entity list and training docs**

    If choosing annotations, enter the URL of the annotations file in Amazon S3. You can also navigate to the bucket or folder in Amazon S3 where the annotation files are located and choose **Browse S3**.

    If choosing entity list, enter the URL of the entity list in Amazon S3. You can also navigate to the bucket or folder in Amazon S3 where the entity list is located and choose **Browse S3**.

12. Enter the URL of an input dataset containing the training documents in Amazon S3. You can also navigate to the bucket or folder in Amazon S3 where the training documents are located and choose **Select folder**.
13. Under **Test dataset** select how you want to evaluate the performance of your trained model - you can do this for both annotations and entity list training types.

    - **Autosplit**: Autosplit automatically selects 10% of your provided training data to use as testing data
    - (Optional) **Customer provided**: When you select customer provided, you can specify exactly what test data you want to use.

14. If you select Customer provided test dataset, enter the URL of the annotations file in Amazon S3. You can also navigate to the bucket or folder in Amazon S3 where the annotation files are located and choose **Select folder**.
15. In the **Choose an IAM role** section, either select an existing IAM role or create a new one.

    - **Choose an existing IAM role** – Select this option if you already have an IAM role with permissions to access the input and output Amazon S3 buckets.
    - **Create a new IAM role** – Select this option when you want to create a new IAM role with the proper permissions for Amazon Comprehend to access the input and output buckets.

> **Note**
>
> If the input documents are encrypted, the IAM role used must have `kms:Decrypt` permission. For more information, see Permissions required to use KMS encryption.

16. (Optional) To launch your resources into Amazon Comprehend from a VPC, enter the VPC ID under **VPC** or choose the ID from the drop-down list.

    1. Choose the subnet under **Subnet(s)**. After you select the first subnet, you can choose additional ones.
    2. Under **Security Group(s)**, choose the security group to use if you specified one. After you select the first security group, you can choose additional ones.

> **Note**
>
> When you use a VPC with your custom entity recognition job, the `DataAccessRole` used for the Create and Start operations must have permissions to the VPC from which the input documents and the output bucket are accessed.

17. (Optional) To add a tag to the custom entity recognizer, enter a key-value pair under **Tags**. Choose **Add tag**. To remove this pair before creating the recognizer, choose **Remove tag**.
18. Choose **Train**.

The new recognizer will then appear in the list, showing its status. It will first show as `Submitted`. It will then show `Training` for a classifier that is processing training documents, `Trained` for a classifier that is ready to use, and `In error` for a classifier that has an error. You can click on a job to get more information about the recognizer, including any error messages.

Creating a custom entity recognizer using the console - augmented manifest

To train a custom entity recognizer with a plaintext, PDF, or word document

1. Sign in to the AWS Management Console and open the Amazon Comprehend console.
2. From the left menu, choose **Customization** and then choose **Custom entity recognition**.
3. Choose **Train recognizer**.
4. Give the recognizer a name. The name must be unique within the Region and account.

5. Select the language. Note: If you're training a PDF or Word document, English is the supported language.
6. Under **Custom entity type**, enter a custom label that you want the recognizer to find in the dataset.

   The entity type must be uppercase, and if it consists of more than one word, separate the words with an underscore.
7. Choose **Add type**.
8. If you want to add an additional entity type, enter it, and then choose **Add type**. If you want to remove one of the entity types you've added, choose **Remove type** and then choose the entity type to remove from the list. A maximum of 25 entity types can be listed.
9. To encrypt your training job, choose **Recognizer encryption** and then choose whether to use a KMS key associated with the current account, or one from another account.

   - If you are using a key associated with the current account, for **KMS key ID** choose the key ID.
   - If you are using a key associated with a different account, for **KMS key ARN** enter the ARN for the key ID.

   > **Note**
   >
   > For more information on creating and using KMS keys and the associated encryption, see AWS Key Management Service.

10. Under **Training data**, choose **Augmented manifest** as your data format:

    - **Augmented manifest** — is a labeled dataset that is produced by Amazon SageMaker Ground Truth. This file is in JSON lines format. Each line in the file is a complete JSON object that contains a training document and its labels. Each label annotates a named entity in the training document. You can provide up to 5 augmented manifest files. If you are using PDF documents for training data, you must select **Augmented manifest**. You can provide up to 5 augmented manifest files. For each file, you can name up to 5 attributes to use as training data.

    For more information about available formats, and for examples, see Training custom entity recognizer models.
11. Select the training model type.

If you selected **Plaintext documents**, under **Input location**, enter the Amazon S3URL of the Amazon SageMakerGround Truth augmented manifest file. You can also navigate to the bucket or folder in Amazon S3 where the augmented manifest(s) is located and choose **Select folder**.

12. Under **Attribute name**, enter the name of the attribute that contains your annotations. If the file contains annotations from multiple chained labeling jobs, add an attribute for each job. In this case, each attribute contains the set of annotations from a labeling job. Note: You can provide up to 5 attribute names for each file.
13. Select **Add**.
14. If you selected **PDF, Word documents** under **Input location**, enter the Amazon S3URL of the Amazon SageMaker Ground Truth augmented manifest file. You can also navigate to the bucket or folder in Amazon S3 where the augmented manifest(s) is located and choose **Select folder**.
15. Enter the S3 prefix for your **Annotation** data files. These are the PDF documents that you labeled.
16. Enter the S3 prefix for your **Source** documents. These are the original PDF documents (data objects) that you provided to Ground Truth for your labeling job.
17. Enter the attribute names that contain your annotations. Note: You can provide up to 5 attribute names for each file. Any attributes in your file that you don't specify are ignored.
18. In the IAM role section, either select an existing IAM role or create a new one.

    - **Choose an existing IAM role** – Select this option if you already have an IAM role with permissions to access the input and output Amazon S3 buckets.
    - **Create a new IAM role** – Select this option when you want to create a new IAM role with the proper permissions for Amazon Comprehend to access the input and output buckets.

    > **Note**
    >
    > If the input documents are encrypted, the IAM role used must have `kms:Decrypt` permission. For more information, see Permissions required to use KMS encryption.

19. (Optional) To launch your resources into Amazon Comprehend from a VPC, enter the VPC ID under **VPC** or choose the ID from the drop-down list.

    1. Choose the subnet under **Subnet(s)**. After you select the first subnet, you can choose additional ones.

2. Under **Security Group(s)**, choose the security group to use if you specified one. After you select the first security group, you can choose additional ones.

> **Note**
>
> When you use a VPC with your custom entity recognition job, the `DataAccessRole` used for the Create and Start operations must have permissions to the VPC from which the input documents and the output bucket are accessed.

20. (Optional) To add a tag to the custom entity recognizer, enter a key-value pair under **Tags**. Choose **Add tag**. To remove this pair before creating the recognizer, choose **Remove tag**.
21. Choose **Train**.

The new recognizer will then appear in the list, showing its status. It will first show as `Submitted`. It will then show `Training` for a classifier that is processing training documents, `Trained` for a classifier that is ready to use, and `In error` for a classifier that has an error. You can click on a job to get more information about the recognizer, including any error messages.

Train custom entity recognizers (API)

To create and train a custom entity recognition model, use the Amazon Comprehend CreateEntityRecognizer API operation

Topics

- Training custom entity recognizers using the AWS Command Line Interface
- Training custom entity recognizers using the AWS SDK for Java
- Training custom entity recognizers using Python (Boto3)

Training custom entity recognizers using the AWS Command Line Interface

The following examples demonstrate using the `CreateEntityRecognizer` operation and other associated APIs with the AWS CLI.

The examples are formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Create a custom entity recognizer using the `create-entity-recognizer` CLI command. For information about the input-data-config parameter, see CreateEntityRecognizer in the *Amazon Comprehend API Reference*.

```
aws comprehend create-entity-recognizer \
     --language-code en \
     --recognizer-name test-6 \
     --data-access-role-arn "arn:aws:iam::account number:role/service-role/
AmazonComprehendServiceRole-role" \
     --input-data-config "EntityTypes=[{Type=PERSON}],Documents={S3Uri=s3://Bucket
 Name/Bucket Path/documents},
                Annotations={S3Uri=s3://Bucket Name/Bucket Path/annotations}" \
     --region region
```

List all entity recognizers in a Region using the `list-entity-recognizers` CLI command..

```
aws comprehend list-entity-recognizers \
     --region region
```

Check Job Status of custom entity recognizers using the `describe-entity-recognizer` CLI command..

```
aws comprehend describe-entity-recognizer \
     --entity-recognizer-arn arn:aws:comprehend:region:account number:entity-
recognizer/test-6 \
     --region region
```

Training custom entity recognizers using the AWS SDK for Java

This example creates a custom entity recognizer and trains the model, using Java

For Amazon Comprehend examples that use Java, see Amazon Comprehend Java examples.

Training custom entity recognizers using Python (Boto3)

Instantiate Boto3 SDK:

```
import boto3
import uuid
```

```
comprehend = boto3.client("comprehend", region_name="region")
```

Create entity recognizer:

```
response = comprehend.create_entity_recognizer(
    RecognizerName="Recognizer-Name-Goes-Here-{}".format(str(uuid.uuid4())),
    LanguageCode="en",
    DataAccessRoleArn="Role ARN",
    InputDataConfig={
        "EntityTypes": [
            {
                "Type": "ENTITY_TYPE"
            }
        ],
        "Documents": {
            "S3Uri": "s3://Bucket Name/Bucket Path/documents"
        },
        "Annotations": {
            "S3Uri": "s3://Bucket Name/Bucket Path/annotations"
        }
    }
)
recognizer_arn = response["EntityRecognizerArn"]
```

List all recognizers:

```
response = comprehend.list_entity_recognizers()
```

Wait for recognizer to reach TRAINED status:

```
while True:
    response = comprehend.describe_entity_recognizer(
        EntityRecognizerArn=recognizer_arn
    )

    status = response["EntityRecognizerProperties"]["Status"]
    if "IN_ERROR" == status:
        sys.exit(1)
    if "TRAINED" == status:
        break

    time.sleep(10)
```

Custom entity recognizer metrics

Amazon Comprehend provides you with metrics to help you estimate how well an entity recognizer should work for your job. They are based on training the recognizer model, and so while they accurately represent the performance of the model during training, they are only an approximation of the API performance during entity discovery.

Metrics are returned any time metadata from a trained entity recognizer is returned.

Amazon Comprehend supports training a model on up to 25 entities at a time. When metrics are returned from a trained entity recognizer, scores are computed against both the recognizer as a whole (global metrics) and for each individual entity (entity metrics).

Three metrics are available, both as global and entity metrics:

- **Precision**

  This indicates the fraction of entities produced by the system that are correctly identified and correctly labeled. This shows how many times the model's entity identification is truly a good identification. It is a percentage of the total number of identifications.

  In other words, precision is based on *true positives (tp)* and *false positives (fp)* and it is calculated as *precision = tp / (tp + fp)*.

  For example, if a model predicts that two examples of an entity are present in a document, where there's actually only one, the result is one true positive and one false positive. In this case, *precision = 1 / (1 + 1)*. The precision is 50%, as one entity is correct out of the two identified by the model.

- **Recall**

  This indicates the fraction of entities present in the documents that are correctly identified and labeled by the system. Mathematically, this is defined in terms of the total number of correct identifications *true positives (tp)* and missed identifcations *false negatives (fn)*.

  It is calculated as *recall = tp / (tp + fn)*. For example if a model correctly identifies one entity, but misses two other instances where that entity is present, the result is one true positive and two false negatives. In this case, *recall = 1 / (1 + 2)*. The recall is 33.33%, as one entity is correct out of a possible three examples.

- **F1 score**

  This is a combination of the Precision and Recall metrics, which measures the overall accuracy of the model for custom entity recognition. The F1 score is the harmonic mean of the Precision and Recall metrics: *F1 = 2 * Precision * Recall / (Precision + Recall)* .

  > **Note**
  >
  > Intuitively, the harmonic mean penalizes the extremes more than the simple average or other means (example: `precision` = 0, `recall` = 1 could be achieved trivially by predicting all possible spans. Here, the simple average would be 0.5, but F1 would penalize it as 0).

  In the examples above, `precision` = 50% and `recall` = 33.33%, therefore F1 = 2 * 0.5 * 0.3333 / (0.5 + 0.3333). The F1 Score is .3975, or 39.75%.

Global and individual entity metrics

The relationship between global and individual entity metrics can be seen when analyzing the following sentence for entities that are either a *place* or a *person*

```
John Washington and his friend Smith live in San Francisco, work in San Diego, and own
    a house in Seattle.
```

In our example, the model makes the following predictions.

```
John Washington = Person
Smith = Place
San Francisco = Place
San Diego = Place
Seattle = Person
```

However, the predictions should have been the following.

```
John Washington = Person
Smith = Person
San Francisco = Place
San Diego = Place
```

```
Seattle = Place
```

The individual entity metrics for this would be:

```
entity:  Person
  True positive (TP) = 1 (because John Washington is correctly predicted to be a
    Person).
  False positive (FP) = 1 (because Seattle is incorrectly predicted to be a Person,
    but is actually a Place).
  False negative (FN) = 1 (because Smith is incorrectly predicted to be a Place, but
    is actually a Person).
  Precision = 1 / (1 + 1) = 0.5 or 50%
  Recall = 1 / (1+1) = 0.5 or 50%
  F1 Score = 2 * 0.5 * 0.5 / (0.5 + 0.5) = 0.5 or 50%

entity:  Place
  TP = 2 (because San Francisco and San Diego are each correctly predicted to be a
    Place).
  FP = 1 (because Smith is incorrectly predicted to be a Place, but is actually a
    Person).
  FN = 1 (because Seattle is incorrectly predicted to be a Person, but is actually a
    Place).
  Precision = 2 / (2+1) = 0.6667 or 66.67%
  Recall = 2 / (2+1) = 0.6667 or 66.67%
  F1 Score = 2 * 0.6667 * 0.6667 / (0.6667 + 0.6667) = 0.6667 or  66.67%
```

The global metrics for this would be:

Global:

```
Global:
  TP = 3 (because John Washington, San Francisco and San Diego are predicted
 correctly.
    This is also the sum of all individual entity TP).
  FP = 2 (because Seattle is predicted as Person and Smith is predicted as Place. This
    is the sum of all individual entity FP).
  FN = 2 (because Seattle is predicted as Person and Smith is predicted as Place. This
    is the sum of all individual FN).
  Global Precision = 3 / (3+2) = 0.6 or 60%
    (Global Precision = Global TP / (Global TP + Global FP))
  Global Recall = 3 / (3+2) = 0.6 or 60%
    (Global Recall = Global TP / (Global TP + Global FN))
  Global F1Score = 2 * 0.6 * 0.6 / (0.6 + 0.6) = 0.6 or 60%
```

```
    (Global F1Score = 2 * Global Precision *  Global Recall / (Global Precision +
    Global Recall))
```

Improving custom entity recognizer performance

These metrics provide an insight into how accurately the trained model will perform when you use it to identify entities. Here are a few options you can use to improve your metrics if they are lower than your expectations:

1. Depending on whether you use Annotations or Entity lists (plaintext only), make sure to follow the guidelines in the respective documentation to improve data quality. If you observe better metrics after improving your data and re-training the model, you can keep iterating and improving data quality to achieve better model performance.
2. If you are using an Entity List, consider using Annotations instead. Manual annotations can often improve your results.
3. If you are sure there is not a data quality issue, and yet the metrics remain unreasonably low, please submit a support request.

Running real-time custom recognizer analysis

Real-time analysis is useful for applications that process small documents as they arrive. For example, you can detect custom entities in social media posts, support tickets, or customer reviews.

Before you begin

You need a custom entity recognition model (also known as a recognizer) before you can detect custom entities. For more information about these models, see the section called "Training recognizer models".

A recognizer that is trained with plain-text annotations supports entity detection for plain-text documents only. A recognizer that is trained with PDF document annotations supports entity detection for plain-text documents, images, PDF files, and Word documents. For information about the input files, see Inputs for real-time custom analysis.

If you plan to analyze image files or scanned PDF documents, your IAM policy must grant permissions to use two Amazon Textract API methods (DetectDocumentText and

AnalyzeDocument). Amazon Comprehend invokes these methods during text extraction. For an example policy, see Permissions required to perform document analysis actions.

Topics

- Real-time analysis for custom entity recognition (console)
- Real-time analysis for custom entity recognition (API)
- Outputs for real-time analysis

Real-time analysis for custom entity recognition (console)

You can use the Amazon Comprehend console to run real-time analysis with a custom model. First, you create an endpoint to run the real-time analysis. After you create the endpoint, you run the real-time analysis.

For information about provisioning endpoint throughput, and the associated costs, see Using Amazon Comprehend endpoints.

Topics

- Creating an endpoint for custom entity detection
- Running real-time custom entity detection

Creating an endpoint for custom entity detection

To create an endpoint (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Endpoints** and choose the **Create endpoint** button. A **Create endpoint** screen opens.
3. Give the endpoint a name. The name must be unique within the current Region and account.
4. Choose a custom model that you want to attach the new endpoint to. From the dropdown, you can search by model name.

> **Note**
>
> You must create a model before you can attach an endpoint to it. If you don't have a model yet, see Training custom entity recognizer models.

5. (Optional) To add a tag to the endpoint, enter a key-value pair under **Tags** and choose **Add tag**. To remove this pair before creating the endpoint, choose **Remove tag**.
6. Enter the number of inference units (IUs) to assign to the endpoint. Each unit represents a throughput of 100 characters per second for up to two documents per second. For more information about endpoint throughput, see Using Amazon Comprehend endpoints.
7. (Optional) If you are creating a new endpoint, you have the option to use the IU estimator. The estimator can help you determine the number of IUs to request. The number of inference units depends on the throughput or the number of characters that you want to analyze per second.
8. From the **Purchase summary**, review your estimated hourly, daily, and monthly endpoint cost.
9. Select the check box if you understand that your account accrues charges for the endpoint from the time it starts until you delete it.
10. Choose **Create endpoint**.

Running real-time custom entity detection

After you create an endpoint for your custom entity recognizer model, you can run real-time analysis to detect entities in individual documents.

Complete the following steps to detect custom entities in your text by using the Amazon Comprehend console.

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Real-time analysis**.
3. In the **Input text** section, for **Analysis type**, choose **Custom**.
4. For **Select endpoint**, choose the endpoint that is associated with the entity-detection model that you want to use.
5. To specify the input data for analysis, you can input text or upload a file.
   - To enter text:

a. Choose **Input text**.

b. Enter the text that you want to analyze.

- To upload a file:

a. Choose **Upload file** and enter the filename to upload.

b. (Optional) Under **Advanced read actions**, you can override the default actions for text extraction. For details, see Setting text extraction options.

6. Choose **Analyze**. The console displays the output of the analysis, along with a confidence assessment.

Real-time analysis for custom entity recognition (API)

You can use the Amazon Comprehend API to run real-time analysis with a custom model. First, you create an endpoint to run the real-time analysis. After you create the endpoint, you run the real-time analysis.

For information about provisioning endpoint throughput, and the associated costs, see Using Amazon Comprehend endpoints.

Topics

- Creating an endpoint for custom entity detection
- Running real-time custom entity detection

Creating an endpoint for custom entity detection

For information about the costs associated with endpoints, see Using Amazon Comprehend endpoints.

Creating an Endpoint with the AWS CLI

To create an endpoint by using the AWS CLI, use the `create-endpoint` command:

```
$ aws comprehend create-endpoint \
> --desired-inference-units number of inference units \
> --endpoint-name endpoint name \
> --model-arn arn:aws:comprehend:region:account-id:model/example \
```

```
> --tags Key=Key,Value=Value
```

If your command succeeds, Amazon Comprehend responds with the endpoint ARN:

```
{
   "EndpointArn": "Arn"
}
```

For more information about this command, its parameter arguments, and its output, see `create-endpoint` in the AWS CLI Command Reference.

Running real-time custom entity detection

After you create an endpoint for your custom entity recognizer model, you use the endpoint to run the DetectEntities API operation. You can provide text input using the `text` or `bytes` parameter. Enter the other input types using the `bytes` parameter.

For image files and PDF files, you can use the `DocumentReaderConfig` parameter to override the default text extraction actions. For details, see Setting text extraction options.

Detecting entities in text using the AWS CLI

To detect custom entities in text, run the `detect-entities` command with the input text in the `text` parameter.

Example : Use the CLI to detect entities in input text

```
$ aws comprehend detect-entities \
> --endpoint-arn arn \
> --language-code en \
> --text  "Andy Jassy is the CEO of Amazon."
```

If your command succeeds, Amazon Comprehend responds with the analysis. For each entity that Amazon Comprehend detects, it provides the entity type, text, location, and confidence score.

Detecting entities in semi-structured documents using the AWS CLI

To detect custom entities in PDF, Word, or image file, run the `detect-entities` command with the input file in the `bytes` parameter.

Example : Use the CLI to detect entities in an image file

This example shows how to pass in the image file using the `fileb` option to base64 encode the image bytes. For more information, see Binary large objects in the AWS Command Line Interface User Guide.

This example also passes in a JSON file named `config.json` to set the text extraction options.

```
$ aws comprehend detect-entities \
> --endpoint-arn arn \
> --language-code en \
> --bytes fileb://image1.jpg   \
> --document-reader-config file://config.json
```

The **config.json** file contains the following content.

```
{
   "DocumentReadMode": "FORCE_DOCUMENT_READ_ACTION",
   "DocumentReadAction": "TEXTRACT_DETECT_DOCUMENT_TEXT"
}
```

For more information about the command syntax, see DetectEntities in the *Amazon Comprehend API Reference*.

Outputs for real-time analysis

Outputs for text inputs

If you input text using the `Text` parameter, the output consists of an array of entities that the analysis detected. The following example shows an analysis that detected two JUDGE entities.

```
{
        "Entities":
        [
            {
                "BeginOffset": 0,
                "EndOffset": 22,
                "Score": 0.9763959646224976,
                "Text": "John Johnson",
```

```
                "Type": "JUDGE"
            },
            {
                "BeginOffset": 11,
                "EndOffset": 15,
                "Score": 0.9615424871444702,
                "Text": "Thomas Kincaid",
                "Type": "JUDGE"
            }
        ]
    }
```

Outputs for semi-structured inputs

For a semi-structured input document, or a text file, the output can include the following additional fields:

- DocumentMetadata – Extraction information about the document. The metadata includes a list of pages in the document, with the number of characters extracted from each page. This field is present in the response if the request included the `Byte` parameter.
- DocumentType – The document type for each page in the input document. This field is present in the response for a request that included the `Byte` parameter.
- Blocks – Information about each block of text in the input document. Blocks are nested. A page block contains a block for each line of text, which contains a block for each word. This field is present in the response for a request that included the `Byte` parameter.
- BlockReferences – A reference to each block for this entity. This field is present in the response for a request that included the `Byte` parameter. The field is not present for text files.
- Errors – Page-level errors that the system detected while processing the input document. The field is empty if the system encountered no errors.

For descriptions of these output fields, see DetectEntities in the *Amazon Comprehend API Reference*. For more information about the layout elements, see Amazon Textract analysis objects in the Amazon Textract Developer Guide.

The following example shows the output for a one-page scanned PDF input document.

```
{
    "Entities": [{
        "Score": 0.9984670877456665,
```

```
        "Type": "DATE-TIME",
        "Text": "September 4,",
        "BlockReferences": [{
            "BlockId": "42dcaaee-c484-4b5d-9e3f-ae0be928b3e1",
            "BeginOffset": 0,
            "EndOffset": 12,
            "ChildBlocks": [{
                    "ChildBlockId": "6e9cbb43-f8be-4da0-9a4b-ff9a6c350a14",
                    "BeginOffset": 0,
                    "EndOffset": 9
                },
                {
                    "ChildBlockId": "599e0d53-ae9f-491b-a762-459b22c79ff5",
                    "BeginOffset": 0,
                    "EndOffset": 2
                },
                {
                    "ChildBlockId": "599e0d53-ae9f-491b-a762-459b22c79ff5",
                    "BeginOffset": 0,
                    "EndOffset": 2
                }
            ]
        }]
    }],
    "DocumentMetadata": {
        "Pages": 1,
        "ExtractedCharacters": [{
            "Page": 1,
            "Count": 609
        }]
    },
    "DocumentType": [{
        "Page": 1,
        "Type": "SCANNED_PDF"
    }],
    "Blocks": [{
        "Id": "ee82edf3-28de-4d63-8883-40e2e4938ccb",
        "BlockType": "LINE",
        "Text": "Your Band",
        "Page": 1,
        "Geometry": {
            "BoundingBox": {
                "Height": 0.024125460535287857,
                "Left": 0.11745482683181763,
```

```
                "Top": 0.06821706146001816,
                "Width": 0.12074867635965347
            },
            "Polygon": [{
                    "X": 0.11745482683181763,
                    "Y": 0.06821706146001816
                },
                {
                    "X": 0.2382034957408905,
                    "Y": 0.06821706146001816
                },
                {
                    "X": 0.2382034957408905,
                    "Y": 0.09234252572059631
                },
                {
                    "X": 0.11745482683181763,
                    "Y": 0.09234252572059631
                }
            ]
        },
        "Relationships": [{
            "Ids": [
                "b105c561-c8d9-485a-a728-7a5b1a308935",
                "60ecb119-3173-4de2-8c5d-de182a5f86a5"
            ],
            "Type": "CHILD"
        }]
    }]
}
```

The following example shows the output for analysis of a native PDF document.

Example Example output from a custom entity recognition analysis of a PDF document

```
{
        "Blocks":
        [
            {
                "BlockType": "LINE",
                "Geometry":
                {
                    "BoundingBox":
                    {
```

```
                    "Height": 0.012575757575757575,
                    "Left": 0.0,
                    "Top": 0.0015063131313131314,
                    "Width": 0.02262091503267974
                },
                "Polygon":
                [
                    {
                        "X": 0.0,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.014082070707070706
                    },
                    {
                        "X": 0.0,
                        "Y": 0.014082070707070706
                    }
                ]
            },
            "Id": "4330efed-6334-4fc4-ba48-e050afa95c8d",
            "Page": 1,
            "Relationships":
            [
                {
                    "ids":
                    [
                        "f343ce48-583d-4abe-b84b-a232e266450f"
                    ],
                    "type": "CHILD"
                }
            ],
            "Text": "S-3"
        },
        {
            "BlockType": "WORD",
            "Geometry":
            {
                "BoundingBox":
```

```
                {
                    "Height": 0.012575757575757575,
                    "Left": 0.0,
                    "Top": 0.0015063131313131314,
                    "Width": 0.02262091503267974
                },
                "Polygon":
                [
                    {
                        "X": 0.0,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.014082070707070706
                    },
                    {
                        "X": 0.0,
                        "Y": 0.014082070707070706
                    }
                ]
            },
            "Id": "f343ce48-583d-4abe-b84b-a232e266450f",
            "Page": 1,
            "Relationships":
            [],
            "Text": "S-3"
        }
    ],
    "DocumentMetadata":
    {
        "PageNumber": 1,
        "Pages": 1
    },
    "DocumentType": "NativePDF",
    "Entities":
    [
        {
            "BlockReferences":
            [
```

```
                    {
                        "BeginOffset": 25,
                        "BlockId": "4330efed-6334-4fc4-ba48-e050afa95c8d",
                        "ChildBlocks":
                        [
                            {
                                "BeginOffset": 1,
                                "ChildBlockId": "cbba5534-ac69-4bc4-beef-306c659f70a6",
                                "EndOffset": 6
                            }
                        ],
                        "EndOffset": 30
                    }
                ],
                "Score": 0.9998825926329088,
                "Text": "0.001",
                "Type": "OFFERING_PRICE"
            },
            {
                "BlockReferences":
                [
                    {
                        "BeginOffset": 41,
                        "BlockId": "f343ce48-583d-4abe-b84b-a232e266450f",
                        "ChildBlocks":
                        [
                            {
                                "BeginOffset": 0,
                                "ChildBlockId": "292a2e26-21f0-401b-a2bf-03aa4c47f787",
                                "EndOffset": 9
                            }
                        ],
                        "EndOffset": 50
                    }
                ],
                "Score": 0.9809727537330395,
                "Text": "6,097,560",
                "Type": "OFFERED_SHARES"
            }
        ],
        "File": "example.pdf",
        "Version": "2021-04-30"
    }
```

Running analysis jobs for custom entity recognition

You can run an asynchronous analysis job to detect custom entities in a set of one or more documents.

Before you begin

You need a custom entity recognition model (also known as a recognizer) before you can detect custom entities. For more information about these models, see the section called "Training recognizer models".

A recognizer that is trained with plain-text annotations supports entity detection for plain-text documents only. A recognizer that is trained with PDF document annotations supports entity detection for plain-text documents, images, PDF files, and Word documents. For files other than text files, Amazon Comprehend performs text extraction before running the analysis. For information about the input files, see Inputs for asynchronous custom analysis.

If you plan to analyze image files or scanned PDF documents, your IAM policy must grant permissions to use two Amazon Textract API methods (DetectDocumentText and AnalyzeDocument). Amazon Comprehend invokes these methods during text extraction. For an example policy, see Permissions required to perform document analysis actions.

To run an async analysis job, you perform the following overall steps:

1. Store the documents in an Amazon S3 bucket.
2. Use the API or console to start the analysis job.
3. Monitor the progress of the analysis job.
4. After the job runs to completion, retrieve the results of the analysis from the S3 bucket that you specified when you started the job.

Topics

Starting a custom entity detection job (console)

You can use the console to start and monitor an async analysis job for custom entity recognition.

To start an async analysis job

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Analysis jobs** and then choose **Create job**.
3. Give the classification job a name. The name must be unique your account and current Region.
4. Under **Analysis type**, choose **Custom entity recognition**.
5. From **Recognizer model**, choose the custom entity recognizer to use.
6. From **Version**, choose the recognizer version to use.
7. (Optional) If you choose to encrypt the data that Amazon Comprehend uses while processing your job, choose **Job encryption**. Then choose whether to use a KMS key associated with the current account, or one from another account.
   - If you are using a key associated with the current account, choose the key ID for **KMS key ID**.
   - If you are using a key associated with a different account, enter the ARN for the key ID under **KMS key ARN**.

   > **Note**
   >
   > For more information on creating and using KMS keys and the associated encryption, see Key management service (KMS).

8. Under **Input data**, enter the location of the Amazon S3 bucket that contains your input documents or navigate to it by choosing **Browse S3**. This bucket must be in the same Region as the API that you are calling. The IAM role that you're using for access permissions for the analysis job must have reading permissions for the S3 bucket.
9. (Optional) for **Input format**, you can choose the format of the input documents. The format can be one document per file, or one document per line in a single file. One document per line applies only to text documents.
10. (Optional) For **Document read mode**, you can override the default text extraction actions. For more information, see Setting text extraction options.

11. Under **Output data**, enter the location of the Amazon S3 bucket where Amazon Comprehend should write the job's output data or navigate to it by choosing **Browse S3**. This bucket must be in the same Region as the API that you are calling. The IAM role you're using for access permissions for the classification job must have write permissions for the S3 bucket.
12. (Optional) if you choose to encrypt the output result from your job, choose **Encryption**. Then choose whether to use a KMS key associated with the current account, or one from another account.

    - If you are using a key associated with the current account, choose the key alias or ID for **KMS key ID**.
    - If you are using a key associated with a different account, enter the ARN for the key alias or ID under **KMS key ID**.
13. (Optional) To launch your resources into Amazon Comprehend from a VPC, enter the VPC ID under **VPC** or choose the ID from the drop-down list.

    1. Choose the subnet under **Subnet(s)**. After you select the first subnet, you can choose additional ones.
    2. Under **Security Group(s)**, choose the security group to use if you specified one. After you select the first security group, you can choose additional ones.

    > **Note**
    >
    > When you use a VPC with your analysis job, the `DataAccessRole` used for the Create and Start operations must have permissions to the VPC that accesses the output bucket.

14. Choose **Create job** to create the entity recognition job.

Starting a custom entity detection job (API)

You can use the API to start and monitor an async analysis job for custom entity recognition.

To start a custom entity detection job with the StartEntitiesDetectionJob operation, you provide the EntityRecognizerArn, which is the Amazon Resource Name (ARN) of the trained model. You can find this ARN in the response to the CreateEntityRecognizer operation.

Topics

- Detecting custom entities using the AWS Command Line Interface
- Detecting custom entities using the AWS SDK for Java
- Detecting custom entities using the AWS SDK for Python (Boto3)
- Overriding API actions for PDF files

Detecting custom entities using the AWS Command Line Interface

Use the following example for Unix, Linux, and macOS environments. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^). To detect custom entities in a document set, use the following request syntax:

```
aws comprehend start-entities-detection-job \
     --entity-recognizer-arn "arn:aws:comprehend:region:account number:entity-
recognizer/test-6" \
     --job-name infer-1 \
     --data-access-role-arn "arn:aws:iam::account number:role/service-role/
AmazonComprehendServiceRole-role" \
     --language-code en \
     --input-data-config "S3Uri=s3://Bucket Name/Bucket Path" \
     --output-data-config "S3Uri=s3://Bucket Name/Bucket Path/" \
     --region region
```

Amazon Comprehend responds with the `JobID` and `JobStatus` and will return the output from the job in the S3 bucket that you specified in the request.

Detecting custom entities using the AWS SDK for Java

For Amazon Comprehend examples that use Java, see Amazon Comprehend Java examples.

Detecting custom entities using the AWS SDK for Python (Boto3)

This example creates a custom entity recognizer, trains the model, and then runs it in an entity recognizer job using the AWS SDK for Python (Boto3).

Instantiate the SDK for Python.

```
import boto3
import uuid
```

```
comprehend = boto3.client("comprehend", region_name="region")
```

Create an entity recognizer:

```
response = comprehend.create_entity_recognizer(
    RecognizerName="Recognizer-Name-Goes-Here-{}".format(str(uuid.uuid4())),
    LanguageCode="en",
    DataAccessRoleArn="Role ARN",
    InputDataConfig={
        "EntityTypes": [
            {
                "Type": "ENTITY_TYPE"
            }
        ],
        "Documents": {
            "S3Uri": "s3://Bucket Name/Bucket Path/documents"
        },
        "Annotations": {
            "S3Uri": "s3://Bucket Name/Bucket Path/annotations"
        }
    }
)
recognizer_arn = response["EntityRecognizerArn"]
```

List all recognizers:

```
response = comprehend.list_entity_recognizers()
```

Wait for the entity recognizer to reach TRAINED status:

```
while True:
    response = comprehend.describe_entity_recognizer(
        EntityRecognizerArn=recognizer_arn
    )

    status = response["EntityRecognizerProperties"]["Status"]
    if "IN_ERROR" == status:
        sys.exit(1)
    if "TRAINED" == status:
        break

    time.sleep(10)
```

Start a custom entities detection job:

```
response = comprehend.start_entities_detection_job(
    EntityRecognizerArn=recognizer_arn,
    JobName="Detection-Job-Name-{}".format(str(uuid.uuid4())),
    LanguageCode="en",
    DataAccessRoleArn="Role ARN",
    InputDataConfig={
        "InputFormat": "ONE_DOC_PER_LINE",
        "S3Uri": "s3://Bucket Name/Bucket Path/documents"
    },
    OutputDataConfig={
        "S3Uri": "s3://Bucket Name/Bucket Path/output"
    }
)
```

Overriding API actions for PDF files

For image files and PDF files, you can override the default extraction actions using the `DocumentReaderConfig` parameter in `InputDataConfig`.

The following example defines a JSON file named myInputDataConfig.json to set the `InputDataConfig` values. It sets `DocumentReadConfig` to use the Amazon Textract `DetectDocumentText` API for all PDF files.

Example

```
"InputDataConfig": {
  "S3Uri": s3://Bucket Name/Bucket Path",
  "InputFormat": "ONE_DOC_PER_FILE",
  "DocumentReaderConfig": {
      "DocumentReadAction": "TEXTRACT_DETECT_DOCUMENT_TEXT",
      "DocumentReadMode": "FORCE_DOCUMENT_READ_ACTION"
  }
}
```

In the `StartEntitiesDetectionJob` operation, specify the myInputDataConfig.json file as the `InputDataConfig` parameter:

```
  --input-data-config file://myInputDataConfig.json
```

For more information about the `DocumentReaderConfig` parameters, see Setting text extraction options.

Outputs for asynchronous analysis jobs

After an analysis job completes, it stores the results in the S3 bucket that you specified in the request.

Outputs for text inputs

For text input files, the output consists of a list of entities for each input document.

The following example shows the output for two documents from an input file named **50_docs**, using one document per line format.

```
{
        "File": "50_docs",
        "Line": 0,
        "Entities":
        [
            {
                "BeginOffset": 0,
                "EndOffset": 22,
                "Score": 0.9763959646224976,
                "Text": "John Johnson",
                "Type": "JUDGE"
            }
        ]
    }
    {
        "File": "50_docs",
        "Line": 1,
        "Entities":
        [
            {
                "BeginOffset": 11,
                "EndOffset": 15,
                "Score": 0.9615424871444702,
                "Text": "Thomas Kincaid",
                "Type": "JUDGE"
            }
        ]
    }
```

Outputs for semi-structured inputs

For semi-structured input documents, the output can include the following additional fields:

- DocumentMetadata – Extraction information about the document. The metadata includes a list of pages in the document, with the number of characters extracted from each page. This field is present in the response if the request included the `Byte` parameter.
- DocumentType – The document type for each page in the input document. This field is present in the response for a request that included the `Byte` parameter.
- Blocks – Information about each block of text in the input document. Blocks can nest within a block. A page block contains a block for each line of text, which contains a block for each word. This field is present in the response for a request that included the `Byte` parameter.
- BlockReferences – A reference to each block for this entity. This field is present in the response for a request that included the `Byte` parameter. The field isn't present for text files.
- Errors – Page-level errors that the system detected while processing the input document. The field is empty if the system encountered no errors.

For more details about these output fields, see DetectEntities in the *Amazon Comprehend API Reference*

The following example shows the output for a one-page native PDF input document.

Example Example output from a custom entity recognition analysis of a PDF document


```
{
        "Blocks":
        [
            {
                "BlockType": "LINE",
                "Geometry":
                {
                    "BoundingBox":
                    {
                        "Height": 0.012575757575757575,
                        "Left": 0.0,
                        "Top": 0.0015063131313131314,
                        "Width": 0.02262091503267974
                    },
                    "Polygon":
                    [
```

```
                    {
                        "X": 0.0,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.014082070707070706
                    },
                    {
                        "X": 0.0,
                        "Y": 0.014082070707070706
                    }
                ]
            },
            "Id": "4330efed-6334-4fc4-ba48-e050afa95c8d",
            "Page": 1,
            "Relationships":
            [
                {
                    "ids":
                    [
                        "f343ce48-583d-4abe-b84b-a232e266450f"
                    ],
                    "type": "CHILD"
                }
            ],
            "Text": "S-3"
        },
        {
            "BlockType": "WORD",
            "Geometry":
            {
                "BoundingBox":
                {
                    "Height": 0.012575757575757575,
                    "Left": 0.0,
                    "Top": 0.0015063131313131314,
                    "Width": 0.02262091503267974
                },
                "Polygon":
```

```
                [
                    {
                        "X": 0.0,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.0015063131313131314
                    },
                    {
                        "X": 0.02262091503267974,
                        "Y": 0.014082070707070706
                    },
                    {
                        "X": 0.0,
                        "Y": 0.014082070707070706
                    }
                ]
            },
            "Id": "f343ce48-583d-4abe-b84b-a232e266450f",
            "Page": 1,
            "Relationships":
            [],
            "Text": "S-3"
        }
    ],
    "DocumentMetadata":
    {
        "PageNumber": 1,
        "Pages": 1
    },
    "DocumentType": "NativePDF",
    "Entities":
    [
        {
            "BlockReferences":
            [
                {
                    "BeginOffset": 25,
                    "BlockId": "4330efed-6334-4fc4-ba48-e050afa95c8d",
                    "ChildBlocks":
                    [
                        {
                            "BeginOffset": 1,
```

```
                            "ChildBlockId": "cbba5534-ac69-4bc4-beef-306c659f70a6",
                            "EndOffset": 6
                        }
                    ],
                    "EndOffset": 30
                }
            ],
            "Score": 0.9998825926329088,
            "Text": "0.001",
            "Type": "OFFERING_PRICE"
        },
        {
            "BlockReferences":
            [
                {
                    "BeginOffset": 41,
                    "BlockId": "f343ce48-583d-4abe-b84b-a232e266450f",
                    "ChildBlocks":
                    [
                        {
                            "BeginOffset": 0,
                            "ChildBlockId": "292a2e26-21f0-401b-a2bf-03aa4c47f787",
                            "EndOffset": 9
                        }
                    ],
                    "EndOffset": 50
                }
            ],
            "Score": 0.9809727537330395,
            "Text": "6,097,560",
            "Type": "OFFERED_SHARES"
        }
    ],
    "File": "example.pdf",
    "Version": "2021-04-30"
}
```

Creating and managing custom models

Amazon Comprehend includes built-in NLP (natural language processing) models that you can use for analyzing insights or topic modeling. You can also use Amazon Comprehend to create custom models for entity recognition and document classification.

You can use model versioning to keep track of your model's history. When you create and train a new model version, you can make changes to the training dataset. Amazon Comprehend displays details (including model performance) for each model version on the model details page. Over time, you can see how model performance changes as you make changes to your training dataset.

You can create model versions using the Amazon Comprehend console or API. As an alternative, Amazon Comprehend provides Flywheels to simplify the tasks associated with training and evaluating new custom model versions.

After you create a custom model, you can share the model with other users by allowing other AWS accounts to import a copy of your model.

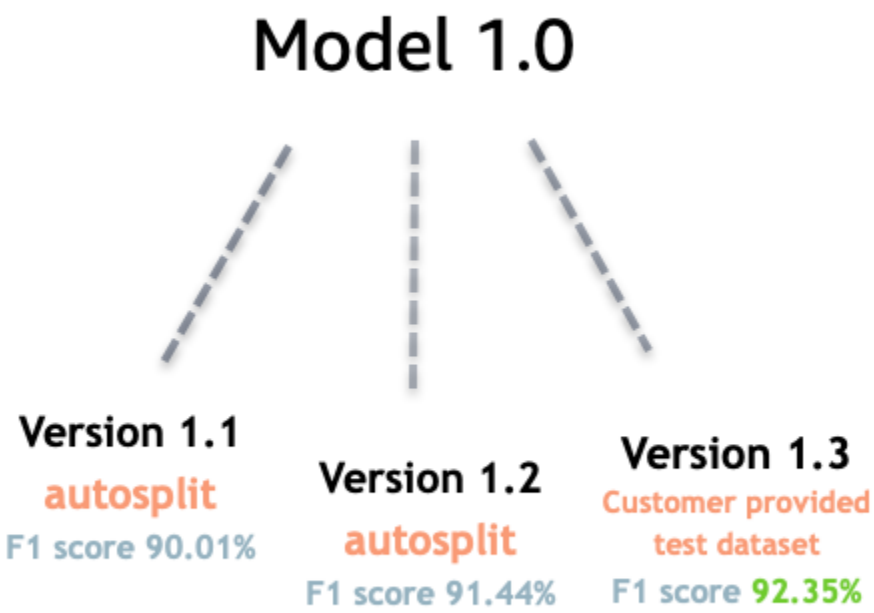
Topics

Model versioning with Amazon Comprehend

Artificial intelligence and machine learning (AI/ML) is all about rapid experimentation. With Amazon Comprehend, you train and build out models which you use to gain insight on your data. With model versioning you can keep track of your modeling history and scores associated with running results of your models as you provide more or different sets of data. You can use versioning with your custom classification models or your custom entity recognition models. Taking a look at your different versions over time you can gain insight on how successful they've performed and gain insight on what parameters you used to get to your state of success.

When you train a new version of an existing custom classifier model or entity recognition model, all you need to do is create a new version from the model details page and all the details populate for you. The new version will have the same name as your earlier model — what we call the versionID — although you will give it a unique version name during creation. As you add new versions to a model, you can see all the previous versions and their details in one view from the model details

page. With versioning, you can see how model performance changes as you make changes to your training dataset.



Create a new Custom classifier version (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Customization** and then choose **Custom classification**.
3. From the **Classifiers** list, choose the name of the custom model from which you want to create a new version. The custom model details page is displayed.
4. On the top right, select **Create new model.** A screen opens with prepopulated details from the parent custom classification model.
5. Under **Version name** add a unique name to the new version.
6. Under version details, you can change the language and number of labels associated with your new model.
7. Under the **Data specifications** section configure how you want to provide the data to your new version— make sure to provide full data, which includes documents from your previous model and your new documents. You can change the **Classifier mode** (single-label, or multi-label),

Data format (CSV file, Augmented manifest), your **Training dataset**, and your **Test dataset** (autosplit, or your custom test data configuration).

8. (Optional) update the S3 location for your output data
9. Under **Access permissions**, create or use an existing IAM role.
10. (Optional) Update your VPC settings
11. (Optional) Add tags to your new version to help keep track of the details.

For more information about creating custom classifiers, see Create a Custom Classifier

Create a new Custom entity recognizer version (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Customization** and then choose **Custom entity recognition**.
3. From the **Recognizer model** list, choose the name of the recognizer from which you want to create a new version. The details page is displayed.
4. On the top right, select **Train new version.** A screen opens with prepopulated details from the parent entity recognizer.
5. Under **Version name** add a unique name to the new version.
6. Under Custom entity type, add the custom labels or label you want the recognizer to identify in your dataset and select **Add type**. Choose a custom entity type from the annotations or entity list you've provided. The recognizer will then use all of the included entity types to identify entities in the data set when running your job. Each entity type must be upper-case and separated by and underscore if it uses multiple words. A maximum of 25 types are allowed.
7. (Optional) Select **Recognizer encryption** to encrypt the data in the storage volume while your job is being processed.
8. Under the Training data section, specify the **Annotation and data format** details (CSV file, Augmented manifest)single-label, or multi-label), **Data format** (CSV, Augmented manifest), your **Training dataset**, and your **Test dataset** (autosplit, or your custom test data configuration).
9. (Optional) update the S3 location for your output data
10. Under **Access permissions**, create or use an existing IAM role.
11. (Optional) Update your VPC settings

12. (Optional) Add tags to your new version to help keep track of the details.

To learn more about custom entity recognizers, see Custom Entity Recognition and Creating a Custom Entity Recognizer Using the Console.

Copying custom models between AWS accounts

Amazon Comprehend users can copy trained custom models between AWS accounts in a two-step process. First, a user in one AWS account (account A), *shares* a custom model that's in their account. Then, a user in another AWS account (account B) *imports* the model into their account. The account B user does not need to train the model, and does not need to copy (or access) the original training data or test data.

To share a custom model in account A, the user attaches an AWS Identity and Access Management (IAM) policy to a model version. This policy authorizes an entity in account B, such as a user or role, to import the model version into Amazon Comprehend in their AWS account. The account B user must import the model into the same AWS Region as the original model.

To import the model in account B, the user of this account provides Amazon Comprehend with the necessary details, such as the Amazon Resource Name (ARN) of the model. By importing the model, this user creates a new custom model in their AWS account that replicates the model that they imported. This model is fully trained and ready for inference jobs, such as document classification or named entity recognition.

Copying a custom model is useful if:

- You belong to an organization that uses multiple AWS accounts. For example, your organization might have an AWS account for each phase of development, such as build, stage, test, and deploy. Or, it might have distinct AWS accounts for business functions, such as data science and engineering.
- Your organization works with another, such as an AWS Partner, that trains custom models in Amazon Comprehend and provides them to you as their client.

In scenarios like these, you can quickly copy a trained custom entity recognizer or document classifier from one AWS account to another. Copying a model in this way is easier than the alternative, where you copy training data between AWS accounts to train duplicate models.

Topics

Sharing a custom model with another AWS account

With Amazon Comprehend, you can share your custom models with others, so they can import your models into their AWS accounts. When a user imports one of your custom models, they create a new custom model in their account. Their new model duplicates the one that you shared.

To share a custom model, you attach a policy to it that authorizes others to import it. Then, you provide those users with the details that they need.

> **Note**
>
> When other users import a custom model that you've shared, they must use the same AWS Region —for example, US East (N. Virginia)— that contains your model.

Topics

Before you begin

Before you can share a model, you must have a trained custom classifier or custom entity recognizer in Amazon Comprehend in your AWS account. For more information about training custom models, see Custom classification or Custom entity recognition.

Required permissions

IAM policy statement

Before you can add a resource-based policy to a custom model, you require permissions in AWS Identity and Access Management (IAM). Your user, group, or role must have a policy attached so you can create, get, and delete model policies, as shown in the following example.

Example IAM policy to manage resource-based policies for custom models

```
{
  "Effect": "Allow",
  "Action": [
    "comprehend:PutResourcePolicy",
    "comprehend:DeleteResourcePolicy",
    "comprehend:DescribeResourcePolicy"
  ],
  "Resource": "arn:aws:comprehend:us-west-2:111122223333:document-classifier/foo/
version/*"
}
```

For information about creating an IAM policy, see Creating IAM policies in the *IAM User Guide*. For information about attaching an IAM policy, see Adding and removing IAM identity permissions in the *IAM User Guide*.

AWS KMS key policy statement

If you are sharing an encrypted model, then you might need to add permissions for AWS KMS. This requirement depends on the type of KMS key that you use to encrypt the model in Amazon Comprehend.

An **AWS owned key** is owned and managed by an AWS service. If you use an AWS owned key, you do not need to add permissions for AWS KMS, and you can skip this section.

A **Customer managed key** is a key that you create, own, and manage in your AWS account. If you use a customer managed key, you must add a statement to your KMS key policy.

The policy statement authorizes one or more entities (such as users or accounts) to perform the AWS KMS operations required to decrypt the model.

You use condition keys to help prevent the confused deputy problem. For more information, see the section called “Cross-service confused deputy prevention”.

Use the following condition keys in the policy to validate the entities that access your KMS key. When a user imports the model, AWS KMS checks that the ARN of the source model version matches the condition. If you do not include a condition in the policy, the specified principals can use your KMS key to decrypt any model version:

- aws:SourceArn – Use this condition key with the `kms:GenerateDataKey` and `kms:Decrypt` actions.

- kms:EncryptionContext – Use this condition key with the `kms:GenerateDataKey`, `kms:Decrypt`, and `kms:CreateGrant` actions.

In the following example, the policy authorizes AWS account 444455556666 to use version 1 of the specified classifier model owned by AWS account 111122223333.

Example KMS key policy to access a specific classifier model version

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
          "AWS":
                "arn:aws:iam::444455556666:root"
      },
      "Action": [
          "kms:Decrypt",
          "kms:GenerateDataKey"
      ],
      "Resource": "*",
      "Condition": {
          "StringEquals": {
              "aws:SourceArn":
                "arn:aws:comprehend:us-west-2:111122223333:document-
classifier/classifierName/version/1"
          }
      }
    },
    {
      "Effect": "Allow",
      "Principal": {
          "AWS":  "arn:aws:iam::444455556666:root"
      },
      "Action": "kms:CreateGrant",
      "Resource": "*",
      "Condition": {
        "StringEquals": {
           "kms:EncryptionContext:aws:comprehend:arn":
              "arn:aws:comprehend:us-west-2:111122223333:document-
classifier/classifierName/version/1"
```

```
        }
      }
    }
  ]
}
```

The following example policy authorizes user **ExampleUser** from AWS account 444455556666 and **ExampleRole** from AWS account 123456789012 to access this KMS key via the Amazon Comprehend service.

Example KMS key policy to allow access to the Amazon Comprehend service (alternative 1).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
        "Effect": "Allow",
        "Principal": {
            "AWS": [
                "arn:aws:iam::444455556666:user/ExampleUser",
                "arn:aws:iam::123456789012:role/ExampleRole"
              ]
        },
        "Action": [
            "kms:Decrypt",
            "kms:GenerateDataKey"
        ],
        "Resource": "*",
        "Condition": {
            "StringLike": {
                "aws:SourceArn": "arn:aws:comprehend:*"

            }
        }
    },
    {
      "Effect": "Allow",
      "Principal": {
         "AWS": [
                "arn:aws:iam::444455556666:user/ExampleUser",
                "arn:aws:iam::123456789012:role/ExampleRole"
              ]
      },
```

```
      "Action": "kms:CreateGrant",
      "Resource": "*",
      "Condition": {
          "StringLike": {
              "kms:EncryptionContext:aws:comprehend:arn": "arn:aws:comprehend:*"
          }
      }
    }
  ]
}
```

The following example policy authorizes AWS account 444455556666 to access this KMS key via the Amazon Comprehend service, using an alternative syntax to the previous example.

Example KMS key policy to allow access to the Amazon Comprehend service (alternative 2).

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
          "AWS": "arn:aws:iam::444455556666:root"
      },
      "Action": [
          "kms:Decrypt",
          "kms:GenerateDataKey",
          "kms:CreateGrant"
      ],
      "Resource": "*",
      "Condition": {
          "StringLike": {
              "kms:EncryptionContext:aws:comprehend:arn": "arn:aws:comprehend:*"
          }
      }
    }
  ]
}
```

For more information, see Key policies in AWS KMS in the *AWS Key Management Service Developer Guide*.

Resource-based policies for custom models

Before an Amazon Comprehend user in another AWS account can import a custom model from your AWS account, you must authorize them to do so. To authorize them, you add a *resource-based policy* to the model version that you want to share. A resource-based policy is an IAM policy that you attach to a resource in AWS.

When you attach a resource policy to a custom model version, the policy authorizes users, groups, or roles to perform the `comprehend:ImportModel` action on the model version.

Example Resource-based policy for a custom model version

This example specifies the authorized entities in the `Principal` attribute. Resource "*" refers to the specific model version that you attach the policy to.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "comprehend:ImportModel",
      "Resource": "*",
      "Principal": {
        "AWS": [
                "arn:aws:iam::111122223333:root",
                "arn:aws:iam::444455556666:user/ExampleUser",
                "arn:aws:iam::123456789012:role/ExampleRole"
         ]
      }
    }
  ]
}
```

For policies that you attach to custom models, `comprehend:ImportModel` is the only action that Amazon Comprehend supports.

For more information about resource-based policies, see Identity-based policies and resource-based policies in the *IAM User Guide*.

Step 1: Add a resource-based policy to a custom model

You can add a resource-based policy by using the AWS Management Console, AWS CLI, or Amazon Comprehend API.

AWS Management Console

You can use Amazon Comprehend in the AWS Management Console.

To add a resource-based policy

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. In the navigation menu on the left, under **Customization**, choose the page that contains your custom model:
   a. If you are sharing a custom document classifier, choose **Custom classification**.
   b. If you are sharing a custom entity recognizer, choose **Custom entity recognition**.
3. In the list of models, choose the model name to open its details page.
4. Under **Versions**, choose the name of the model version that you want to share.
5. On the version details page, choose the **Tags, VPC & Policy** tab.
6. In the **Resource-based policy** section, choose **Edit**.
7. On the **Edit resource-based policy page**, do the following:
   a. For **Policy name**, enter a name that will help you recognize the policy after you create it.
   b. Under **Authorize**, specify one or more of the following entities to authorize them to import your model:

Field	Definition and examples
Service principals	Service principal identifiers for the services that can access this model version. For example: **comprehend.amazonaws.com**
AWS account IDs	AWS accounts that can access this model version. Authorizes all users who belong to the account. For example:

Field	Definition and examples
	111122223333, **123456789012**
IAM entities	ARNs for users or roles that can access this model version. For example: **arn:aws:iam::111122223333:user/ExampleUser, arn:aws:iam::444455556666:role/ExampleRole**

8. Under **Share**, you can copy the ARN of the model version to help you share it with the person who will import your model. When someone imports a custom model from a different AWS account, the model version ARN is required.
9. Choose **Save**. Amazon Comprehend creates your resource-based policy and attaches it to your model.

AWS CLI

To add a resource-based policy to a custom model with the AWS CLI, use the PutResourcePolicy command. The command takes the following parameters:

- `resource-arn` – The ARN of the custom model, including the model version.
- `resource-policy` – A JSON file that defines the resource-based policy to attach to your custom model.

  You can also provide the policy as an inline JSON string. To provide valid JSON for your policy, enclose the attribute names and values in double quotes. If the JSON body is also enclosed in double quotes, you escape the double quotes that are inside the policy.
- `policy-revision-id` – The revision ID that Amazon Comprehend assigned to the policy that you are updating. If you are creating a new policy that has no prior version, don't use this parameter. Amazon Comprehend creates the revision ID for you.

Example Add a resource-based policy to a custom model using the `put-resource-policy` command

This example defines a policy in a JSON file named **policyFile.json** and associates the policy to a model. The model is version **v2** of a classifier named **mycf1**.

```
$ aws comprehend put-resource-policy \
> --resource-arn arn:aws:comprehend:us-west-2:111122223333:document-classifier/mycf1/
version/v2 \
> --resource-policy file://policyFile.json \
> --policy-revision-id revision-id
```

The JSON file for the resource policy contains the following contents:

- *Action* – The policy authorizes the named principals to use `comprehend:ImportModel`.
- *Resource* – The ARN of the custom model. Resource "*" refers to the model version that you specify in the `put-resource-policy` command.
- *Principal* – The policy authorizes user `jane` from AWS account 444455556666 and all users from AWS account 123456789012.

```
{
"Version":"2012-10-17",
 "Statement":[
    {"Sid":"ResourcePolicyForImportModel",
     "Effect":"Allow",
     "Action":["comprehend:ImportModel"],
     "Resource":"*",
     "Principal":
         {"AWS":
            ["arn:aws:iam::444455556666:user/jane",
             "123456789012"]
         }
   }
 ]
}
```

Amazon Comprehend API

To add a resource-based policy to a custom model by using the Amazon Comprehend API, use the PutResourcePolicy API operation.

You can also add a policy to a custom model in the API request that creates the model. To do this, provide the policy JSON for the ModelPolicy parameter when you submit a CreateDocumentClassifier or CreateEntityRecognizer request.

Step 2: Provide the details that others need to import

Now that you have added the resource-based policy to your custom model, you have authorized other Amazon Comprehend users to import your model into their AWS accounts. However, before they can import, you must provide them with the following details:

- The Amazon Resource Name (ARN) of the model version.
- The AWS Region that contains the model. Anyone who imports your model must use the same AWS Region .
- Whether the model is encrypted, and if it is, the type of AWS KMS key that you use: AWS owned key or customer managed key.
- If your model is encrypted with a customer managed key, then you must provide the ARN of the KMS key. Anyone who imports your model must include the ARN in an IAM service role in their AWS account. This role authorizes Amazon Comprehend to use the KMS key to decrypt the model during the import.

For more information about how other users import your model, see Importing a custom model from another AWS account.

Importing a custom model from another AWS account

In Amazon Comprehend, you can import a custom model that's in another AWS account. When you import a model, you create a new custom model in your account. Your new custom model is a fully-trained duplicate of the model that you imported.

Topics

Before you begin

Before you can import a custom model from another AWS account, ensure that the person who shared the model with you does the following:

- Authorizes you to do the import. This authorization is granted in the resource-based policy that is attached to the model version. For more information, see Resource-based policies for custom models.

- Provides you with the following information:
  - The Amazon Resource Name (ARN) of the model version.
  - The AWS Region that contains the model. You must use the same AWS Region when you import.
  - Whether the model is encrypted with an AWS KMS key and, if it is, the type of key that is used.

If the model is encrypted, you might need to take additional steps, depending on the type of KMS key that is used:

- **AWS owned key** – This type of KMS key is owned and managed by AWS. If the model is encrypted with an AWS owned key, no additional steps are needed.
- **Customer managed key** – This type of KMS key is created, owned, and managed by an AWS customer in their AWS account. If the model is encrypted with a customer managed key, then the person who shared the model must:
  - Authorize you to decrypt the model. This authorization is granted in the KMS key policy for the customer managed key. For more information, see AWS KMS key policy statement.
  - Provide the ARN of the customer managed key. You use this ARN when you create an IAM service role. This role authorizes Amazon Comprehend to use the KMS key to decrypt the model.

Required permissions

Before you can import a custom model, you or your administrator must authorize the required actions in AWS Identity and Access Management (IAM). As an Amazon Comprehend user, you must be authorized to import by an IAM policy statement. If encryption or decryption is required during the import, then Amazon Comprehend must be authorized to use the necessary AWS KMS keys.

IAM policy statement

Your user, group or role must have a policy attached that allows the `ImportModel` action, as shown in the following example.

Example IAM policy to import a custom model

```
{
  "Effect": "Allow",
  "Action": [
```

```
    "comprehend:ImportModel"
  ],
  "Resource": "arn:aws:comprehend:us-west-2:111122223333:document-classifier/foo/
version/*"
}
```

For information about creating an IAM policy, see Creating IAM policies in the *IAM User Guide*. For information about attaching an IAM policy, see Adding and removing IAM identity permissions in the *IAM User Guide*.

IAM service role for AWS KMS encryption

When you import a custom model, you must authorize Amazon Comprehend to use AWS KMS keys in either of the following cases:

- You are importing a custom model that is encrypted with a customer managed key in AWS KMS. In this case, Amazon Comprehend needs access to the KMS key so that it can decrypt the model during the import.
- You want to encrypt the new custom model that you create with the import, and you want to use a customer managed key. In this case, Amazon Comprehend needs access to your KMS key so that it can encrypt the new model.

To authorize Amazon Comprehend to use these AWS KMS keys, you create an *IAM service role*. This type of IAM role allows an AWS service to access resources in other services on your behalf. For more information about service roles, see Creating a role to delegate permissions to an AWS service in the *IAM User Guide*.

If you use the Amazon Comprehend console to import, you can have Amazon Comprehend create the service role for you. Otherwise, you must create a service role in IAM before you import.

The IAM service role must have a permissions policy and a trust policy, as shown by the following examples.

Example permissions policy

The following permissions policy allows the AWS KMS operations that Amazon Comprehend uses to encrypt and decrypt custom models. It grants access to two KMS keys:

- One KMS key is in the AWS account that contains the model to import. It was used to encrypt the model, and Amazon Comprehend uses it to decrypt the model during the import.

- The other KMS key is in the AWS account that imports the model. Amazon Comprehend uses this key to encrypt the new custom model that is created by the import.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
        "Effect": "Allow",
        "Action": [
            "kms:CreateGrant"
        ],
        "Resource": [
            "arn:aws:kms:us-west-2:111122223333:key/key-id",
            "arn:aws:kms:us-west-2:444455556666:key/key-id"
        ]
    },
    {
        "Effect": "Allow",
        "Action": [
            "kms:Decrypt",
            "kms:GenerateDatakey"
        ],
        "Resource": [
            "arn:aws:kms:us-west-2:111122223333:key/key-id",
            "arn:aws:kms:us-west-2:444455556666:key/key-id"
        ],
        "Condition": {
            "StringEquals": {
                "kms:ViaService": [
                    "s3.us-west-2.amazonaws.com"
                ]
            }
        }
    }
  ]
}
```

Example trust policy

The following trust policy allows Amazon Comprehend to assume the role and gain its permissions. It allows the `comprehend.amazonaws.com` service principal to perform the `sts:AssumeRole` operation. To help with confused deputy prevention, you restrict the scope of the permission by

using one or more global condition context keys. For `aws:SourceAccount`, specify the account Id of the user who is importing the model.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "comprehend.amazonaws.com"
      },
      "Action": "sts:AssumeRole",
      "Condition": {
        "StringEquals": {
          "aws:SourceAccount": "444455556666"
        }
      }
    }
  ]
}
```

Importing a custom model

You can import a custom model by using the AWS Management Console, AWS CLI, or Amazon Comprehend API.

AWS Management Console

You can use Amazon Comprehend in the AWS Management Console.

To import a custom model

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. In the navigation menu on the left, under **Customization**, choose the page for the type of model that you are importing:
   a. If you are importing a custom document classifier, choose **Custom classification**.
   b. If you are importing a custom entity recognizer, choose **Custom entity recognition**.
3. Choose **Import version**.
4. On the **Import model version** page, enter the following details:

- **Model version ARN** – The ARN of the model version to import.
- **Model name** – A custom name for the new model that is created by the import.
- **Version name** – A custom name for the new model version that is created by the import.

5. For **Model encryption**, choose the type of KMS key to use to encrypt the new custom model that you create with the import:

   - **Use AWS owned key** – Amazon Comprehend encrypts your model by using a key in AWS Key Management Service (AWS KMS) that is created, managed, and used on your behalf by AWS.
   - **Choose a different AWS KMS key (advanced)** – Amazon Comprehend encrypts your model by using a customer managed key that you manage in AWS KMS.

     If you choose this option, select a KMS key that's in your AWS account, or create a new one by choosing **Create an AWS KMS key**.

6. In the **Service access** section, grant Amazon Comprehend access to any AWS KMS keys that it needs to:

   - Decrypt the custom model that you import.
   - Encrypt that the new custom model that you create with the import.

   You grant access with an IAM service role that allows Amazon Comprehend to use the KMS keys.

   For **Service role**, do one of the following:

   - If you have an existing service role that you want to use, choose **Use an existing IAM role**. Then, select it under **Role name**.
   - If you want Amazon Comprehend to create a role for you, choose **Create an IAM role**.

7. If you chose to have Amazon Comprehend create the role for you, do the following:

   a. For **Role name**, enter a role name suffix that will help you recognize the role later.

   b. For **Source KMS key ARN**, enter the ARN of the KMS key that is used to encrypt the model that you're importing. Amazon Comprehend uses this key to decrypt the model during the import.

8. (Optional) In the **Tags** section, you can add tags to the new custom model that you create by importing. For more information about tagging custom models, see Tagging a new resource.

9. Choose **Confirm**.

AWS CLI

You can use Amazon Comprehend by running commands with the AWS CLI.

Example Import-model command

To import a custom model, use the `import-model` command:

```
$ aws comprehend import-model \
> --source-model arn:aws:comprehend:us-west-2:111122223333:document-classifier/foo/
version/bar \
> --model-name importedDocumentClassifier \
> --version-name versionOne \
> --data-access-role-arn arn:aws:iam::444455556666:role/comprehendAccessRole \
> --model-kms-key-id kms-key-id
```

This example uses the following parameters:

- `source-model` – The ARN of the custom model to import.
- `model-name` – A custom name for the new model that is created by the import.
- `version-name` – A custom name for the new model version that is created by the import.
- `data-access-role-arn` – The ARN of the IAM service role that allows Amazon Comprehend to use the necessary AWS KMS keys to encrypt or decrypt the custom model.
- `model-kms-key-id` – The ARN or ID of the KMS key that Amazon Comprehend uses to encrypt the custom model that you create with this import. This key must be in AWS KMS in your AWS account.

Amazon Comprehend API

To import a custom model by using the Amazon Comprehend API, use the ImportModel API action.

Flywheels

An Amazon Comprehend *flywheel* simplifies the process of improving a custom model over time. You can use a flywheel to orchestrate the tasks associated with training and evaluating new custom model versions. Flywheels support plain text custom models for custom classification and custom entity recognition.

Topics

Flywheel overview

A *flywheel* is an Amazon Comprehend resource that orchestrates the training and evaluation of new versions of a custom model. You can create a flywheel to use an existing trained model, or Amazon Comprehend can create and train a new model for the flywheel. Use flywheels with plain-text custom models for custom classification or custom entity recognition.

You can configure and manage flywheels using the Amazon Comprehend console or API. You can also configure flywheels using AWS CloudFormation.

When you create a flywheel, Amazon Comprehend creates a *data lake* in your account. The data lake stores and manages all the flywheel data, such as the training data and test data for all versions of the model.

You set the *active model version* to be the version of the flywheel model that you want to use for inference jobs or Amazon Comprehend endpoints. Initially, the flywheel contains one version of the model. Over time, as you train new model versions, you select the best-performing version to be

the active model version. When a user specifies the flywheel ARN to run an inference job, Amazon Comprehend runs the job using the flywheel's active model version.

Periodically, you obtain new labeled data (training data or test data) for the model. You make new data available to the flywheel by creating one or more *datasets*. A dataset contains input data for training or testing the custom model associated with a flywheel. Amazon Comprehend uploads the input data to the flywheel's data lake.

To incorporate the new datasets into your custom model, you create and run a flywheel *iteration*. A flywheel iteration is a workflow that uses the new datasets to evaluate the active model version and to train a new model version. Based on the metrics for the existing and new model versions, you can decide whether to promote the new model version to be the active version.

You can use the flywheel active model version to run custom analysis (real time or asynchronous jobs). To use the flywheel model for real-time analysis, you must create an endpoint for the flywheel.

There is no additional charge for using flywheels. However, when you run a flywheel iteration, you incur the standard charges for training a new model version and storing the model data. For detailed pricing information, see Amazon Comprehend Pricing.

Topics

Flywheel datasets

To add new labeled data to a flywheel, you create a dataset. You configure each dataset as training data or test data. You associate the dataset with a specific flywheel and custom model.

After you create a dataset, Amazon Comprehend uploads the data to the flywheel's data lake. For more information, see Flywheel data lakes.

Flywheel creation

When you create a flywheel, you can associate the flywheel with an existing trained model, or the flywheel can create a new model.

When you create a flywheel with an existing model, you specify the active model version. Amazon Comprehend copies the model's training data and test data into the flywheel's data lake. Make sure that the model training and test data exist in the same Amazon S3 location as when you created the model.

To create a flywheel for a new model, you provide a dataset for training data (and an optional dataset for test data) when you create the flywheel. When you run the flywheel to create the first flywheel iteration, the flywheel trains the new model.

When you train a custom model, you specify a list of custom labels (custom classification) or custom entities (custom entity recognition) for the model to recognize. Note the following important points about custom labels/entities:

- When you create a flywheel for a new model, the list of labels/entities that you provide during flywheel creation is the final list for the flywheel.
- When you create a flywheel from an existing model, the list of labels/entities associated with that model becomes the final list for the flywheel.
- If you associate a new dataset with the flywheel, and that dataset contains additional labels/entities, Amazon Comprehend ignores the new labels/entities.
- You can review a flywheel's label/entity list using the DescribeFlywheel API operation.

> **Note**
>
> For custom classification, Amazon Comprehend populates the label list after the flywheel status becomes ACTIVE. Wait until the flywheel is active before calling the DescribeFlywheel API operation.

Flywheel states

A flywheel transitions between the following states:

- CREATING - Amazon Comprehend is creating the flywheel resources. You can perform read operations on the flywheel, such as `DescribeFlywheel`.
- ACTIVE - The flywheel is active. You can determine if a flywheel iteration in progress and view the status of the iteration. You can perform read actions on the flywheel and actions such as `DeleteFlywheel` and `UpdateFlywheel`.

- UPDATING - Amazon Comprehend is updating the flywheel. You can perform read operations on the flywheel.
- DELETING - Amazon Comprehend is deleting the flywheel. You can perform read operations on the flywheel.
- FAILED - the flywheel create operation failed.

After Amazon Comprehend deletes a flywheel, you retain access to all the model data in the flywheel data lake. Amazon Comprehend deletes all the internal metadata required for managing the flywheel resources. Amazon Comprehend also deletes the datasets associated with this flywheel (the model data is saved in the data lake).

Flywheel iterations

When you obtain new training or test data for a flywheel model, you create one or more new datasets to upload the new data to the flywheel's data lake.

You then run the flywheel to create a new flywheel iteration. The flywheel iteration evaluates the current active model version using the new data and stores the results in the data lake. The flywheel also creates and trains a new model version.

If the new model exhibits better performance than the current active model version, you can promote the new model version to be the active model version. You can use the console or the UpdateFlywheel API operation to update the active model version.

Flywheel data lakes

When you create a flywheel, Amazon Comprehend creates a data lake in your account to contain all the flywheel data, such as the input and output data required for the model versions.

Amazon Comprehend creates the data lake in the Amazon S3 location that you specify when you create the flywheel. You can specify the location as an Amazon S3 bucket or as a new folder in an Amazon S3 bucket.

Data lake folder structure

When Amazon Comprehend creates the data lake, it sets up the following folder structure in the Amazon S3 location.

> **Warning**
>
> Amazon Comprehend manages the data lake folder organization and contents. Always use the Amazon Comprehend API operations to modify the data lake folders, or your flywheel may not operate correctly.

```
Document Pool
Annotations Pool
Staging
Model Datasets
  (data for each version of the model)
  VersionID-1
    Training
    Test
    ModelStats
  VersionID-2
    Training
    Test
    ModelStats
```

To view the training assessment of a model version, perform these steps:

1. Open the folder named **Model Datasets** at the root level of the data lake. This folder contains a subfolder for each version of the model.
2. Open the folder for the model version of interest.
3. Open the folder named **ModelStats** to view the statistics for the model.

Data lake management

Amazon Comprehend performs the following tasks to manage the data lake on your behalf:

- Defines the folder structure of the data lake and ingests datasets into the appropriate folders.
- Manages the input documents (such as text files and annotation files) required to train the model.
- Manages the training and evaluation output data associated with each version of the model.
- Manages encryption for files stored in the data lake.

Amazon Comprehend performs all the data creation and update operations for the data lake. You retain full access to the data in the data lake. For example:

- You have full access to the contents of the data lake.
- The data lake remains available after you delete the flywheel.
- You can configure access logs for the Amazon S3 bucket that contains the data lake.
- You can provide encryption keys for the data. You specify these when you create the flywheel.

We recommend the following best practices:

- Don't manually add your own folders or files into the data lake. Don't modify or delete any files in the data lake.
- Always use the Amazon Comprehend creation and update operations to add or modify data in the data lake. For example, use `CreateDataset` to provide training or test data and `StartFlywheelIteration` to generate evaluation data for model versions.
- The data lake structure may evolve over time. Don't create downstream scripts or programs that rely explicitly on the data lake structure.
- When you provide a data lake location for the flywheel, we recommend creating a common prefix for data related to all flywheels or using a different prefix for each flywheel. We don't recommend using the complete data lake path of one flywheel as the prefix for another flywheel.

IAM policies and permissions

You configure the following policies and permissions to use flywheels:

- the section called “Configure IAM user permissions” for users to access flywheel operations.
- (Optional) the section called “Configure permissions for AWS KMS keys” for the data lake.
- the section called “Create a data access role” that authorizes Amazon Comprehend to access the data lake.

Configure IAM user permissions

To use flywheel capabilities, add appropriate permissions policies to your AWS Identity and Access Management (IAM) identities (users, groups, and roles).

The following example shows permissions policy to create datasets, to create and manage flywheels, and to run the flywheel.

Example IAM policy to manage flywheels

```
{
  "Effect": "Allow",
  "Action": [
    "comprehend:CreateFlywheel",
    "comprehend:DeleteFlywheel",
    "comprehend:UpdateFlywheel",
    "comprehend:ListFlywheels",
    "comprehend:DescribeFlywheel",
    "comprehend:CreateDataset",
    "comprehend:DescribeDataset",
    "comprehend:ListDatasets",
    "comprehend:StartFlywheelIteration",
    "comprehend:DescribeFlywheelIteration",
    "comprehend:ListFlywheelIterationHistory"
  ],
  "Resource": "*"
}
```

For information about creating IAM policies for Amazon Comprehend, see How Amazon Comprehend works with IAM.

Configure permissions for AWS KMS keys

If you are using AWS KMS keys for your data in the data lake, set up the required permissions. For information, see Permissions required to use KMS encryption .

Create a data access role

You create a data access role in IAM for Amazon Comprehend to access flywheel data in the data lake. If you use the console to create a flywheel, the system can optionally create a new role for this purpose. For more information, see Role-based permissions required for asynchronous operations.

Configuring flywheels using the console

You can use the Amazon Comprehend console to create, update, and delete flywheels.

When you create a flywheel, Amazon Comprehend creates a data lake to hold all the data that the flywheel needs, such as the training data and test data for each version of the model.

When you delete a flywheel, Amazon Comprehend doesn't delete the data lake or the model associated with the flywheel.

Review the information in section Flywheel creation before you create a new flywheel.

Topics

Create a flywheel

When you create a flywheel, the required configuration fields depend on whether the flywheel is for an existing custom model or a new model.

To create a flywheel

1. Sign in to the AWS Management Console and open the Amazon Comprehend console.
2. From the left menu, choose **Flywheels**.
3. From the **Flywheels** table, choose **Create new flywheel**.
4. Under **Flywheel name**, enter a name for the flywheel.
5. (Optional) To create a flywheel for an existing model, configure the fields under **Active model version**.
    a. From the **Model** drop-down list, select a model
    b. From the **Version** drop-down list, select the model version.
6. (Optional) To create a new classifier model for the flywheel, under **Custom model type**, choose a **Custom classification** and configure the parameters in following steps.
    a. Under **Language**, select the language for the model.
    b. Under **Classifier mode**, choose single-label mode or multi-label mode.
    c. Under **Custom labels**, enter one or more custom labels to use for training the model. Each label must match one of the classes in your input training data.

7. (Optional) To create a new entity recognition model for the flywheel, under **Custom model type**, choose a **Custom entity recognition** and configure the parameters in following steps.

   a. Under **Language**, select the language for the model.

   b. Under **Custom entity type**, enter up to 25 custom entities to use for training the model. Each label must match one of the entity types in your input training data.

      To create more than one label, perform the following steps multiple times.

      i. Enter a custom label. The label must be all uppercase. Use an underscore as a separator between words in the label.

      ii. Choose **Add type**.

      To remove one of the labels that you've added, choose **X** to the right of the label name.

8. Configure your choices for volume encryption, model encryption, and data lake encryption. For each of these, choose whether to use an AWS owned KMS key or a key that you have permission to use.

   - If you are using an AWS owned KMS key, there are no additional parameters.
   - If you are using another existing key, for **KMS key ARN** enter the ARN for the key ID.
   - If you want to create a new key, choose **Create an AWS KMS key**.

   For more information on creating and using KMS keys and the associated encryption, see AWS Key Management Service.

   a. Configure the **Volume encryption** key. Amazon Comprehend uses this key to encrypt the data in the storage volume while your job is being processed. choose whether to use an AWS owned KMS key or a key that you have permission to use.

   b. Configure the **Model encryption** key. Amazon Comprehend uses this key to encrypt the model data for this model version.

9. Configure the **Data lake location**. For more information, see Data lake management.

10. (Optional) Configure **Data lake encryption** key. Amazon Comprehend uses this key to encrypt all files in the data lake.

11. (Optional) Configure **VPC settings**. Enter the VPC ID under **VPC** or choose the ID from the drop-down list.

1. Choose the subnet under **Subnets(s)**. After you select the first subnet, you can choose additional ones.
2. Under **Security Group(s)**, choose the security group to use if you specified one. After you select the first security group, you can choose additional ones.

12. Configure the **Service access** permissions.
    1. If you select **Use an existing IAM role**, select the role name in the drop-down list.
    2. If you select **Create an IAM role**, Amazon Comprehend creates a new role. The console displays the permissions that Amazon Comprehend configures for the role. Under **Role name**, enter a descriptive name for the role.
13. (Optional) Configure **Tags** settings. To add a tag, enter a key-value pair under **Tags**. Choose **Add tag**. To remove this pair before creating the flywheel, choose **Remove tag**. For more information, see Tagging your resources.
14. Choose **Create**.

Update a flywheel

You can configure the flywheel name, data lake location, model type, and model configuration only when you create the flywheel.

When you update a flywheel, you can specify a different model if the model type and configuration options are the same as the current model. You can configure a new active model version. You can also update encryption details, service access permissions, and VPC settings.

To update a flywheel

1. Sign in to the AWS Management Console and open the Amazon Comprehend console.
2. From the left menu, choose **Flywheels**.
3. From the **Flywheels** table, choose the flywheel to update.
4. Under **Active model version**, choose a model from the **Model** drop-down list and choose a model version.

   The form populates the model type and model configuration.
5. (Optional) Configure **Volume encryption** and **Model encryption** settings.
6. (Optional) Configure **Data lake encryption** settings.

7. Configure the **Service access** permissions.
8. (Optional) Configure **VPC settings**.
9. (Optional) Configure **Tags** settings.
10. Choose **Save**.

Delete a flywheel

To delete a flywheel

1. Sign in to the AWS Management Console and open the Amazon Comprehend console.
2. From the left menu, choose **Flywheels**.
3. From the **Flywheels** table, choose the flywheel to delete.
4. Choose **Delete**.

Configuring flywheels using the API

You can use the Amazon Comprehend API to create, update, and delete flywheels.

When you create a flywheel, Amazon Comprehend creates a data lake to hold all the data that the flywheel needs, such as the training data and test data for each version of the model.

When you delete a flywheel, Amazon Comprehend doesn't delete the data lake or the model associated with the flywheel.

The flywheel delete operation fails if the flywheel is running an iteration or creating a dataset.

Review the information in section Flywheel creation before you create a new flywheel.

Create a flywheel for an existing model

Use the CreateFlywheel operation to create a flywheel for an existing model.

Example

```
aws comprehend create-flywheel  \
    --flywheel-name "myFlywheel2"  \
```

```
    --active-model-arn  "modelArn"  \
    --data-access-role-arn   arn:aws::iam::111122223333:role/testFlywheelDataAccess \
    --data-lake-s3-uri": "https://s3-bucket-endpoint"   \
```

If the operation is successful, the response includes the flywheel ARN.

```
{
  "FlywheelArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/name",
  "ActiveModelArn": "modelArn"
}
```

Create a flywheel for a new model

Use the CreateFlywheel operation to create a flywheel for a new custom classification model.

Example

```
aws comprehend create-flywheel \
    --flywheel-name "myFlywheel2" \
    --data-access-role-arn  arn:aws::iam::111122223333:role/testFlywheelDataAccess \
    --model-type "DOCUMENT_CLASSIFIER" \
    --data-lake-s3-uri  "s3Uri"  \
    --task-config  file://taskConfig.json
```

The taskConfig.json file contains the following content.

```
{
    "LanguageCode": "en",
    "DocumentClassificationConfig": {
        "Mode": "MULTI_LABEL",
        "Labels": ["optimism", "anger"]
    }
}
```

The API response body includes the following content.

```
{
  "FlywheelArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/name",
  "ActiveModelArn": "modelArn"
}
```

Describe a flywheel

Use the Amazon Comprehend DescribeFlywheel operation to retrieve configured information about a flywheel.

```
aws comprehend describe-flywheel \
    --flywheel-arn  "flywheelArn"
```

The API response body includes the following content.

```
{
  "FlywheelProperties": {
      "FlywheelArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/
myTestFlywheel",
      "DataAccessRoleArn": "arn:aws::iam::111122223333:role/Admin",
      "TaskConfig": {
          "LanguageCode": "en",
          "DocumentClassificationConfig": {
              "Mode": "MULTI_LABEL"
          }
      },
      "DataLakeS3Uri": "s3://my-test-datalake/flywheelbasictest/myTestFlywheel/
schemaVersion=1/20220801T014326Z",
      "Status": "ACTIVE",
      "ModelType":  "DOCUMENT_CLASSIFIER",
      "CreationTime": 1659318206.102,
      "LastModifiedTime": 1659318249.05
  }
}
```

Update a flywheel

Use the UpdateFlywheel operation to update the modifiable configuration values of the flywheel.

Some configuration fields are JSON structures with subfields. To update one or more subfields, provide values for all the subfields (Amazon Comprehend sets the value to null for any subfield missing in the request).

If you omit a top-level parameter in the `UpdateFlywheel` request, Amazon Comprehend doesn't change the values of the parameter or any of its subfields in the flywheel.

To add or remove tags on the flywheel, use the TagResource and UntagResource operations.

You can promote a model version by setting the `ActiveModelArn` parameter, as shown in the following example.

```
aws comprehend update-flywheel \
    --region aws-region \
    --flywheel-arn  "flywheelArn" \
    --active-model-arn  "modelArn" \
```

The API response body includes the following content.

```
{
  "FlywheelArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/name",
  "ActiveModelArn": "modelArn"
}
```

Delete a flywheel

Use the Amazon Comprehend DeleteFlywheel operation to delete flywheels.

```
aws comprehend delete-flywheel \
    --flywheel-arn  "flywheelArn"
```

A successful API response contains an empty response message body

List the flywheels

Use the Amazon Comprehend ListFlywheels operation to retrieve a list of flywheels in the current region.

```
aws comprehend list-flywheel \
    --region aws-region \
    --endpoint-url  "uri"
```

The API response body includes the following content.

```
{
    "FlywheelSummaryList": [
        {
            "FlywheelArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/
myTestFlywheel",
```

```
            "DataLakeS3Uri": "s3://my-test-datalake/flywheelbasictest/myTestFlywheel/
schemaVersion=1/20220801T014326Z",
            "Status": "ACTIVE",
            ""ModelType":  "DOCUMENT_CLASSIFIER",
            "CreationTime": 1659318206.102,
            "LastModifiedTime": 1659318249.05
        }
    ]
}
```

Configuring datasets

To add labeled training or test data to a flywheel, use the Amazon Comprehend console or API to create a dataset.

You configure each dataset as training data or test data. You associate the dataset with a specific flywheel and custom model. When you create a dataset, Amazon Comprehend uploads the data to the flywheel's data lake. For details about file formats for the training data, see Preparing classifier training data or Preparing entity recognizer training data.

When you delete the flywheel, Amazon Comprehend deletes the datasets. The uploaded data remains available in the data lake.

Creating a dataset (console)

Create a dataset

1. Sign in to the AWS Management Console and open the Amazon Comprehend console.
2. From the left menu, choose **Flywheels** and choose the flywheel where you want to add the data.
3. Choose the **Datasets** tab.
4. In the **Training datasets** or **Test datasets** table, choose **Create dataset**.
5. Under **Dataset details**, enter a name for the dataset and an optional description.
6. Under **Data specifications**, choose the **Data format** and the **Dataset type** configuration fields.
7. (Optional) Under **Input format**, choose the format of the input documents.
8. Under **Annotation location on S3**, enter the Amazon S3 location of the annotations file.
9. Under **Training data location on S3**, enter the Amazon S3 location of the document files.

10. Choose **Create**.

Creating a dataset (API)

You can use the CreateDataset operation to create a dataset.

Example

```
aws comprehend create-dataset \
    --flywheel-arn "myFlywheel2" \
    --dataset-name "my-training-dataset"
    --dataset-type "TRAIN"
    --description "my training dataset"
    --cli-input-json file://inputConfig.json
}
```

The `inputConfig.json` file contains the following content.

```
{
    "DataFormat": "COMPREHEND_CSV",
    "DocumentClassifierInputDataConfig": {
        "S3Uri": "s3://my-comprehend-datasets/multilabel_train.csv"
    }
}
```

To add or remove tags on the dataset, use the TagResource and UntagResource operations.

Describe a dataset

Use the Amazon Comprehend DescribeDataset operation to retrieve configured information about a flywheel.

```
aws comprehend describe-dataset \
    --dataset-arn  "datasetARN"
```

The response contains the following content.

```
{
   "DatasetProperties": {
      "DatasetArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/
myTestFlywheel/dataset/train-dataset",
```

```
      "DatasetName": "train-dataset",
      "DatasetType": "TRAIN",
      "DatasetS3Uri": "s3://my-test-datalake/flywheelbasictest/myTestFlywheel/
schemaVersion=1/20220801T014326Z/datasets/train-dataset/20220801T194844Z",
      "Description": "Good Dataset",
      "Status": "COMPLETED",
      "NumberOfDocuments": 90,
      "CreationTime": 1659383324.297
  }
}
```

Flywheel iterations

Use flywheel iterations to help you create and manage new model versions.

Topics

Iteration workflow

A flywheel starts out with a trained model version or uses an initial dataset to train a model version.

Over time, as you obtain new labeled data, you train new model versions to improve the performance of your flywheel model. When you run the flywheel, it creates a new iteration that trains and evaluates a new model version. You can promote the new model version if its performance is superior to the existing active model version.

The flywheel iteration workflow includes the following steps:

1. You create datasets for the new labeled data.
2. You run the flywheel to create a new iteration. The iteration follows these steps to train and evaluate a new model version:
   a. Evaluates the active model version using the new data.
   b. Trains a new model version using the new data.
   c. Stores the evaluation and training results in the data lake.

d. Returns the F1 scores for both models.

3. After the iteration completes, you can compare the F1 scores for the existing active model and the new model.
4. If the new model version has superior performance, you promote it to be the active model version. You can use the console or the API to promote the new model version.

Managing iterations (console)

You can use the console to start a new iteration and query the status of an in-progress iteration. You can also view the results of completed iterations.

Start a flywheel iteration (console)

Before you can start a new iteration, create one or more new training or test datasets. See Configuring datasets

Start a flywheel iteration (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console.
2. From the left menu, choose **Flywheels**.
3. From the **Flywheels** table, choose a flywheel.
4. Choose **Run flywheel**.

Analyze iteration results (Console)

After it runs the flywheel iteration, the console displays the results in the **Flywheels iterations** table.

Promote new model version (Console)

From the model details page in the console, you can promote a new model version to be the active model version.

Promote a flywheel model version to active model version (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console.
2. From the left menu, choose **Flywheels**.

3. From the **Flywheels** table, choose a flywheel.
4. From the **Flywheel details page** table, choose the version to promote from the **Flywheels iterations** table.
5. Choose **Make active model**.

Managing iterations (API)

You can use the Amazon Comprehend API to start a new iteration and query the status of an in-progress iteration. You can also view the results of completed iterations.

Start flywheel iteration (API)

Use the Amazon Comprehend StartFlywheelIteration operation to start a flywheel iteration.

```
aws comprehend start-flywheel-iteration \
    --flywheel-arn  "flywheelArn"
```

The response contains the following content.

```
{
  "FlywheelIterationArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/name"
}
```

Promote new model version (API)

Use the UpdateFlywheel operation to promote a model version to be the active model version.

Send the `UpdateFlywheel` request with the `ActiveModelArn` parameter set to the ARN of the new active model version.

```
aws comprehend update-flywheel \
    --active-model-arn  "modelArn" \
```

The response contains the following content.

```
{
  "FlywheelArn": "arn:aws::comprehend:aws-region:111122223333:flywheel/name",
  "ActiveModelArn": "modelArn"
```

```
}
```

Describe flywheel iteration results (API)

The Amazon Comprehend DescribeFlywheelIteration operation returns information about an iteration after it runs to completion.

```
aws comprehend describe-flywheel-iteration \
 --flywheel-arn "flywheelArn" \
 --flywheel-iteration-id  "flywheelIterationId" \
 --region aws-region
```

The response contains the followng content.

```
{
    "FlywheelIterationProperties": {
        "FlywheelArn": "flywheelArn",
        "FlywheelIterationId": "iterationId",
        "CreationTime": <createdAt>,
        "EndTime": <endedAt>,
        "Status": <status>,
        "Message": <message>,
        "EvaluatedModelArn": "modelArn",
        "EvaluatedModelMetrics": {
            "AverageF1Score": <value>,
            "AveragePrecision": <value>,
            "AverageRecall": <value>,
            "AverageAccuracy": <value>
        },
        "TrainedModelArn": "modelArn",
        "TrainedModelMetrics": {
            "AverageF1Score": <value>,
            "AveragePrecision": <value>,
            "AverageRecall": <value>,
            "AverageAccuracy": <value>
        }
    }
}
```

Get iteration history (API)

Use the ListFlywheelIterationHistory operation to get information about iteration history.

```
aws comprehend list-flywheel-iteration-history \
 --flywheel-arn "flywheelArn"
```

The response contains the followng content.

```
{
    "FlywheelIterationPropertiesList": [
        {
            "FlywheelArn": "<flywheelArn>",
            "FlywheelIterationId": "20220907T214613Z",
            "CreationTime": 1662587173.224,
            "EndTime": 1662592043.02,
            "Status": "<status>",
            "Message": "<message>",
            "EvaluatedModelArn": "modelArn",
            "EvaluatedModelMetrics": {
                "AverageF1Score": 0.8333333333333333,
                "AveragePrecision": 0.75,
                "AverageRecall": 0.9375,
                "AverageAccuracy": 0.8125
            },
            "TrainedModelArn": "modelArn",
            "TrainedModelMetrics": {
                "AverageF1Score": 0.865497076023392,
                "AveragePrecision": 0.7636363636363637,
                "AverageRecall": 1.0,
                "AverageAccuracy": 0.84375
            }
        }
    ]
}
```

Using flywheels for analysis

You can use the flywheel's active model version to run analysis for custom classification or entity recognition. The active model version is configurable. You can use the console or the UpdateFlywheel API operation to set a new version of the model to be the active model version.

To use the flywheel, specify the flywheel ARN instead of a custom model ARN when you configure the analysis task. Amazon Comprehend runs the analysis using the flywheel's active model version.

Real-time analysis

You use an endpoint to run real-time analysis. When you create or update an endpoint, you can configure it with the flywheel ARN instead of a model ARN. When you run the real-time analysis, select the endpoint associated with the flywheel. Amazon Comprehend runs the analysis using the active model version of the flywheel.

When you use UpdateFlywheel to set a new active model version for the flywheel, the endpoint updates automatically to start using the new active model version. If you don't want the endpoint to update automatically, configure the endpoint (using UpdateEndpoint) to use the model version ARN directly. The endpoint continues to use this model version if the flywheel active model version changes.

For custom classification, use the ClassifyDocument API operation. For custom entity recognition, use the DetectEntities API request. Provide the endpoint of the flywheel in the **EndpointArn** parameter.

You can also use the console to run real-time analysis for custom classification or custom entity recognition.

Asynchronous jobs

For custom classification, use the StartDocumentClassificationJob API request to start an aysnchronous job. Provide the **FlywheelArn** parameter instead of the **DocumentClassifierArn**.

For custom entity recognition, use the StartEntitiesDetectionJob API request. Provide the **FlywheelArn** parameter instead of the **EntityRecognizerArn**.

You can use the console to run asynchronous analysis jobs for custom classification or custom entity recognition. When you create the job, enter the flywheel ARN in the **Recognizer model** or **Classifier model** field.

Managing Amazon Comprehend endpoints

In Amazon Comprehend, endpoints make your custom models available for real-time classification or entity detection. After you create an endpoint, you can make changes to it as your business needs evolve. For example, you can monitor your endpoint utilization and apply auto scaling to automatically set endpoint provisioning to fit your capacity needs. You can manage all your endpoints from a single view, and when you no longer need an endpoint you can delete it to save costs.

Before you can manage an endpoint, you must create one. For more information, see the following procedures:

- Creating an endpoint for custom classification
- Creating an endpoint for custom entity detection

Topics

Overview of Amazon Comprehend endpoints

The endpoints page from Amazon Comprehend console provides you a global view of your endpoints. From the endpoints overview page, you can view all of your endpoints in one place to understand your endpoint usage versus your actual resource usage. On the top right of the endpoints page you can specify what endpoints you want to view— all of them, custom classifier endpoints, or your custom entity endpoints.

You can create, update, monitor, and delete endpoints from this page. From the endpoints overview section, you can view a list of your endpoints, what custom models the endpoints are

hosting, their creation time, the provisioned throughput, and the status of the endpoint. When you select a specific endpoint from the endpoint overview table, the endpoint details are displayed.

Also, if you are a AWS Business Support or an AWS Enterprise Support customer, you have access to Trusted Advisor checks specific to your endpoints. To learn more, see Using Trusted Advisor with Amazon Comprehend. For a complete list of checks and descriptions, see the Trusted Advisor Best Practices.

For more information on managing your endpoints, see the following topics.

- Using Amazon Comprehend endpoints
- Monitoring Amazon Comprehend endpoints
- Updating Amazon Comprehend endpoints
- Using Trusted Advisor with Amazon Comprehend
- Deleting Amazon Comprehend endpoints

> **Important**
>
> The cost for real-time custom classification is based on both the throughput you set and the length of time the endpoint is active. If you are no longer using the endpoint, or are not using it for an extended period, you should set up an auto scaling policy to reduce your costs. Or, if you are no longer using an endpoint you can delete the endpoint to avoid incurring additional cost. For more information, see Auto scaling with endpoints.

Using Amazon Comprehend endpoints

You create an endpoint to run real-time analysis using a custom model. An endpoint includes managed resources that makes your custom model available for real-time inference.

Amazon Comprehend assigns throughput to an endpoint using *Inference units* (IU). An IU represents data throughput of 100 characters per second. You can provision the endpoint with up to 10 inference units. You can scale the endpoint throughput either up or down by updating the endpoint.

If your input documents include semi-structured documents or image files, the throughput of 100 characters per second is for the characters extracted from the input file. The number of IUs that you provision for an endpoint depends on character density of the input documents.

The ClassifyDocument and DetectEntities API responses include the character count for each page of input. You can use this information to estimate the number of inference units to provision to achieve the desired throughput.

After you have completed your real-time analysis, delete the endpoint because the charge for it continues as long as it's active. You can create another endpoint when you are ready to run further real-time analysis.

For more information on endpoint cost, see Amazon Comprehend Pricing.

After you create an endpoint, you can monitor it with Amazon CloudWatch, update it to change its inference units, or delete it when no longer needed. For more information, see Monitoring Amazon Comprehend endpoints.

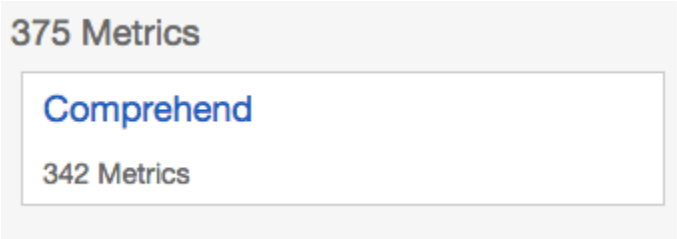
Monitoring Amazon Comprehend endpoints

You can adjust the throughput of your endpoint by increasing or decreasing the number of inference units (IUs). For more information on updating your endpoint, see the section called "Updating endpoints".

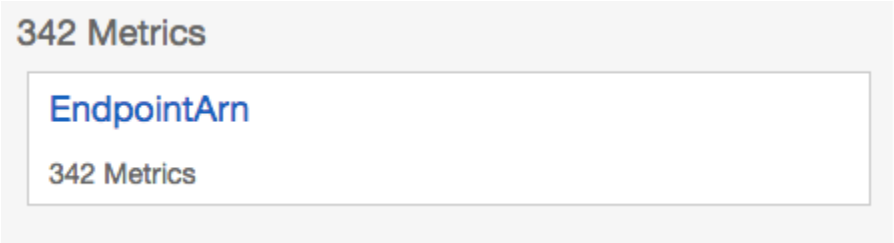
You can determine how to best adjust your endpoint's throughput by monitoring its usage with the Amazon CloudWatch console.

Monitor your endpoint usage with CloudWatch

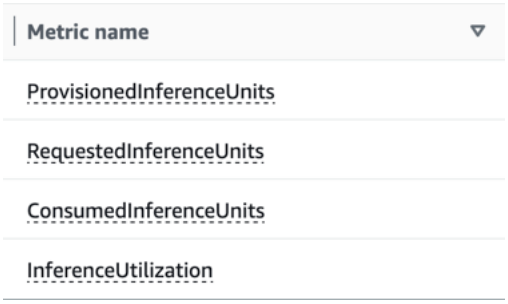
1. Sign in to the AWS Management Console and open the CloudWatch console.
2. On the left, choose **Metrics** and select **All metrics**.
3. Under **All metrics**, choose **Comprehend**.



4. The CloudWatch console displays the dimensions for the **Comprehend** metrics. Choose the **EndpointArn** dimension.



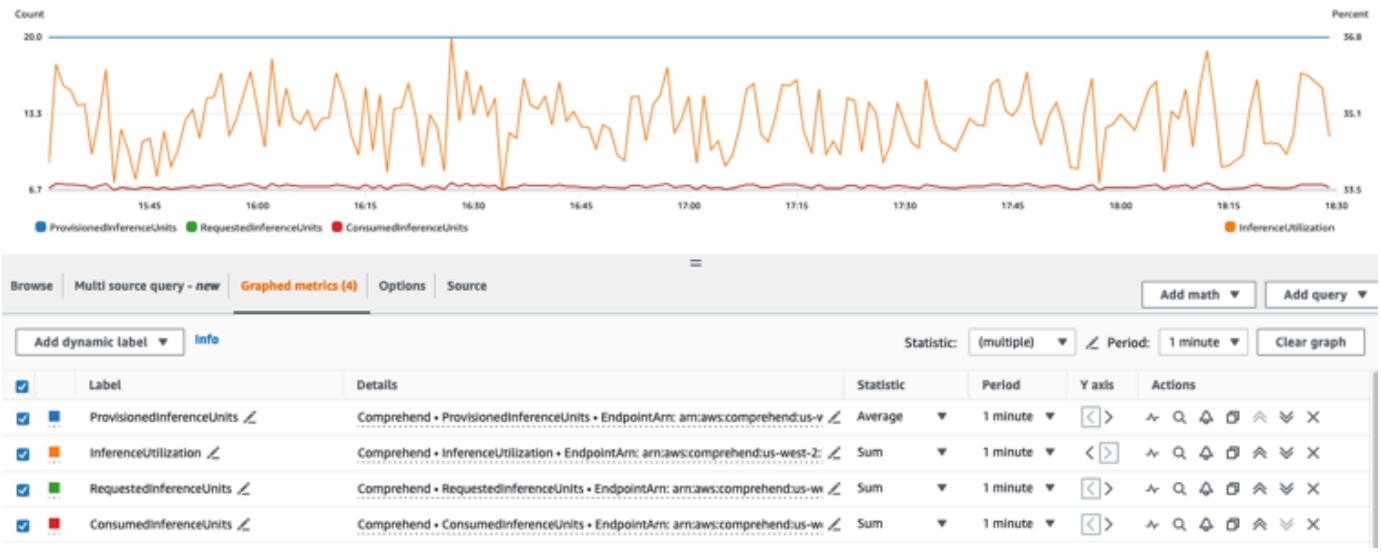
The console displays **ProvisionedInferenceUnits**, **RequestedInferenceUnits**, **ConsumedInferenceUnits**, and **InferenceUtilization** for each of your endpoints.



Select the four metrics and navigate to the **Graphed metrics** tab.

5. Set the Statistic columns for **RequestedInferenceUnits** and **ConsumedInferenceUnits** to **Sum**.
6. Set the Statistic column for **InferenceUtilization** to **Sum**.
7. Set the Statistic column for **ProvisionedInferenceUnits** to **Average**.
8. Change the Period column for all metrics to **1 Minute**.
9. Select **InferenceUtilization** and select the arrow to move it to a separate **Y Axis**.

Your graph is ready for analysis.



Based on the CloudWatch metrics, you can also set up auto scaling to automatically adjust the throughput of your endpoint. For more information about using auto scaling with your endpoints, see Auto scaling with endpoints.

- **ProvisionedInferenceUnits** - This metric represents the number of average provisioned IUs at the time the request was made.
- **RequestedInferenceUnits** - This is based on the usage of each request submitted to the service that was sent to be processed. This can be helpful to compare the request sent to be processed to what was actually processed without getting throttling (ConsumedInferenceUnits). The value for this metric is calculated by taking the number of characters sent to be processed and dividing it by the number of characters that can be processed in a minute for 1 IU.
- **ConsumedInferenceUnits** - This is based on the usage of each request submitted to the service that was successfully processed (not throttled). This can be helpful when you compare what you're consuming against your provisioned IUs. The value for this metric is calculated by taking the number of characters processed and dividing it by the number of characters that can be processed in a minute for 1 IU.
- **InferenceUtilization** - This is emitted per request. This value is calculated by taking the consumed IUs defined in **ConsumedInferenceUnits** and dividing it by **ProvisionedInferenceUnits** and converting to a percentage out of 100.

> **Note**
>
> All of the metrics are emitted only for successful requests. The metric won't appear if it's from a request that is throttled or fails with an internal server error or a customer error.

Updating Amazon Comprehend endpoints

Frequently, the level of throughput you need changes after creating an endpoint, or your first estimation of your needs changes. When this happens, it may be necessary to update your endpoint to adjust the throughput up or down. Throughput is governed by the number of inference units with which you've provisioned your endpoint. Each inference unit represents a throughput of 100 characters per second for up to 2 documents per second. You might also want to update the version of the model associated with the endpoint. When you edit an endpoint, you can choose a different version of the model for the endpoint.

It can also be helpful to add tags to your endpoint to help keep them organized. This can also be done while updating your endpoint. For more information on endpoints, see Tagging your resources

To update an endpoint (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Endpoints**.
3. From the **Classifiers** list, choose the name of the custom model from which you want to update the endpoint and follow the link. The model details page displays.
4. From the model details page, select the version details. The endpoints list displays.
5. Select the endpoint checkbox for your endpoint. At the top right of the endpoints table, select the **Actions** icon.
6. Choose **Edit**. You can update provisioned IUs and edit tags.
7. Save your changes.
8. To edit the number of inference units with which the endpoint is provisioned, choose **Edit**.
9. Enter the updated number of inference units to assign to the endpoint. Each unit represents a throughput of 100 characters per second. You can assign up to a maximum of 10 inference units per endpoint.

> **Note**
>
> The cost of using an endpoint is based on the amount of time operating and the throughput (based on the number of inference units. Increasing the number of inference units will thus increase the cost of operation. For more information, see Amazon Comprehend pricing.

10. Choose **Edit endpoint**. The endpoint details page is displayed.
11. Confirm that the endpoint is updating by choosing the model name from the breadcrumbs at the top of the page. On the custom model details page, navigate to the **Endpoints** list and verify that it shows **Updating** next to the endpoint. When the update is complete, it will show **Ready**.

    The following example demonstrates using the *UpdateEndpoint* operation with the AWS CLI.

    The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

    ```
    aws comprehend update-endpoint \
        --desired-inference-units updated number of inference units \
        --desired-model-arn arn:aws:comprehend:region:account-id:model type/model name
     \
        --desired-data-access-role-arn arn:aws:iam:account id:role/role name
        --endpoint-arn arn:aws:comprehend:region:account id:endpoint/endpoint name
    ```

    If the action is successful, Amazon Comprehend responds with an HTTP 200 response with an empty HTTP body.
12. To edit the custom model attached to your endpoint, from the custom model details page, navigate to the **Endpoints** list.
13. Select the endpoint you want to change and select **Edit**.
14. From the endpoint settings page, under **Select classifier model** or **Select recognizer model** depending on your endpoint, you can search for a model in the dropdown. Select the model you want.
15. Under **Select version** you can search for the model version you want. Select the version.
16. Select **Edit endpoint** to save.

Using Trusted Advisor with Amazon Comprehend

AWS Trusted Advisor is an online tool that provides recommendations to help you provision your resources following AWS best practices.

If you have a Basic or Developer Support plan, you can use the Trusted Advisor console to access all checks in the Service Limits category and six checks in the Security category. If you have a Business or Enterprise Support Plan, you can use the Trusted Advisor console and the AWS Support API to access all of the Trusted Advisor checks.

Amazon Comprehend supports the following Trusted Advisor checks to help customers optimize the cost and the security of their Amazon Comprehend endpoints by providing actionable recommendations.

Amazon Comprehend underutilized endpoints

The **Amazon Comprehend underutilized endpoints** check evaluates the throughput configuration of your endpoints. This check alerts you when endpoints are not actively used for real-time inference requests. An endpoint that isn't used for more than 15 days is considered underutilized. All endpoints accrue charges based on both the throughput set and the length of time that the endpoint is active. For the endpoint not used in last 15 days, we recommend that you define a scaling policy for the resource using Application Autoscaling. For an endpoint that hasn't been used in the last 30 days and does have an auto scaling policy defined we recommend that you use asynchronous inference or delete it. These check results are automatically refreshed once every day and can be viewed under the **CostOptimization** category on the Trusted Advisor console.

To view the utilization status of all your endpoints and the corresponding recommendations

1. Sign in to the AWS Management Console and open the Trusted Advisor console.
2. In the navigation pane, choose the **CostOptimization** check category.
3. On the category page, you can view the summary for each check category:

   - **Action recommended (red)** – Trusted Advisor recommends an action for the check.
   - **Investigation recommended (yellow)** – Trusted Advisor detects a possible issue for the check.
   - **No problems detected (green)** – Trusted Advisor doesn't detect an issue for the check.
   - **Excluded items (gray)**– The number of checks that have excluded items, such as resources that you want a check to ignore.

4. Choose **Amazon Comprehend Underutilized Endpoints check** to view the check description and the following details:

   - **Alert Criteria** – Describes the threshold when a check will change status.
   - **Recommended Action** – Describes the recommended actions for this check.
   - **Resource Table:** A table that lists your endpoint details and the status for each based on your recommendations.

5. In the Resource table, if an endpoint is flagged with a **Investigation Recommended** because of a **Not used in last 30 days** warning, you can navigate to the Endpoint Details page on the Amazon Comprehend console.

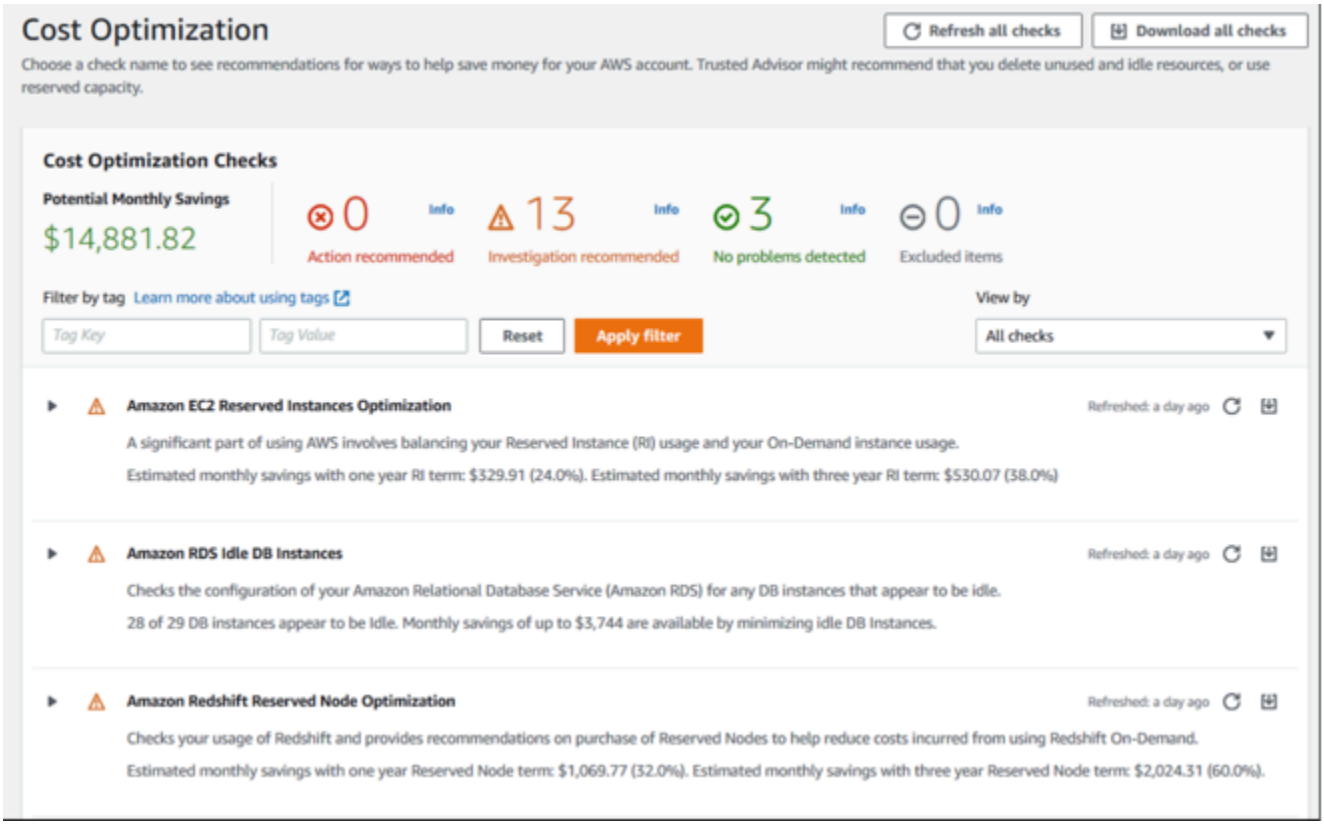
   - If you do not want to use this endpoint anymore, choose **Delete**.
   - Choose **Delete** again to confirm the deletion. The custom model details page is displayed. Confirm that the endpoint you deleted shows **deleting** next to it. When it has deleted, the endpoint is removed from the **Endpoints** list.

6. In the Resource table on the Trusted Advisor console, if an endpoint is flagged with an **Investigation Recommended** status because it hasn't been used in the last 15 days, and if it has AutoScaling disabled, you can navigate to the Endpoint Details page on the Amazon Comprehend console to adjust the endpoint.

   - If you want to reduce the throughput configured for this endpoint, click **Edit**. Enter the updated number of inference units to assign to the endpoint, then select the checkbox to acknowledge and then choose **Edit Endpoint.** When the update is complete, the status will show as **Ready**.
   - If you want to automatically set endpoint provisioning on your endpoint instead of manually adjusting the throughput configuration, we recommend you use **Application Autoscaling**.

7. In the Resource table on the Trusted Advisor console, if an endpoint is flagged with the **No problems detected** status because of the **Used Actively** reason, then it implies the endpoint is being utilized actively for running real-time inference requests and no actions are recommended.

Here's an example which shows the CostOptimization category view on the Trusted Advisor console:



Amazon Comprehend endpoint access risk

The **Amazon Comprehend endpoint access risk** check evaluates the AWS Key Management Service (AWS KMS) key permissions for an endpoint where the underlying model was encrypted using customer managed keys. If the customer managed key is disabled or the key policy was changed to alter the allowed permissions for Amazon Comprehend, the endpoint availability might be affected. If the key has been disabled, we recommend that you enable it. If the key policy has been altered and you wish to continue using this endpoint, we recommend that you update the key policy. The check results are automatically refreshed multiple times during the day. This check can be viewed under the **Fault Tolerance** category of the Trusted Advisor console.

To view the AWS KMS key status of your Amazon Comprehend endpoints

1. Sign in to the AWS Management Console and open the Trusted Advisor console.
2. In the navigation pane, choose the **FaultTolerance** check category.
3. On the category page, you can view the summary for each check category:

   - **Action recommended (red)** – Trusted Advisor recommends an action for the check.

- **Investigation recommended (yellow)**– Trusted Advisor detects a possible issue for the check.
- **No problems detected (green)** – Trusted Advisor doesn't detect an issue for the check.
- **Excluded items (gray)** – The number of checks that have excluded items, such as resources that you want a check to ignore.

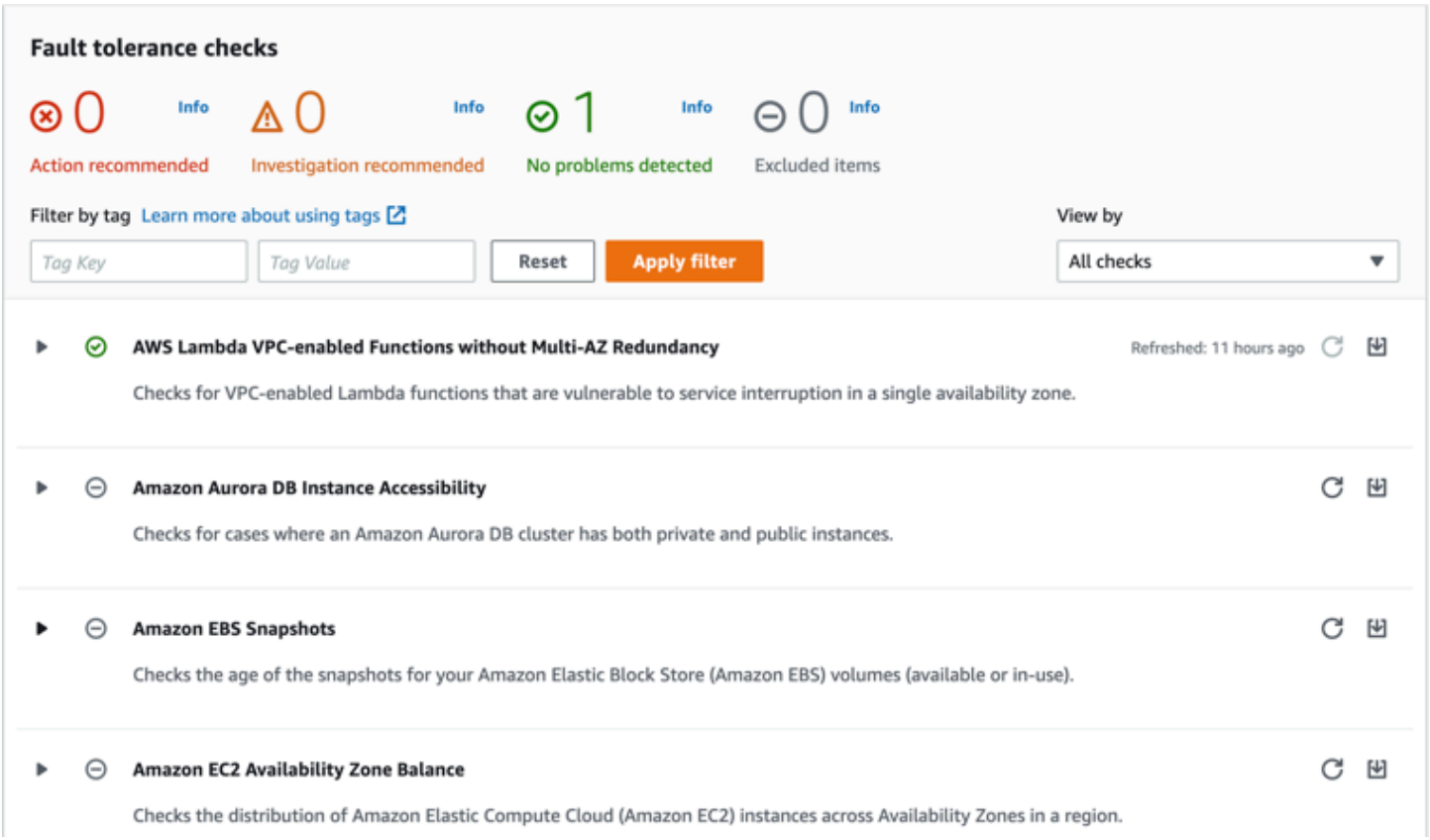
4. Choose Amazon Comprehend Endpoint Access Risk Check and you can view the check description and the following details:

   - **Alert Criteria**– Describes the threshold when a check will change status.
   - **Recommended Action** – Describes the recommended actions for this check.
   - **Resource Table:** A table that lists your KMS encrypted endpoint details and the status for each one based on if there are recommended actions.

5. In the Resource table, if an endpoint is flagged with an **Action Recommended** status, select the link in the KMS KeyId column and you will be redirected to the corresponding AWS KMS key page.

   - To enable a disabled AWS KMS key, choose **Key Actions**, and select **Enable**.
   - If the Key Status is listed as **Enabled**, update the key policy by choosing **Switch to policy view** in the Key Policy section. Edit the key policy document to provide the necessary permissions to Amazon Comprehend and then choose **Save changes**.

Here's an example of the FaultTolerance category view on the Trusted Advisor console:



These checks and their results can also be viewed by referring the Trusted Advisor section of the AWS Support API.

To learn more about setting up alarms using CloudWatch, see: Creating Trusted Advisor alarms using CloudWatch. For a full set of Trusted Advisor Best Practice Checks, see: AWS Trusted Advisor best practice checklist.

Deleting Amazon Comprehend endpoints

Once you no longer need your endpoint, you should delete it so that you stop incurring costs from it. You can easily create another endpoint whenever you need it from the **Endpoints** section.

To delete an endpoint (console)

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. From the left menu, choose **Endpoints**.
3. From the **Endpoints table** locate the endpoint you want to delete. You can search or filter all of the endpoints to find the one you need.

4. Select the endpoint checkbox for the endpoint you want to delete. At the top right of the endpoints table, select the **Actions** icon.
5. Choose **Delete**.
6. Choose **Delete** again to confirm the deletion. The endpoints page is displayed. Confirm that the endpoint you deleted shows **Deleting** next to it. When it's deleted, the endpoint is removed from the **Endpoints** list.

To delete an endpoint (AWS CLI)

The following example demonstrates using the *DeleteEndpoint* operation with the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws comprehend delete-endpoint \
    --endpoint-arn arn:aws:comprehend:region:account-id endpoint/endpoint name
```

If the action is successful, Amazon Comprehend responds with an HTTP 200 response with an empty HTTP body.

Auto scaling with endpoints

Instead of manually adjusting the number of inference units provisioned for your document classification endpoints and entity recognizer endpoints, you can use auto scaling to automatically set endpoint provisioning to fit your capacity needs.

There are two ways to use auto scaling to adjust the number of inference units provisioned for your endpoint:

- Target tracking: Set auto scaling to adjust endpoint provisioning to fit capacity needs based on usage.
- Scheduled scaling: Set auto scaling to adjust endpoint provisioning to fit capacity needs on a specified schedule.

You can set auto scaling only with the AWS Command Line Interface (AWS CLI). For more information about auto scaling, see What is Application Auto Scaling?

Target tracking

With target tracking, you can adjust endpoint provisioning to fit your capacity needs based on usage. The number of inference units automatically adjust so that the utilized capacity is within a target percentage of the provisioned capacity. You can use target tracking to accommodate temporary surges of use for your document classification endpoints and entity recognizer endpoints. For more information, see Target tracking scaling policies for Application Auto Scaling.

> **Note**
>
> The following examples are formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Setting up target tracking

To set up target tracking for an endpoint, you use AWS CLI commands to register a scalable target and then create a scaling policy. The scalable target defines inference units as the resource used to adjust endpoint provisioning, and the scaling policy defines the metrics that control the auto scaling of the provisioned capacity.

To set up target tracking

1. Register a scalable target. The following examples register a scalable target to adjust endpoint provisioning with a minimum capacity of 1 inference unit and a maximum capacity of 2 inference units.

   For a document classification endpoint, use the following AWS CLI command:

   ```
   aws application-autoscaling register-scalable-target \
       --service-namespace comprehend \
       --resource-id arn:aws:comprehend:region:account-id:document-classifier-
   endpoint/name \
       --scalable-dimension comprehend:document-classifier-
   endpoint:DesiredInferenceUnits \
       --min-capacity 1 \
       --max-capacity 2
   ```

   For an entity recognizer endpoint, use the following AWS CLI command:

```
aws application-autoscaling register-scalable-target \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
endpoint/name \
    --scalable-dimension comprehend:entity-recognizer-
endpoint:DesiredInferenceUnits \
    --min-capacity 1 \
    --max-capacity 2
```

2. To verify the registration of the scalable target, use the following AWS CLI command:

```
aws application-autoscaling describe-scalable-targets \
    --service-namespace comprehend \
    --resource-id endpoint ARN
```

3. Create a target tracking configuration for the scaling policy and save the configuration in a file called `config.json`. The following is an example of a target tracking configuration that automatically adjusts the number of inference units so that utilized capacity is always 70% of the provisioned capacity.

```
{
  "TargetValue": 70,
  "PredefinedMetricSpecification":
  {
  "PredefinedMetricType": "ComprehendInferenceUtilization"
  }
}
```

4. Create a scaling policy. The following examples create a scaling policy based on the target tracking configuration defined in the `config.json` file.

   For a document classification endpoint, use the following AWS CLI command:

```
aws application-autoscaling put-scaling-policy \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:document-classifier-
endpoint/name \
    --scalable-dimension comprehend:document-classifier-
endpoint:DesiredInferenceUnits \
```

```
    --policy-name TestPolicy \
    --policy-type TargetTrackingScaling \
    --target-tracking-scaling-policy-configuration file://config.json
```

For an entity recognizer endpoint, use the following AWS CLI command:

```
aws application-autoscaling put-scaling-policy \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
endpoint/name \
    --scalable-dimension comprehend:entity-recognizer-
endpoint:DesiredInferenceUnits \
    --policy-name TestPolicy \
    --policy-type TargetTrackingScaling \
    --target-tracking-scaling-policy-configuration file://config.json
```

Removing target tracking

To remove target tracking for an endpoint, you use AWS CLI commands to delete the scaling policy and then deregister the scalable target.

To remove target tracking

1. Delete the scaling policy. The following examples delete a specified scaling policy.

   For a document classification endpoint, use the following AWS CLI command:

   ```
   aws application-autoscaling delete-scaling-policy \
       --service-namespace comprehend \
       --resource-id arn:aws:comprehend:region:account-id:document-classifier-
   endpoint/name \
       --scalable-dimension comprehend:document-classifier-
   endpoint:DesiredInferenceUnits \
       --policy-name TestPolicy \
   ```

   For an entity recognizer endpoint, use the following AWS CLI command:

   ```
   aws application-autoscaling delete-scaling-policy \
       --service-namespace comprehend \
   ```

```
    --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
endpoint/name \
    --scalable-dimension comprehend:entity-recognizer-
endpoint:DesiredInferenceUnits \
    --policy-name TestPolicy
```

2. Deregister the scalable target. The following examples deregister a specified scalable target.

   For a document classification endpoint, use the following AWS CLI command:

```
aws application-autoscaling deregister-scalable-target \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:document-classifier-
endpoint/name \
    --scalable-dimension comprehend:document-classifier-
endpoint:DesiredInferenceUnits
```

   For an entity recognizer endpoint, use the following AWS CLI command:

```
aws application-autoscaling deregister-scalable-target \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
endpoint/name \
    --scalable-dimension comprehend:entity-recognizer-
endpoint:DesiredInferenceUnits
```

Scheduled scaling

With scheduled scaling, you can adjust endpoint provisioning to fit your capacity needs on a specified schedule. Scheduled scaling automatically adjusts the number of inference units to accommodate surges of use at specific times. You can use scheduled scaling for document classification endpoints and entity recognizer endpoints. For additional information about scheduled scaling, see Scheduled scaling for Application Auto Scaling.

> **Note**
>
> The following examples are formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Setting up scheduled scaling

To set up scheduled scaling for an endpoint, you use AWS CLI commands to register a scalable target and then create a scheduled action. The scalable target defines inference units as the resource used to adjust endpoint provisioning, and the scheduled action controls the auto scaling of the provisioned capacity at specific times.

To set up scheduled scaling

1. Register a scalable target. The following examples register a scalable target to adjust endpoint provisioning with a minimum capacity of 1 inference unit and a maximum capacity of 2 inference units.

   For a document classification endpoint, use the following AWS CLI command:

   ```
   aws application-autoscaling register-scalable-target \
       --service-namespace comprehend \
       --resource-id arn:aws:comprehend:region:account-id:document-classifier-
   endpoint/name \
       --scalable-dimension comprehend:document-classifier-
   endpoint:DesiredInferenceUnits \
       --min-capacity 1 \
       --max-capacity 2
   ```

   For an entity recognizer endpoint, use the following AWS CLI command:

   ```
   aws application-autoscaling register-scalable-target \
       --service-namespace comprehend \
       --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
   endpoint/name \
       --scalable-dimension comprehend:entity-recognizer-
   endpoint:DesiredInferenceUnits \
       --min-capacity 1 \
       --max-capacity 2
   ```

2. Create a scheduled action. The following examples create a scheduled action to automatically adjust the provisioned capacity every day at 12:00 UTC with a minimum of 2 inference units and a maximum of 5 inference units. For more information about chronological expressions and scheduled scaling, see Schedule expressions.

   For a document classification endpoint, use the following AWS CLI command:

```
aws application-autoscaling put-scheduled-action \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:document-classifier-
endpoint/name \
    --scalable-dimension comprehend:document-classifier-
endpoint:DesiredInferenceUnits \
    --scheduled-action-name TestScheduledAction \
    --schedule "cron(0 12 * * ? *)" \
    --scalable-target-action MinCapacity=2,MaxCapacity=5
```

For an entity recognizer endpoint, use the following AWS CLI command:

```
aws application-autoscaling put-scheduled-action \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
endpoint/name \
    --scalable-dimension comprehend:entity-recognizer-
endpoint:DesiredInferenceUnits \
    --scheduled-action-name TestScheduledAction \
    --schedule "cron(0 12 * * ? *)" \
    --scalable-target-action MinCapacity=2,MaxCapacity=5
```

Removing scheduled scaling

To remove scheduled scaling for an endpoint, you use AWS CLI commands to delete the scheduled action and then deregister the scalable target.

To remove scheduled scaling

1. Delete the scheduled action. The following examples delete a specified scheduled action.

   For a document classification endpoint, use the following AWS CLI command:

```
aws application-autoscaling delete-scheduled-action \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:document-classifier-
endpoint/name \
    --scalable-dimension comprehend:document-classifier-
endpoint:DesiredInferenceUnits \
    --scheduled-action-name TestScheduledAction
```

For an entity recognizer endpoint, use the following AWS CLI command:

```
aws application-autoscaling delete-scheduled-action \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
endpoint/name \
    --scalable-dimension comprehend:entity-recognizer-
endpoint:DesiredInferenceUnits \
    --scheduled-action-name TestScheduledAction
```

2. Deregister the scalable target. The following examples deregister a specified scalable target.

   For a document classification endpoint, use the following AWS CLI command:

```
aws application-autoscaling deregister-scalable-target \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:document-classifier-
endpoint/name \
    --scalable-dimension comprehend:document-classifier-
endpoint:DesiredInferenceUnits
```

   For an entity recognizer endpoint, use the following AWS CLI command:

```
aws application-autoscaling deregister-scalable-target \
    --service-namespace comprehend \
    --resource-id arn:aws:comprehend:region:account-id:entity-recognizer-
endpoint/name \
    --scalable-dimension comprehend:entity-recognizer-
endpoint:DesiredInferenceUnits
```

Tagging your resources

A tag is a key-value pair that you can add to an Amazon Comprehend resource as metadata. You can use tags on **Analysis jobs**, **Custom classification** models, **Custom entity recognition** models, and **endpoints**. Tags have two major functions: organizing your resources and providing tag-based access control.

To organize your resources with tags, you could add the tag key 'Department' and tag values ‘Sales’ or ‘Legal'. You can then search and filter for resources that are pertinent to your company's legal department.

To provide tag-based access control, create IAM policies with permissions based on tags. A policy can allow or disallow an operation based on the tags provided in your request (request-tags) or tags associated with the resource you're calling (resource-tags). For more information on using tags with IAM, see Controlling access using tags in the *IAM User Guide*.

Considerations for using tags with Amazon Comprehend:

- You can add up to 50 tags per resource, and tags can be added at the time you create the resource, or retroactively.
- A tag *key* is a required field but a tag *value* is optional.
- Tags do not have to be unique between resources, but a given resource cannot have duplicate tag keys.
- Tag keys and values are case sensitive.
- A tag key can have a maximum of 127 characters; a tag value can have a maximum of 255 characters.
- The '`aws:`' prefix is reserved for AWS use; you cannot add, edit, or delete tags whose key begins with `aws:`. These tags don't count against your tags-per-resource limit of 50.

> **Note**
>
> If you plan to use your tagging schema across multiple AWS services and resources, remember that other services may have different requirements for allowed characters.

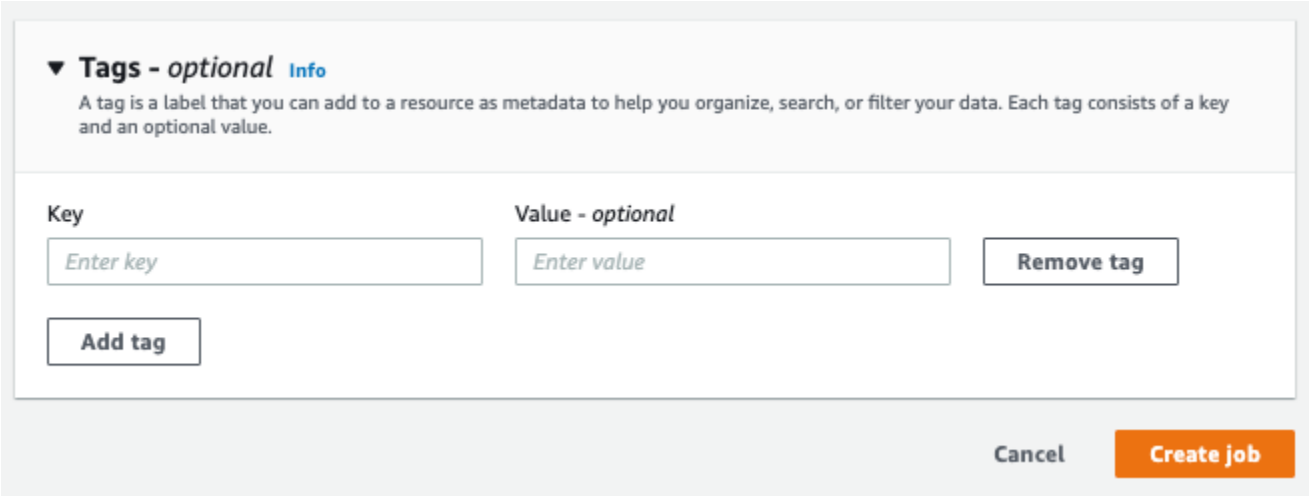
Topics

- Tagging a new resource
- Viewing, editing, and deleting tags associated with a resource

Tagging a new resource

You can add tags to an **Analysis job**, a **Custom classification** model, a **Custom entity recognition** model, or **endpoints**.

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. Select the resource (Analysis job, Custom classification, or Custom entity recognition) you want to create from the left navigation pane.
3. Click **Create job** (or **Create new model**). This takes you to the main 'create' page for your resource. At the bottom of this page, you'll see a '**Tags -** *optional*' panel.



   Enter a tag key and, optionally, a tag value. Choose **Add tag** to add another tag to the resource. Repeat this process until all your tags are added. Note that tag keys must be unique per resource.
4. Select the **Create** or **Create job** button to continue creating your resource.

You can also add tags using the AWS CLI. This example shows how to add tags with the start-entities-detection-job command.

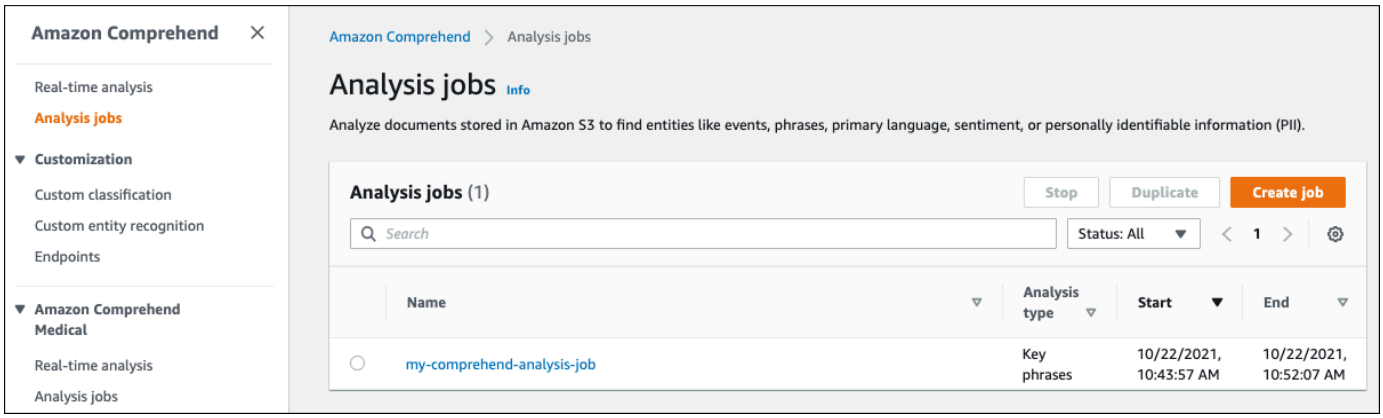
```
aws comprehend start-entities-detection-job \
--language-code "en" \
```

```
--input-data-config "{\"S3Uri\": \"s3://test-input/TEST.csv\"}" \
--output-data-config "{\"S3Uri\": \"s3://test-output\"}" \
--data-access-role-arn arn:aws:iam::123456789012:role/test \
--tags "[{\"Key\": \"color\",\"Value\": \"orange\"}]"
```

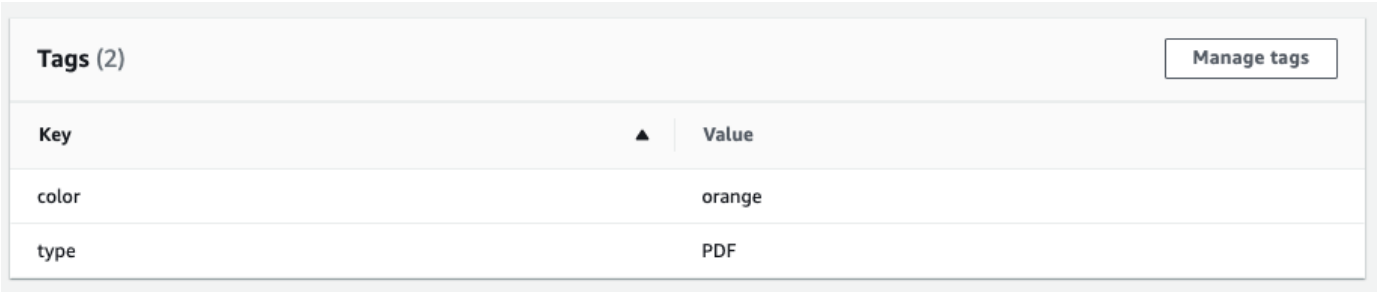
Viewing, editing, and deleting tags associated with a resource

You can view tags associated with an **Analysis job**, a **Custom classification** model, or a **Custom entity recognition** model.

1. Sign in to the AWS Management Console and open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
2. Select the resource (Analysis job, Custom classification, or Custom entity recognition) that contains the file with the tags you want to view, modify, or delete. This displays the list of existing files for your selected resource.

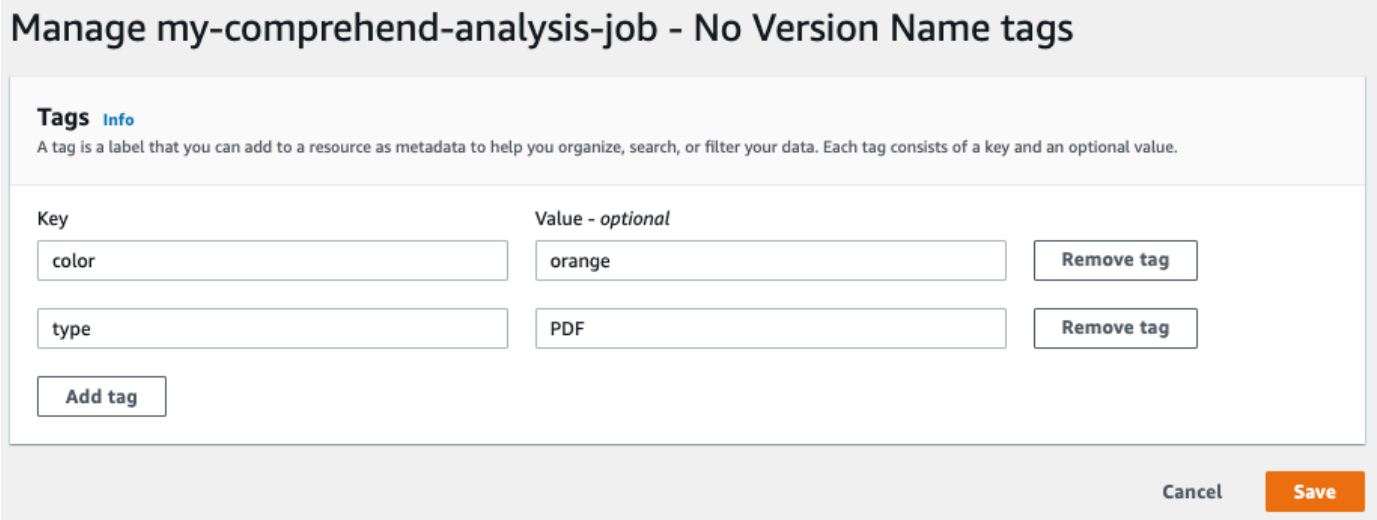


3. Click the name of the file (or model) whose tags you want to view, modify, or delete. This takes you to the details page for that file (or model). Scroll down until you see a **Tags** box. Here, you can see all the tags associated with your selected file (or model).



Select **Manage tags** to edit or remove tags from your resource.

4. Click on the text you want to modify, then edit your tag. You can also remove the tag by selecting **Remove tag**. To add a new tag, select **Add tag**, then enter your desired text in the blank fields.



When you're finished modifying your tags, select **Save**.

Code examples for Amazon Comprehend using AWS SDKs

The following code examples show how to use Amazon Comprehend with an AWS software development kit (SDK).

Actions are code excerpts from larger programs and must be run in context. While actions show you how to call individual service functions, you can see actions in context in their related scenarios and cross-service examples.

Scenarios are code examples that show you how to accomplish a specific task by calling multiple functions within the same service.

Cross-service examples are sample applications that work across multiple AWS services.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Code examples

- Actions for Amazon Comprehend using AWS SDKs
  - Use CreateDocumentClassifier with an AWS SDK or CLI
  - Use DeleteDocumentClassifier with an AWS SDK or CLI
  - Use DescribeDocumentClassificationJob with an AWS SDK or CLI
  - Use DescribeDocumentClassifier with an AWS SDK or CLI
  - Use DescribeTopicsDetectionJob with an AWS SDK or CLI
  - Use DetectDominantLanguage with an AWS SDK or CLI
  - Use DetectEntities with an AWS SDK or CLI
  - Use DetectKeyPhrases with an AWS SDK or CLI
  - Use DetectPiiEntities with an AWS SDK or CLI
  - Use DetectSentiment with an AWS SDK or CLI
  - Use DetectSyntax with an AWS SDK or CLI
  - Use ListDocumentClassificationJobs with an AWS SDK or CLI
  - Use ListDocumentClassifiers with an AWS SDK or CLI

- Use ListTopicsDetectionJobs with an AWS SDK or CLI
- Use StartDocumentClassificationJob with an AWS SDK or CLI
- Use StartTopicsDetectionJob with an AWS SDK or CLI
- Scenarios for Amazon Comprehend using AWS SDKs
  - Detect document elements with Amazon Comprehend and an AWS SDK
  - Run an Amazon Comprehend topic modeling job on sample data using an AWS SDK
  - Train a custom Amazon Comprehend classifier and classify documents using an AWS SDK
- Cross-service examples for Amazon Comprehend using AWS SDKs
  - Build an Amazon Transcribe streaming app
  - Create an Amazon Lex chatbot to engage your website visitors
  - Create a web application that sends and retrieves messages by using Amazon SQS
  - Create an application that analyzes customer feedback and synthesizes audio
  - Detect entities in text extracted from an image using an AWS SDK

Actions for Amazon Comprehend using AWS SDKs

The following code examples demonstrate how to perform individual Amazon Comprehend actions with AWS SDKs. These excerpts call the Amazon Comprehend API and are code excerpts from larger programs that must be run in context. Each example includes a link to GitHub, where you can find instructions for setting up and running the code.

The following examples include only the most commonly used actions. For a complete list, see the Amazon Comprehend API Reference.

Examples

- Use CreateDocumentClassifier with an AWS SDK or CLI
- Use DeleteDocumentClassifier with an AWS SDK or CLI
- Use DescribeDocumentClassificationJob with an AWS SDK or CLI
- Use DescribeDocumentClassifier with an AWS SDK or CLI
- Use DescribeTopicsDetectionJob with an AWS SDK or CLI
- Use DetectDominantLanguage with an AWS SDK or CLI
- Use DetectEntities with an AWS SDK or CLI
- Use DetectKeyPhrases with an AWS SDK or CLI

- Use DetectPiiEntities with an AWS SDK or CLI
- Use DetectSentiment with an AWS SDK or CLI
- Use DetectSyntax with an AWS SDK or CLI
- Use ListDocumentClassificationJobs with an AWS SDK or CLI
- Use ListDocumentClassifiers with an AWS SDK or CLI
- Use ListTopicsDetectionJobs with an AWS SDK or CLI
- Use StartDocumentClassificationJob with an AWS SDK or CLI
- Use StartTopicsDetectionJob with an AWS SDK or CLI

Use `CreateDocumentClassifier` with an AWS SDK or CLI

The following code examples show how to use `CreateDocumentClassifier`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Train a custom classifier and classify documents

CLI

AWS CLI

To create a document classifier to categorize documents

The following `create-document-classifier` example begins the training process for a document classifier model. The training data file, `training.csv`, is located at the `--input-data-config` tag. `training.csv` is a two column document where the labels, or, classifications are provided in the first column and the documents are provided in the second column.

```
aws comprehend create-document-classifier \
    --document-classifier-name example-classifier \
    --data-access-arn arn:aws:comprehend:us-west-2:111122223333:pii-entities-
detection-job/123456abcdeb0e11022f22a11EXAMPLE \
    --input-data-config "S3Uri=s3://DOC-EXAMPLE-BUCKET/" \
    --language-code en
```

Output:

```
{
    "DocumentClassifierArn": "arn:aws:comprehend:us-west-2:111122223333:document-
classifier/example-classifier"
}
```

For more information, see Custom Classification in the *Amazon Comprehend Developer Guide*.

- For API details, see CreateDocumentClassifier in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.comprehend.ComprehendClient;
import software.amazon.awssdk.services.comprehend.model.ComprehendException;
import
 software.amazon.awssdk.services.comprehend.model.CreateDocumentClassifierRequest;
import
 software.amazon.awssdk.services.comprehend.model.CreateDocumentClassifierResponse;
import
 software.amazon.awssdk.services.comprehend.model.DocumentClassifierInputDataConfig;

/**
 * Before running this code example, you can setup the necessary resources, such
 * as the CSV file and IAM Roles, by following this document:
 * https://aws.amazon.com/blogs/machine-learning/building-a-custom-classifier-
using-amazon-comprehend/
 *
 * Also, set up your development environment, including your credentials.
 *
 * For more information, see the following documentation topic:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
```

```
 */
public class DocumentClassifierDemo {
    public static void main(String[] args) {
        final String usage = """

                Usage:    <dataAccessRoleArn> <s3Uri> <documentClassifierName>

                Where:
                  dataAccessRoleArn - The ARN value of the role used for this
 operation.
                  s3Uri - The Amazon S3 bucket that contains the CSV file.
                  documentClassifierName - The name of the document classifier.
                """;

        if (args.length != 3) {
            System.out.println(usage);
            System.exit(1);
        }

        String dataAccessRoleArn = args[0];
        String s3Uri = args[1];
        String documentClassifierName = args[2];

        Region region = Region.US_EAST_1;
        ComprehendClient comClient = ComprehendClient.builder()
                .region(region)
                .build();

        createDocumentClassifier(comClient, dataAccessRoleArn, s3Uri,
 documentClassifierName);
        comClient.close();
    }

    public static void createDocumentClassifier(ComprehendClient comClient,
 String dataAccessRoleArn, String s3Uri,
            String documentClassifierName) {
        try {
            DocumentClassifierInputDataConfig config =
 DocumentClassifierInputDataConfig.builder()
                    .s3Uri(s3Uri)
                    .build();

            CreateDocumentClassifierRequest createDocumentClassifierRequest =
 CreateDocumentClassifierRequest.builder()
```

```
                    .documentClassifierName(documentClassifierName)
                    .dataAccessRoleArn(dataAccessRoleArn)
                    .languageCode("en")
                    .inputDataConfig(config)
                    .build();

            CreateDocumentClassifierResponse createDocumentClassifierResult =
 comClient
                    .createDocumentClassifier(createDocumentClassifierRequest);
            String documentClassifierArn =
 createDocumentClassifierResult.documentClassifierArn();
            System.out.println("Document Classifier ARN: " +
 documentClassifierArn);

        } catch (ComprehendException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- For API details, see CreateDocumentClassifier in *AWS SDK for Java 2.x API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendClassifier:
    """Encapsulates an Amazon Comprehend custom classifier."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client
```

```
        self.classifier_arn = None


    def create(
        self,
        name,
        language_code,
        training_bucket,
        training_key,
        data_access_role_arn,
        mode,
    ):
        """
        Creates a custom classifier. After the classifier is created, it
immediately
        starts training on the data found in the specified Amazon S3 bucket.
Training
        can take 30 minutes or longer. The `describe_document_classifier`
function
        can be used to get training status and returns a status of TRAINED when
the
        classifier is ready to use.

        :param name: The name of the classifier.
        :param language_code: The language the classifier can operate on.
        :param training_bucket: The Amazon S3 bucket that contains the training
data.
        :param training_key: The prefix used to find training data in the
training
                             bucket. If multiple objects have the same prefix,
all
                             of them are used.
        :param data_access_role_arn: The Amazon Resource Name (ARN) of a role
that
                                     grants Comprehend permission to read from
the
                                     training bucket.
        :return: The ARN of the newly created classifier.
        """
        try:
            response = self.comprehend_client.create_document_classifier(
                DocumentClassifierName=name,
                LanguageCode=language_code,
```

```
                InputDataConfig={"S3Uri": f"s3://{training_bucket}/
{training_key}"},
                DataAccessRoleArn=data_access_role_arn,
                Mode=mode.value,
            )
            self.classifier_arn = response["DocumentClassifierArn"]
            logger.info("Started classifier creation. Arn is: %s.",
 self.classifier_arn)
        except ClientError:
            logger.exception("Couldn't create classifier %s.", name)
            raise
        else:
            return self.classifier_arn
```

- For API details, see CreateDocumentClassifier in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DeleteDocumentClassifier` with an AWS SDK or CLI

The following code examples show how to use `DeleteDocumentClassifier`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Train a custom classifier and classify documents

CLI

AWS CLI

To delete a custom document classifier

The following `delete-document-classifier` example deletes a custom document classifier model.

```
aws comprehend delete-document-classifier \
    --document-classifier-arn arn:aws:comprehend:us-west-2:111122223333:document-
classifier/example-classifier-1
```

This command produces no output.

For more information, see Managing Amazon Comprehend endpoints in the *Amazon Comprehend Developer Guide*.

- For API details, see DeleteDocumentClassifier in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendClassifier:
    """Encapsulates an Amazon Comprehend custom classifier."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client
        self.classifier_arn = None


    def delete(self):
        """
        Deletes the classifier.
        """
        try:
            self.comprehend_client.delete_document_classifier(
                DocumentClassifierArn=self.classifier_arn
            )
            logger.info("Deleted classifier %s.", self.classifier_arn)
```

```
            self.classifier_arn = None
        except ClientError:
            logger.exception("Couldn't deleted classifier %s.",
 self.classifier_arn)
            raise
```

- For API details, see DeleteDocumentClassifier in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DescribeDocumentClassificationJob` with an AWS SDK or CLI

The following code examples show how to use `DescribeDocumentClassificationJob`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Train a custom classifier and classify documents

CLI

AWS CLI

To describe a document classification job

The following `describe-document-classification-job` example gets the properties of an asynchronous document classification job.

```
aws comprehend describe-document-classification-job \
    --job-id 123456abcdeb0e11022f22a11EXAMPLE
```

Output:

```
{
    "DocumentClassificationJobProperties": {
        "JobId": "123456abcdeb0e11022f22a11EXAMPLE",
```

```
        "JobArn": "arn:aws:comprehend:us-west-2:111122223333:document-
classification-job/123456abcdeb0e11022f22a11EXAMPLE",
        "JobName": "exampleclassificationjob",
        "JobStatus": "COMPLETED",
        "SubmitTime": "2023-06-14T17:09:51.788000+00:00",
        "EndTime": "2023-06-14T17:15:58.582000+00:00",
        "DocumentClassifierArn": "arn:aws:comprehend:us-
west-2:111122223333:document-classifier/mymodel/version/1",
        "InputDataConfig": {
            "S3Uri": "s3://DOC-EXAMPLE-BUCKET/jobdata/",
            "InputFormat": "ONE_DOC_PER_LINE"
        },
        "OutputDataConfig": {
            "S3Uri": "s3://DOC-EXAMPLE-DESTINATION-BUCKET/
testfolder/111122223333-CLN-123456abcdeb0e11022f22a11EXAMPLE/output/
output.tar.gz"
        },
        "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-servicerole"
    }
}
```

For more information, see Custom Classification in the *Amazon Comprehend Developer Guide*.

- For API details, see DescribeDocumentClassificationJob in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendClassifier:
    """Encapsulates an Amazon Comprehend custom classifier."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
```

```
        """
        self.comprehend_client = comprehend_client
        self.classifier_arn = None


    def describe_job(self, job_id):
        """
        Gets metadata about a classification job.

        :param job_id: The ID of the job to look up.
        :return: Metadata about the job.
        """
        try:
            response =
self.comprehend_client.describe_document_classification_job(
                JobId=job_id
            )
            job = response["DocumentClassificationJobProperties"]
            logger.info("Got classification job %s.", job["JobName"])
        except ClientError:
            logger.exception("Couldn't get classification job %s.", job_id)
            raise
        else:
            return job
```

- For API details, see DescribeDocumentClassificationJob in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DescribeDocumentClassifier` with an AWS SDK or CLI

The following code examples show how to use `DescribeDocumentClassifier`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Train a custom classifier and classify documents

CLI

AWS CLI

To describe a document classifier

The following `describe-document-classifier` example gets the properties of a custom document classifier model.

```
aws comprehend describe-document-classifier \
    --document-classifier-arn arn:aws:comprehend:us-west-2:111122223333:document-
classifier/example-classifier-1
```

Output:

```
{
    "DocumentClassifierProperties": {
        "DocumentClassifierArn": "arn:aws:comprehend:us-
west-2:111122223333:document-classifier/example-classifier-1",
        "LanguageCode": "en",
        "Status": "TRAINED",
        "SubmitTime": "2023-06-13T19:04:15.735000+00:00",
        "EndTime": "2023-06-13T19:42:31.752000+00:00",
        "TrainingStartTime": "2023-06-13T19:08:20.114000+00:00",
        "TrainingEndTime": "2023-06-13T19:41:35.080000+00:00",
        "InputDataConfig": {
            "DataFormat": "COMPREHEND_CSV",
            "S3Uri": "s3://DOC-EXAMPLE-BUCKET/trainingdata"
        },
        "OutputDataConfig": {},
        "ClassifierMetadata": {
            "NumberOfLabels": 3,
            "NumberOfTrainedDocuments": 5016,
            "NumberOfTestDocuments": 557,
            "EvaluationMetrics": {
                "Accuracy": 0.9856,
                "Precision": 0.9919,
                "Recall": 0.9459,
                "F1Score": 0.9673,
                "MicroPrecision": 0.9856,
                "MicroRecall": 0.9856,
                "MicroF1Score": 0.9856,
                "HammingLoss": 0.0144
```

```
            }
        },
        "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-example-role",
        "Mode": "MULTI_CLASS"
    }
}
```

For more information, see Creating and managing custom models in the *Amazon Comprehend Developer Guide*.

- For API details, see DescribeDocumentClassifier in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendClassifier:
    """Encapsulates an Amazon Comprehend custom classifier."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client
        self.classifier_arn = None


    def describe(self, classifier_arn=None):
        """
        Gets metadata about a custom classifier, including its current status.

        :param classifier_arn: The ARN of the classifier to look up.
        :return: Metadata about the classifier.
        """
        if classifier_arn is not None:
```

```
            self.classifier_arn = classifier_arn
        try:
            response = self.comprehend_client.describe_document_classifier(
                DocumentClassifierArn=self.classifier_arn
            )
            classifier = response["DocumentClassifierProperties"]
            logger.info("Got classifier %s.", self.classifier_arn)
        except ClientError:
            logger.exception("Couldn't get classifier %s.", self.classifier_arn)
            raise
        else:
            return classifier
```

- For API details, see DescribeDocumentClassifier in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DescribeTopicsDetectionJob` with an AWS SDK or CLI

The following code examples show how to use `DescribeTopicsDetectionJob`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Run a topic modeling job on sample data

CLI

AWS CLI

To describe a topics detection job

The following `describe-topics-detection-job` example gets the properties of an asynchronous topics detection job.

```
aws comprehend describe-topics-detection-job \
```

```
    --job-id 123456abcdeb0e11022f22a11EXAMPLE
```

Output:

```
{
    "TopicsDetectionJobProperties": {
        "JobId": "123456abcdeb0e11022f22a11EXAMPLE",
        "JobArn": "arn:aws:comprehend:us-west-2:111122223333:topics-detection-
job/123456abcdeb0e11022f22a11EXAMPLE",
        "JobName": "example_topics_detection",
        "JobStatus": "IN_PROGRESS",
        "SubmitTime": "2023-06-09T18:44:43.414000+00:00",
        "InputDataConfig": {
            "S3Uri": "s3://DOC-EXAMPLE-BUCKET",
            "InputFormat": "ONE_DOC_PER_LINE"
        },
        "OutputDataConfig": {
            "S3Uri": "s3://DOC-EXAMPLE-DESTINATION-BUCKET/
testfolder/111122223333-TOPICS-123456abcdeb0e11022f22a11EXAMPLE/output/
output.tar.gz"
        },
        "NumberOfTopics": 10,
        "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-examplerole"
    }
}
```

For more information, see Async analysis for Amazon Comprehend insights in the *Amazon Comprehend Developer Guide*.

- For API details, see DescribeTopicsDetectionJob in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendTopicModeler:
    """Encapsulates a Comprehend topic modeler."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def describe_job(self, job_id):
        """
        Gets metadata about a topic modeling job.

        :param job_id: The ID of the job to look up.
        :return: Metadata about the job.
        """
        try:
            response = self.comprehend_client.describe_topics_detection_job(
                JobId=job_id
            )
            job = response["TopicsDetectionJobProperties"]
            logger.info("Got topic detection job %s.", job_id)
        except ClientError:
            logger.exception("Couldn't get topic detection job %s.", job_id)
            raise
        else:
            return job
```

- For API details, see DescribeTopicsDetectionJob in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DetectDominantLanguage` with an AWS SDK or CLI

The following code examples show how to use `DetectDominantLanguage`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Detect document elements

.NET

AWS SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    using System;
    using System.Threading.Tasks;
    using Amazon.Comprehend;
    using Amazon.Comprehend.Model;

    /// <summary>
    /// This example calls the Amazon Comprehend service to determine the
    /// dominant language.
    /// </summary>
    public static class DetectDominantLanguage
    {
        /// <summary>
        /// Calls Amazon Comprehend to determine the dominant language used in
        /// the sample text.
        /// </summary>
        public static async Task Main()
        {
            string text = "It is raining today in Seattle.";

            var comprehendClient = new
AmazonComprehendClient(Amazon.RegionEndpoint.USWest2);

            Console.WriteLine("Calling DetectDominantLanguage\n");
            var detectDominantLanguageRequest = new
DetectDominantLanguageRequest()
            {
```

```
                Text = text,
            };

            var detectDominantLanguageResponse = await
 comprehendClient.DetectDominantLanguageAsync(detectDominantLanguageRequest);
            foreach (var dl in detectDominantLanguageResponse.Languages)
            {
                Console.WriteLine($"Language Code: {dl.LanguageCode}, Score:
 {dl.Score}");
            }

            Console.WriteLine("Done");
        }
    }
```

- For API details, see DetectDominantLanguage in *AWS SDK for .NET API Reference*.

CLI

AWS CLI

To detect the dominant language of input text

The following `detect-dominant-language` analyzes the input text and identifies the dominant language. The pre-trained model's confidence score is also output.

```
aws comprehend detect-dominant-language \
    --text "It is a beautiful day in Seattle."
```

Output:

```
{
    "Languages": [
        {
            "LanguageCode": "en",
            "Score": 0.9877256155014038
        }
    ]
}
```

For more information, see Dominant Language in the *Amazon Comprehend Developer Guide*.

- For API details, see DetectDominantLanguage in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.comprehend.ComprehendClient;
import software.amazon.awssdk.services.comprehend.model.ComprehendException;
import
 software.amazon.awssdk.services.comprehend.model.DetectDominantLanguageRequest;
import
 software.amazon.awssdk.services.comprehend.model.DetectDominantLanguageResponse;
import software.amazon.awssdk.services.comprehend.model.DominantLanguage;
import java.util.List;

/**
 * Before running this Java V2 code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation topic:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DetectLanguage {
    public static void main(String[] args) {
        // Specify French text - "It is raining today in Seattle".
        String text = "Il pleut aujourd'hui à Seattle";
        Region region = Region.US_EAST_1;

        ComprehendClient comClient = ComprehendClient.builder()
                .region(region)
                .build();
```

```
        System.out.println("Calling DetectDominantLanguage");
        detectTheDominantLanguage(comClient, text);
        comClient.close();
    }

    public static void detectTheDominantLanguage(ComprehendClient comClient,
 String text) {
        try {
            DetectDominantLanguageRequest request =
 DetectDominantLanguageRequest.builder()
                    .text(text)
                    .build();

            DetectDominantLanguageResponse resp =
 comClient.detectDominantLanguage(request);
            List<DominantLanguage> allLanList = resp.languages();
            for (DominantLanguage lang : allLanList) {
                System.out.println("Language is " + lang.languageCode());
            }

        } catch (ComprehendException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- For API details, see DetectDominantLanguage in *AWS SDK for Java 2.x API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendDetect:
```

```
    """Encapsulates Comprehend detection functions."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def detect_languages(self, text):
        """
        Detects languages used in a document.

        :param text: The document to inspect.
        :return: The list of languages along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_dominant_language(Text=text)
            languages = response["Languages"]
            logger.info("Detected %s languages.", len(languages))
        except ClientError:
            logger.exception("Couldn't detect languages.")
            raise
        else:
            return languages
```

- For API details, see DetectDominantLanguage in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DetectEntities` with an AWS SDK or CLI

The following code examples show how to use `DetectEntities`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Detect document elements

.NET

AWS SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    using System;
    using System.Threading.Tasks;
    using Amazon.Comprehend;
    using Amazon.Comprehend.Model;

    /// <summary>
    /// This example shows how to use the AmazonComprehend service detect any
    /// entities in submitted text.
    /// </summary>
    public static class DetectEntities
    {
        /// <summary>
        /// The main method calls the DetectEntitiesAsync method to find any
        /// entities in the sample code.
        /// </summary>
        public static async Task Main()
        {
            string text = "It is raining today in Seattle";

            var comprehendClient = new AmazonComprehendClient();

            Console.WriteLine("Calling DetectEntities\n");
            var detectEntitiesRequest = new DetectEntitiesRequest()
            {
                Text = text,
                LanguageCode = "en",
            };
            var detectEntitiesResponse = await
comprehendClient.DetectEntitiesAsync(detectEntitiesRequest);
```

```
            foreach (var e in detectEntitiesResponse.Entities)
            {
                Console.WriteLine($"Text: {e.Text}, Type: {e.Type}, Score:
 {e.Score}, BeginOffset: {e.BeginOffset}, EndOffset: {e.EndOffset}");
            }

            Console.WriteLine("Done");
        }
    }
```

- For API details, see DetectEntities in *AWS SDK for .NET API Reference*.

CLI

AWS CLI

To detect named entities in input text

The following `detect-entities` example analyzes the input text and returns the named entities. The pre-trained model's confidence score is also output for each prediction.

```
aws comprehend detect-entities \
    --language-code en \
    --text "Hello Zhang Wei, I am John. Your AnyCompany Financial Services, LLC
 credit card \
    account 1111-XXXX-1111-XXXX has a minimum payment of $24.53 that is due by
 July 31st. Based on your autopay settings, \
    we will withdraw your payment on the due date from your bank account number
 XXXXXX1111 with the routing number XXXXX0000. \
    Customer feedback for Sunshine Spa, 123 Main St, Anywhere. Send comments to
 Alice at AnySpa@example.com."
```

Output:

```
{
    "Entities": [
        {
            "Score": 0.9994556307792664,
            "Type": "PERSON",
            "Text": "Zhang Wei",
```

```
            "BeginOffset": 6,
            "EndOffset": 15
        },
        {
            "Score": 0.9981022477149963,
            "Type": "PERSON",
            "Text": "John",
            "BeginOffset": 22,
            "EndOffset": 26
        },
        {
            "Score": 0.9986887574195862,
            "Type": "ORGANIZATION",
            "Text": "AnyCompany Financial Services, LLC",
            "BeginOffset": 33,
            "EndOffset": 67
        },
        {
            "Score": 0.9959119558334351,
            "Type": "OTHER",
            "Text": "1111-XXXX-1111-XXXX",
            "BeginOffset": 88,
            "EndOffset": 107
        },
        {
            "Score": 0.9708039164543152,
            "Type": "QUANTITY",
            "Text": ".53",
            "BeginOffset": 133,
            "EndOffset": 136
        },
        {
            "Score": 0.9987268447875977,
            "Type": "DATE",
            "Text": "July 31st",
            "BeginOffset": 152,
            "EndOffset": 161
        },
        {
            "Score": 0.9858865737915039,
            "Type": "OTHER",
            "Text": "XXXXXX1111",
            "BeginOffset": 271,
            "EndOffset": 281
```

```
        },
        {
            "Score": 0.9700471758842468,
            "Type": "OTHER",
            "Text": "XXXXX0000",
            "BeginOffset": 306,
            "EndOffset": 315
        },
        {
            "Score": 0.9591118693351746,
            "Type": "ORGANIZATION",
            "Text": "Sunshine Spa",
            "BeginOffset": 340,
            "EndOffset": 352
        },
        {
            "Score": 0.9797496795654297,
            "Type": "LOCATION",
            "Text": "123 Main St",
            "BeginOffset": 354,
            "EndOffset": 365
        },
        {
            "Score": 0.994929313659668,
            "Type": "PERSON",
            "Text": "Alice",
            "BeginOffset": 394,
            "EndOffset": 399
        },
        {
            "Score": 0.9949769377708435,
            "Type": "OTHER",
            "Text": "AnySpa@example.com",
            "BeginOffset": 403,
            "EndOffset": 418
        }
    ]
}
```

For more information, see Entities in the *Amazon Comprehend Developer Guide*.

- For API details, see DetectEntities in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.comprehend.ComprehendClient;
import software.amazon.awssdk.services.comprehend.model.DetectEntitiesRequest;
import software.amazon.awssdk.services.comprehend.model.DetectEntitiesResponse;
import software.amazon.awssdk.services.comprehend.model.Entity;
import software.amazon.awssdk.services.comprehend.model.ComprehendException;
import java.util.List;

/**
 * Before running this Java V2 code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation topic:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DetectEntities {
    public static void main(String[] args) {
        String text = "Amazon.com, Inc. is located in Seattle, WA and was founded
 July 5th, 1994 by Jeff Bezos, allowing customers to buy everything from books to
 blenders. Seattle is north of Portland and south of Vancouver, BC. Other notable
 Seattle - based companies are Starbucks and Boeing.";
       Region region = Region.US_EAST_1;
       ComprehendClient comClient = ComprehendClient.builder()
               .region(region)
               .build();

       System.out.println("Calling DetectEntities");
       detectAllEntities(comClient, text);
       comClient.close();
    }
```

```
    public static void detectAllEntities(ComprehendClient comClient, String text)
 {
        try {
            DetectEntitiesRequest detectEntitiesRequest =
 DetectEntitiesRequest.builder()
                    .text(text)
                    .languageCode("en")
                    .build();

            DetectEntitiesResponse detectEntitiesResult =
 comClient.detectEntities(detectEntitiesRequest);
            List<Entity> entList = detectEntitiesResult.entities();
            for (Entity entity : entList) {
                System.out.println("Entity text is " + entity.text());
            }

        } catch (ComprehendException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- For API details, see DetectEntities in *AWS SDK for Java 2.x API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendDetect:
    """Encapsulates Comprehend detection functions."""

    def __init__(self, comprehend_client):
        """
```

```
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def detect_entities(self, text, language_code):
        """
        Detects entities in a document. Entities can be things like people and
places
        or other common terms.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of entities along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_entities(
                Text=text, LanguageCode=language_code
            )
            entities = response["Entities"]
            logger.info("Detected %s entities.", len(entities))
        except ClientError:
            logger.exception("Couldn't detect entities.")
            raise
        else:
            return entities
```

- For API details, see DetectEntities in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DetectKeyPhrases` with an AWS SDK or CLI

The following code examples show how to use `DetectKeyPhrases`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Detect document elements

.NET

AWS SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    using System;
    using System.Threading.Tasks;
    using Amazon.Comprehend;
    using Amazon.Comprehend.Model;

    /// <summary>
    /// This example shows how to use the Amazon Comprehend service to
    /// search text for key phrases.
    /// </summary>
    public static class DetectKeyPhrase
    {
        /// <summary>
        /// This method calls the Amazon Comprehend method DetectKeyPhrasesAsync
        /// to detect any key phrases in the sample text.
        /// </summary>
        public static async Task Main()
        {
            string text = "It is raining today in Seattle";

            var comprehendClient = new
AmazonComprehendClient(Amazon.RegionEndpoint.USWest2);

            // Call DetectKeyPhrases API
            Console.WriteLine("Calling DetectKeyPhrases");
            var detectKeyPhrasesRequest = new DetectKeyPhrasesRequest()
            {
                Text = text,
                LanguageCode = "en",
            };
```

```
            var detectKeyPhrasesResponse = await
 comprehendClient.DetectKeyPhrasesAsync(detectKeyPhrasesRequest);
            foreach (var kp in detectKeyPhrasesResponse.KeyPhrases)
            {
                Console.WriteLine($"Text: {kp.Text}, Score: {kp.Score},
 BeginOffset: {kp.BeginOffset}, EndOffset: {kp.EndOffset}");
            }

            Console.WriteLine("Done");
        }
    }
```

- For API details, see DetectKeyPhrases in *AWS SDK for .NET API Reference*.

CLI

AWS CLI

To detect key phrases in input text

The following `detect-key-phrases` example analyzes the input text and identifies the key noun phrases. The pre-trained model's confidence score is also output for each prediction.

```
aws comprehend detect-key-phrases \
    --language-code en \
    --text "Hello Zhang Wei, I am John. Your AnyCompany Financial Services, LLC
 credit card \
        account 1111-XXXX-1111-XXXX has a minimum payment of $24.53 that is due
 by July 31st. Based on your autopay settings, \
        we will withdraw your payment on the due date from your bank account
 number XXXXXX1111 with the routing number XXXXX0000. \
        Customer feedback for Sunshine Spa, 123 Main St, Anywhere. Send comments
 to Alice at AnySpa@example.com."
```

Output:

```
{
    "KeyPhrases": [
        {
            "Score": 0.8996376395225525,
```

```
            "Text": "Zhang Wei",
            "BeginOffset": 6,
            "EndOffset": 15
        },
        {
            "Score": 0.9992469549179077,
            "Text": "John",
            "BeginOffset": 22,
            "EndOffset": 26
        },
        {
            "Score": 0.988385021686554,
            "Text": "Your AnyCompany Financial Services",
            "BeginOffset": 28,
            "EndOffset": 62
        },
        {
            "Score": 0.8740853071212769,
            "Text": "LLC credit card account 1111-XXXX-1111-XXXX",
            "BeginOffset": 64,
            "EndOffset": 107
        },
        {
            "Score": 0.9999437928199768,
            "Text": "a minimum payment",
            "BeginOffset": 112,
            "EndOffset": 129
        },
        {
            "Score": 0.9998900890350342,
            "Text": ".53",
            "BeginOffset": 133,
            "EndOffset": 136
        },
        {
            "Score": 0.9979453086853027,
            "Text": "July 31st",
            "BeginOffset": 152,
            "EndOffset": 161
        },
        {
            "Score": 0.9983011484146118,
            "Text": "your autopay settings",
            "BeginOffset": 172,
```

```
            "EndOffset": 193
        },
        {
            "Score": 0.9996572136878967,
            "Text": "your payment",
            "BeginOffset": 211,
            "EndOffset": 223
        },
        {
            "Score": 0.9995037317276001,
            "Text": "the due date",
            "BeginOffset": 227,
            "EndOffset": 239
        },
        {
            "Score": 0.9702621698379517,
            "Text": "your bank account number XXXXXX1111",
            "BeginOffset": 245,
            "EndOffset": 280
        },
        {
            "Score": 0.9179925918579102,
            "Text": "the routing number XXXXX0000.Customer feedback",
            "BeginOffset": 286,
            "EndOffset": 332
        },
        {
            "Score": 0.9978160858154297,
            "Text": "Sunshine Spa",
            "BeginOffset": 337,
            "EndOffset": 349
        },
        {
            "Score": 0.9706913232803345,
            "Text": "123 Main St",
            "BeginOffset": 351,
            "EndOffset": 362
        },
        {
            "Score": 0.9941995143890381,
            "Text": "comments",
            "BeginOffset": 379,
            "EndOffset": 387
        },
```

```
        {
            "Score": 0.9759287238121033,
            "Text": "Alice",
            "BeginOffset": 391,
            "EndOffset": 396
        },
        {
            "Score": 0.8376792669296265,
            "Text": "AnySpa@example.com",
            "BeginOffset": 400,
            "EndOffset": 415
        }
    ]
}
```

For more information, see Key Phrases in the *Amazon Comprehend Developer Guide*.

- For API details, see DetectKeyPhrases in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.comprehend.ComprehendClient;
import software.amazon.awssdk.services.comprehend.model.DetectKeyPhrasesRequest;
import software.amazon.awssdk.services.comprehend.model.DetectKeyPhrasesResponse;
import software.amazon.awssdk.services.comprehend.model.KeyPhrase;
import software.amazon.awssdk.services.comprehend.model.ComprehendException;
import java.util.List;

/**
 * Before running this Java V2 code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation topic:
```

```
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DetectKeyPhrases {
    public static void main(String[] args) {
        String text = "Amazon.com, Inc. is located in Seattle, WA and was founded
 July 5th, 1994 by Jeff Bezos, allowing customers to buy everything from books to
 blenders. Seattle is north of Portland and south of Vancouver, BC. Other notable
 Seattle - based companies are Starbucks and Boeing.";
        Region region = Region.US_EAST_1;
        ComprehendClient comClient = ComprehendClient.builder()
                .region(region)
                .build();

        System.out.println("Calling DetectKeyPhrases");
        detectAllKeyPhrases(comClient, text);
        comClient.close();
    }

    public static void detectAllKeyPhrases(ComprehendClient comClient, String
 text) {
        try {
            DetectKeyPhrasesRequest detectKeyPhrasesRequest =
 DetectKeyPhrasesRequest.builder()
                    .text(text)
                    .languageCode("en")
                    .build();

            DetectKeyPhrasesResponse detectKeyPhrasesResult =
 comClient.detectKeyPhrases(detectKeyPhrasesRequest);
            List<KeyPhrase> phraseList = detectKeyPhrasesResult.keyPhrases();
            for (KeyPhrase keyPhrase : phraseList) {
                System.out.println("Key phrase text is " + keyPhrase.text());
            }

        } catch (ComprehendException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- For API details, see DetectKeyPhrases in *AWS SDK for Java 2.x API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendDetect:
    """Encapsulates Comprehend detection functions."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def detect_key_phrases(self, text, language_code):
        """
        Detects key phrases in a document. A key phrase is typically a noun and
 its
        modifiers.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of key phrases along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_key_phrases(
                Text=text, LanguageCode=language_code
            )
            phrases = response["KeyPhrases"]
            logger.info("Detected %s phrases.", len(phrases))
        except ClientError:
            logger.exception("Couldn't detect phrases.")
            raise
        else:
```

```
            return phrases
```

- For API details, see DetectKeyPhrases in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DetectPiiEntities` with an AWS SDK or CLI

The following code examples show how to use `DetectPiiEntities`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Detect document elements

.NET

AWS SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
using System;
using System.Threading.Tasks;
using Amazon.Comprehend;
using Amazon.Comprehend.Model;

/// <summary>
/// This example shows how to use the Amazon Comprehend service to find
/// personally identifiable information (PII) within text submitted to the
/// DetectPiiEntitiesAsync method.
/// </summary>
```

```
    public class DetectingPII
    {
        /// <summary>
        /// This method calls the DetectPiiEntitiesAsync method to locate any
        /// personally dientifiable information within the supplied text.
        /// </summary>
        public static async Task Main()
        {
            var comprehendClient = new AmazonComprehendClient();
            var text = @"Hello Paul Santos. The latest statement for your
                         credit card account 1111-0000-1111-0000 was
                         mailed to 123 Any Street, Seattle, WA 98109.";

            var request = new DetectPiiEntitiesRequest
            {
                Text = text,
                LanguageCode = "EN",
            };

            var response = await
comprehendClient.DetectPiiEntitiesAsync(request);

            if (response.Entities.Count > 0)
            {
                foreach (var entity in response.Entities)
                {
                    var entityValue = text.Substring(entity.BeginOffset,
entity.EndOffset - entity.BeginOffset);
                    Console.WriteLine($"{entity.Type}: {entityValue}");
                }
            }
        }
    }
```

- For API details, see DetectPiiEntities in *AWS SDK for .NET API Reference*.

CLI

AWS CLI

To detect pii entities in input text

The following `detect-pii-entities` example analyzes the input text and identifies entities that contain personally identifiable information (PII). The pre-trained model's confidence score is also output for each prediction.

```
aws comprehend detect-pii-entities \
    --language-code en \
    --text "Hello Zhang Wei, I am John. Your AnyCompany Financial Services, LLC
 credit card \
        account 1111-XXXX-1111-XXXX has a minimum payment of $24.53 that is due
 by July 31st. Based on your autopay settings, \
        we will withdraw your payment on the due date from your bank account
 number XXXXXX1111 with the routing number XXXXX0000. \
        Customer feedback for Sunshine Spa, 123 Main St, Anywhere. Send comments
 to Alice at AnySpa@example.com."
```

Output:

```
{
    "Entities": [
        {
            "Score": 0.9998322129249573,
            "Type": "NAME",
            "BeginOffset": 6,
            "EndOffset": 15
        },
        {
            "Score": 0.9998878240585327,
            "Type": "NAME",
            "BeginOffset": 22,
            "EndOffset": 26
        },
        {
            "Score": 0.9994089603424072,
            "Type": "CREDIT_DEBIT_NUMBER",
            "BeginOffset": 88,
            "EndOffset": 107
        },
        {
            "Score": 0.9999760985374451,
            "Type": "DATE_TIME",
            "BeginOffset": 152,
            "EndOffset": 161
        },
```

```
        {
            "Score": 0.9999449253082275,
            "Type": "BANK_ACCOUNT_NUMBER",
            "BeginOffset": 271,
            "EndOffset": 281
        },
        {
            "Score": 0.9999847412109375,
            "Type": "BANK_ROUTING",
            "BeginOffset": 306,
            "EndOffset": 315
        },
        {
            "Score": 0.999925434589386,
            "Type": "ADDRESS",
            "BeginOffset": 354,
            "EndOffset": 365
        },
        {
            "Score": 0.9989161491394043,
            "Type": "NAME",
            "BeginOffset": 394,
            "EndOffset": 399
        },
        {
            "Score": 0.9994171857833862,
            "Type": "EMAIL",
            "BeginOffset": 403,
            "EndOffset": 418
        }
    ]
}
```

For more information, see Personally Identifiable Information (PII) in the *Amazon Comprehend Developer Guide*.

- For API details, see DetectPiiEntities in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendDetect:
    """Encapsulates Comprehend detection functions."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def detect_pii(self, text, language_code):
        """
        Detects personally identifiable information (PII) in a document. PII can
 be
        things like names, account numbers, or addresses.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of PII entities along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_pii_entities(
                Text=text, LanguageCode=language_code
            )
            entities = response["Entities"]
            logger.info("Detected %s PII entities.", len(entities))
        except ClientError:
            logger.exception("Couldn't detect PII entities.")
            raise
        else:
            return entities
```

- For API details, see DetectPiiEntities in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DetectSentiment` with an AWS SDK or CLI

The following code examples show how to use `DetectSentiment`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Detect document elements

.NET

AWS SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
using System;
using System.Threading.Tasks;
using Amazon.Comprehend;
using Amazon.Comprehend.Model;

/// <summary>
/// This example shows how to detect the overall sentiment of the supplied
/// text using the Amazon Comprehend service.
/// </summary>
public static class DetectSentiment
{
    /// <summary>
```

```
        /// This method calls the DetetectSentimentAsync method to analyze the
        /// supplied text and determine the overal sentiment.
        /// </summary>
        public static async Task Main()
        {
            string text = "It is raining today in Seattle";

            var comprehendClient = new
AmazonComprehendClient(Amazon.RegionEndpoint.USWest2);

            // Call DetectKeyPhrases API
            Console.WriteLine("Calling DetectSentiment");
            var detectSentimentRequest = new DetectSentimentRequest()
            {
                Text = text,
                LanguageCode = "en",
            };
            var detectSentimentResponse = await
comprehendClient.DetectSentimentAsync(detectSentimentRequest);
            Console.WriteLine($"Sentiment: {detectSentimentResponse.Sentiment}");
            Console.WriteLine("Done");
        }
    }
```

- For API details, see DetectSentiment in *AWS SDK for .NET API Reference*.

CLI

AWS CLI

To detect the sentiment of an input text

The following `detect-sentiment` example analyzes the input text and returns an inference of the prevailing sentiment (`POSITIVE`, `NEUTRAL`, `MIXED`, or `NEGATIVE`).

```
aws comprehend detect-sentiment \
    --language-code en \
    --text "It is a beautiful day in Seattle"
```

Output:

```
{
    "Sentiment": "POSITIVE",
    "SentimentScore": {
        "Positive": 0.9976957440376282,
        "Negative": 9.653854067437351e-05,
        "Neutral": 0.002169104292988777,
        "Mixed": 3.857641786453314e-05
    }
}
```

For more information, see Sentiment in the *Amazon Comprehend Developer Guide*

- For API details, see DetectSentiment in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.comprehend.ComprehendClient;
import software.amazon.awssdk.services.comprehend.model.ComprehendException;
import software.amazon.awssdk.services.comprehend.model.DetectSentimentRequest;
import software.amazon.awssdk.services.comprehend.model.DetectSentimentResponse;

/**
 * Before running this Java V2 code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation topic:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DetectSentiment {
    public static void main(String[] args) {
```

```
        String text = "Amazon.com, Inc. is located in Seattle, WA and was founded
 July 5th, 1994 by Jeff Bezos, allowing customers to buy everything from books to
 blenders. Seattle is north of Portland and south of Vancouver, BC. Other notable
 Seattle - based companies are Starbucks and Boeing.";
        Region region = Region.US_EAST_1;
        ComprehendClient comClient = ComprehendClient.builder()
                .region(region)
                .build();

        System.out.println("Calling DetectSentiment");
        detectSentiments(comClient, text);
        comClient.close();
    }

    public static void detectSentiments(ComprehendClient comClient, String text)
 {
        try {
            DetectSentimentRequest detectSentimentRequest =
 DetectSentimentRequest.builder()
                    .text(text)
                    .languageCode("en")
                    .build();

            DetectSentimentResponse detectSentimentResult =
 comClient.detectSentiment(detectSentimentRequest);
            System.out.println("The Neutral value is " +
 detectSentimentResult.sentimentScore().neutral());

        } catch (ComprehendException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- For API details, see DetectSentiment in *AWS SDK for Java 2.x API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendDetect:
    """Encapsulates Comprehend detection functions."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def detect_sentiment(self, text, language_code):
        """
        Detects the overall sentiment expressed in a document. Sentiment can
        be positive, negative, neutral, or a mixture.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The sentiments along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_sentiment(
                Text=text, LanguageCode=language_code
            )
            logger.info("Detected primary sentiment %s.", response["Sentiment"])
        except ClientError:
            logger.exception("Couldn't detect sentiment.")
            raise
        else:
            return response
```

- For API details, see DetectSentiment in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `DetectSyntax` with an AWS SDK or CLI

The following code examples show how to use `DetectSyntax`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Detect document elements

.NET

AWS SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
using System;
using System.Threading.Tasks;
using Amazon.Comprehend;
using Amazon.Comprehend.Model;

/// <summary>
/// This example shows how to use Amazon Comprehend to detect syntax
/// elements by calling the DetectSyntaxAsync method.
/// </summary>
public class DetectingSyntax
{
    /// <summary>
    /// This method calls DetectSynaxAsync to identify the syntax elements
    /// in the sample text.
    /// </summary>
```

```
        public static async Task Main()
        {
            string text = "It is raining today in Seattle";

            var comprehendClient = new AmazonComprehendClient();

            // Call DetectSyntax API
            Console.WriteLine("Calling DetectSyntaxAsync\n");
            var detectSyntaxRequest = new DetectSyntaxRequest()
            {
                Text = text,
                LanguageCode = "en",
            };
            DetectSyntaxResponse detectSyntaxResponse = await
 comprehendClient.DetectSyntaxAsync(detectSyntaxRequest);
            foreach (SyntaxToken s in detectSyntaxResponse.SyntaxTokens)
            {
                Console.WriteLine($"Text: {s.Text}, PartOfSpeech:
 {s.PartOfSpeech.Tag}, BeginOffset: {s.BeginOffset}, EndOffset: {s.EndOffset}");
            }

            Console.WriteLine("Done");
        }
    }
```

- For API details, see DetectSyntax in *AWS SDK for .NET API Reference*.

CLI

AWS CLI

To detect the parts of speech in an input text

The following `detect-syntax` example analyzes the syntax of the input text and returns the different parts of speech. The pre-trained model's confidence score is also output for each prediction.

```
aws comprehend detect-syntax \
    --language-code en \
    --text "It is a beautiful day in Seattle."
```

Output:


```
{
    "SyntaxTokens": [
        {
            "TokenId": 1,
            "Text": "It",
            "BeginOffset": 0,
            "EndOffset": 2,
            "PartOfSpeech": {
                "Tag": "PRON",
                "Score": 0.9999740719795227
            }
        },
        {
            "TokenId": 2,
            "Text": "is",
            "BeginOffset": 3,
            "EndOffset": 5,
            "PartOfSpeech": {
                "Tag": "VERB",
                "Score": 0.999901294708252
            }
        },
        {
            "TokenId": 3,
            "Text": "a",
            "BeginOffset": 6,
            "EndOffset": 7,
            "PartOfSpeech": {
                "Tag": "DET",
                "Score": 0.9999938607215881
            }
        },
        {
            "TokenId": 4,
            "Text": "beautiful",
            "BeginOffset": 8,
            "EndOffset": 17,
            "PartOfSpeech": {
                "Tag": "ADJ",
                "Score": 0.9987351894378662
            }
        },
```

```
        {
            "TokenId": 5,
            "Text": "day",
            "BeginOffset": 18,
            "EndOffset": 21,
            "PartOfSpeech": {
                "Tag": "NOUN",
                "Score": 0.9999796748161316
            }
        },
        {
            "TokenId": 6,
            "Text": "in",
            "BeginOffset": 22,
            "EndOffset": 24,
            "PartOfSpeech": {
                "Tag": "ADP",
                "Score": 0.9998047947883606
            }
        },
        {
            "TokenId": 7,
            "Text": "Seattle",
            "BeginOffset": 25,
            "EndOffset": 32,
            "PartOfSpeech": {
                "Tag": "PROPN",
                "Score": 0.9940530061721802
            }
        }
    ]
}
```

For more information, see Syntax Analysis in the *Amazon Comprehend Developer Guide*.

- For API details, see DetectSyntax in *AWS CLI Command Reference*.

Java

SDK for Java 2.x

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
import software.amazon.awssdk.regions.Region;
import software.amazon.awssdk.services.comprehend.ComprehendClient;
import software.amazon.awssdk.services.comprehend.model.ComprehendException;
import software.amazon.awssdk.services.comprehend.model.DetectSyntaxRequest;
import software.amazon.awssdk.services.comprehend.model.DetectSyntaxResponse;
import software.amazon.awssdk.services.comprehend.model.SyntaxToken;
import java.util.List;

/**
 * Before running this Java V2 code example, set up your development
 * environment, including your credentials.
 *
 * For more information, see the following documentation topic:
 *
 * https://docs.aws.amazon.com/sdk-for-java/latest/developer-guide/get-
started.html
 */
public class DetectSyntax {
    public static void main(String[] args) {
        String text = "Amazon.com, Inc. is located in Seattle, WA and was founded
 July 5th, 1994 by Jeff Bezos, allowing customers to buy everything from books to
 blenders. Seattle is north of Portland and south of Vancouver, BC. Other notable
 Seattle - based companies are Starbucks and Boeing.";
       Region region = Region.US_EAST_1;
       ComprehendClient comClient = ComprehendClient.builder()
               .region(region)
               .build();

       System.out.println("Calling DetectSyntax");
       detectAllSyntax(comClient, text);
       comClient.close();
    }
```

```
    public static void detectAllSyntax(ComprehendClient comClient, String text) {
        try {
            DetectSyntaxRequest detectSyntaxRequest =
 DetectSyntaxRequest.builder()
                    .text(text)
                    .languageCode("en")
                    .build();

            DetectSyntaxResponse detectSyntaxResult =
 comClient.detectSyntax(detectSyntaxRequest);
            List<SyntaxToken> syntaxTokens = detectSyntaxResult.syntaxTokens();
            for (SyntaxToken token : syntaxTokens) {
                System.out.println("Language is " + token.text());
                System.out.println("Part of speech is " +
 token.partOfSpeech().tagAsString());
            }

        } catch (ComprehendException e) {
            System.err.println(e.awsErrorDetails().errorMessage());
            System.exit(1);
        }
    }
}
```

- For API details, see DetectSyntax in *AWS SDK for Java 2.x API Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendDetect:
    """Encapsulates Comprehend detection functions."""

    def __init__(self, comprehend_client):
```

```
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def detect_syntax(self, text, language_code):
        """
        Detects syntactical elements of a document. Syntax tokens are portions of
        text along with their use as parts of speech, such as nouns, verbs, and
        interjections.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of syntax tokens along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_syntax(
                Text=text, LanguageCode=language_code
            )
            tokens = response["SyntaxTokens"]
            logger.info("Detected %s syntax tokens.", len(tokens))
        except ClientError:
            logger.exception("Couldn't detect syntax.")
            raise
        else:
            return tokens
```

- For API details, see DetectSyntax in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `ListDocumentClassificationJobs` with an AWS SDK or CLI

The following code examples show how to use `ListDocumentClassificationJobs`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Train a custom classifier and classify documents

CLI

AWS CLI

To list of all document classification jobs

The following `list-document-classification-jobs` example lists all document classification jobs.

```
aws comprehend list-document-classification-jobs
```

Output:

```
{
    "DocumentClassificationJobPropertiesList": [
        {
            "JobId": "123456abcdeb0e11022f22a11EXAMPLE",
            "JobArn": "arn:aws:comprehend:us-west-2:1234567890101:document-
classification-job/123456abcdeb0e11022f22a11EXAMPLE",
            "JobName": "exampleclassificationjob",
            "JobStatus": "COMPLETED",
            "SubmitTime": "2023-06-14T17:09:51.788000+00:00",
            "EndTime": "2023-06-14T17:15:58.582000+00:00",
            "DocumentClassifierArn": "arn:aws:comprehend:us-
west-2:1234567890101:document-classifier/mymodel/version/12",
            "InputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-BUCKET/jobdata/",
                "InputFormat": "ONE_DOC_PER_LINE"
            },
            "OutputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-DESTINATION-BUCKET/
thefolder/1234567890101-CLN-e758dd56b824aa717ceab551f11749fb/output/
output.tar.gz"
            },
            "DataAccessRoleArn": "arn:aws:iam::1234567890101:role/service-role/
AmazonComprehendServiceRole-example-role"
        },
```

```
        {
            "JobId": "123456abcdeb0e11022f22a1EXAMPLE2",
            "JobArn": "arn:aws:comprehend:us-west-2:1234567890101:document-
classification-job/123456abcdeb0e11022f22a1EXAMPLE2",
            "JobName": "exampleclassificationjob2",
            "JobStatus": "COMPLETED",
            "SubmitTime": "2023-06-14T17:22:39.829000+00:00",
            "EndTime": "2023-06-14T17:28:46.107000+00:00",
            "DocumentClassifierArn": "arn:aws:comprehend:us-
west-2:1234567890101:document-classifier/mymodel/version/12",
            "InputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-BUCKET/jobdata/",
                "InputFormat": "ONE_DOC_PER_LINE"
            },
            "OutputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-DESTINATION-BUCKET/
thefolder/1234567890101-CLN-123456abcdeb0e11022f22a1EXAMPLE2/output/
output.tar.gz"
            },
            "DataAccessRoleArn": "arn:aws:iam::1234567890101:role/service-role/
AmazonComprehendServiceRole-example-role"
        }
    ]
}
```

For more information, see Custom Classification in the *Amazon Comprehend Developer Guide*.

- For API details, see ListDocumentClassificationJobs in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendClassifier:
    """Encapsulates an Amazon Comprehend custom classifier."""
```

```
    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client
        self.classifier_arn = None


    def list_jobs(self):
        """
        Lists the classification jobs for the current account.

        :return: The list of jobs.
        """
        try:
            response = self.comprehend_client.list_document_classification_jobs()
            jobs = response["DocumentClassificationJobPropertiesList"]
            logger.info("Got %s document classification jobs.", len(jobs))
        except ClientError:
            logger.exception(
                "Couldn't get document classification jobs.",
            )
            raise
        else:
            return jobs
```

- For API details, see ListDocumentClassificationJobs in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `ListDocumentClassifiers` with an AWS SDK or CLI

The following code examples show how to use `ListDocumentClassifiers`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Train a custom classifier and classify documents

CLI

AWS CLI

To list of all document classifiers

The following `list-document-classifiers` example lists all trained and in-training document classifier models.

```
aws comprehend list-document-classifiers
```

Output:

```
{
    "DocumentClassifierPropertiesList": [
        {
            "DocumentClassifierArn": "arn:aws:comprehend:us-
west-2:111122223333:document-classifier/exampleclassifier1",
            "LanguageCode": "en",
            "Status": "TRAINED",
            "SubmitTime": "2023-06-13T19:04:15.735000+00:00",
            "EndTime": "2023-06-13T19:42:31.752000+00:00",
            "TrainingStartTime": "2023-06-13T19:08:20.114000+00:00",
            "TrainingEndTime": "2023-06-13T19:41:35.080000+00:00",
            "InputDataConfig": {
                "DataFormat": "COMPREHEND_CSV",
                "S3Uri": "s3://DOC-EXAMPLE-BUCKET/trainingdata"
            },
            "OutputDataConfig": {},
            "ClassifierMetadata": {
                "NumberOfLabels": 3,
                "NumberOfTrainedDocuments": 5016,
                "NumberOfTestDocuments": 557,
                "EvaluationMetrics": {
                    "Accuracy": 0.9856,
                    "Precision": 0.9919,
                    "Recall": 0.9459,
                    "F1Score": 0.9673,
                    "MicroPrecision": 0.9856,
                    "MicroRecall": 0.9856,
```

```
                    "MicroF1Score": 0.9856,
                    "HammingLoss": 0.0144
                }
            },
            "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-testorle",
            "Mode": "MULTI_CLASS"
        },
        {
            "DocumentClassifierArn": "arn:aws:comprehend:us-
west-2:111122223333:document-classifier/exampleclassifier2",
            "LanguageCode": "en",
            "Status": "TRAINING",
            "SubmitTime": "2023-06-13T21:20:28.690000+00:00",
            "InputDataConfig": {
                "DataFormat": "COMPREHEND_CSV",
                "S3Uri": "s3://DOC-EXAMPLE-BUCKET/trainingdata"
            },
            "OutputDataConfig": {},
            "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-testorle",
            "Mode": "MULTI_CLASS"
        }
    ]
}
```

For more information, see Creating and managing custom models in the *Amazon Comprehend Developer Guide*.

- For API details, see ListDocumentClassifiers in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendClassifier:
```

```
    """Encapsulates an Amazon Comprehend custom classifier."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client
        self.classifier_arn = None


    def list(self):
        """
        Lists custom classifiers for the current account.

        :return: The list of classifiers.
        """
        try:
            response = self.comprehend_client.list_document_classifiers()
            classifiers = response["DocumentClassifierPropertiesList"]
            logger.info("Got %s classifiers.", len(classifiers))
        except ClientError:
            logger.exception(
                "Couldn't get classifiers.",
            )
            raise
        else:
            return classifiers
```

- For API details, see ListDocumentClassifiers in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `ListTopicsDetectionJobs` with an AWS SDK or CLI

The following code examples show how to use `ListTopicsDetectionJobs`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Run a topic modeling job on sample data

CLI

AWS CLI

To list all topic detection jobs

The following `list-topics-detection-jobs` example lists all in-progress and completed asynchronous topics detection jobs.

```
aws comprehend list-topics-detection-jobs
```

Output:

```
{
    "TopicsDetectionJobPropertiesList": [
        {
            "JobId": "123456abcdeb0e11022f22a11EXAMPLE",
            "JobArn": "arn:aws:comprehend:us-west-2:111122223333:topics-
detection-job/123456abcdeb0e11022f22a11EXAMPLE",
            "JobName" "topic-analysis-1"
            "JobStatus": "IN_PROGRESS",
            "SubmitTime": "2023-06-09T18:40:35.384000+00:00",
            "EndTime": "2023-06-09T18:46:41.936000+00:00",
            "InputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-BUCKET",
                "InputFormat": "ONE_DOC_PER_LINE"
            },
            "OutputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-DESTINATION-BUCKET/
thefolder/111122223333-TOPICS-123456abcdeb0e11022f22a11EXAMPLE/output/
output.tar.gz"
            },
            "NumberOfTopics": 10,
            "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-example-role"
        },
        {
            "JobId": "123456abcdeb0e11022f22a1EXAMPLE2",
            "JobArn": "arn:aws:comprehend:us-west-2:111122223333:topics-
detection-job/123456abcdeb0e11022f22a1EXAMPLE2",
```

```
            "JobName": "topic-analysis-2",
            "JobStatus": "COMPLETED",
            "SubmitTime": "2023-06-09T18:44:43.414000+00:00",
            "EndTime": "2023-06-09T18:50:50.872000+00:00",
            "InputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-BUCKET",
                "InputFormat": "ONE_DOC_PER_LINE"
            },
            "OutputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-DESTINATION-BUCKET/
thefolder/111122223333-TOPICS-123456abcdeb0e11022f22a1EXAMPLE2/output/
output.tar.gz"
            },
            "NumberOfTopics": 10,
            "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-example-role"
        },
        {
            "JobId": "123456abcdeb0e11022f22a1EXAMPLE3",
            "JobArn": "arn:aws:comprehend:us-west-2:111122223333:topics-
detection-job/123456abcdeb0e11022f22a1EXAMPLE3",
            "JobName": "topic-analysis-2",
            "JobStatus": "IN_PROGRESS",
            "SubmitTime": "2023-06-09T18:50:56.737000+00:00",
            "InputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-BUCKET",
                "InputFormat": "ONE_DOC_PER_LINE"
            },
            "OutputDataConfig": {
                "S3Uri": "s3://DOC-EXAMPLE-DESTINATION-BUCKET/
thefolder/111122223333-TOPICS-123456abcdeb0e11022f22a1EXAMPLE3/output/
output.tar.gz"
            },
            "NumberOfTopics": 10,
            "DataAccessRoleArn": "arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-example-role"
        }
    ]
}
```

For more information, see Async analysis for Amazon Comprehend insights in the *Amazon Comprehend Developer Guide*.

- For API details, see ListTopicsDetectionJobs in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendTopicModeler:
    """Encapsulates a Comprehend topic modeler."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def list_jobs(self):
        """
        Lists topic modeling jobs for the current account.

        :return: The list of jobs.
        """
        try:
            response = self.comprehend_client.list_topics_detection_jobs()
            jobs = response["TopicsDetectionJobPropertiesList"]
            logger.info("Got %s topic detection jobs.", len(jobs))
        except ClientError:
            logger.exception("Couldn't get topic detection jobs.")
            raise
        else:
            return jobs
```

- For API details, see ListTopicsDetectionJobs in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `StartDocumentClassificationJob` with an AWS SDK or CLI

The following code examples show how to use `StartDocumentClassificationJob`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Train a custom classifier and classify documents

CLI

AWS CLI

To start document classification job

The following `start-document-classification-job` example starts a document classification job with a custom model on all of the files at the address specified by the `--input-data-config` tag. In this example, the input S3 bucket contains `SampleSMStext1.txt`, `SampleSMStext2.txt`, and `SampleSMStext3.txt`. The model was previously trained on document classifications of spam and non-spam, or, "ham", SMS messages. When the job is complete, `output.tar.gz` is put at the location specified by the `--output-data-config` tag. `output.tar.gz` contains `predictions.jsonl` which lists the classification of each document. The Json output is printed on one line per file, but is formatted here for readability.

```
aws comprehend start-document-classification-job \
    --job-name exampleclassificationjob \
    --input-data-config "S3Uri=s3://DOC-EXAMPLE-BUCKET-INPUT/jobdata/" \
    --output-data-config "S3Uri=s3://DOC-EXAMPLE-DESTINATION-BUCKET/testfolder/"
 \
    --data-access-role-arn arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-example-role \
    --document-classifier-arn arn:aws:comprehend:us-west-2:111122223333:document-
classifier/mymodel/version/12
```

Contents of `SampleSMStext1.txt`:

```
"CONGRATULATIONS! TXT 2155550100 to win $5000"
```

Contents of `SampleSMStext2.txt`:

```
"Hi, when do you want me to pick you up from practice?"
```

Contents of `SampleSMStext3.txt`:

```
"Plz send bank account # to 2155550100 to claim prize!!"
```

Output:

```
{
    "JobId": "e758dd56b824aa717ceab551fEXAMPLE",
    "JobArn": "arn:aws:comprehend:us-west-2:111122223333:document-classification-
job/e758dd56b824aa717ceab551fEXAMPLE",
    "JobStatus": "SUBMITTED"
}
```

Contents of `predictions.jsonl`:

```
{"File": "SampleSMSText1.txt", "Line": "0", "Classes": [{"Name": "spam", "Score":
 0.9999}, {"Name": "ham", "Score": 0.0001}]}
{"File": "SampleSMStext2.txt", "Line": "0", "Classes": [{"Name": "ham", "Score":
 0.9994}, {"Name": "spam", "Score": 0.0006}]}
{"File": "SampleSMSText3.txt", "Line": "0", "Classes": [{"Name": "spam", "Score":
 0.9999}, {"Name": "ham", "Score": 0.0001}]}
```

For more information, see Custom Classification in the *Amazon Comprehend Developer Guide*.

- For API details, see StartDocumentClassificationJob in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendClassifier:
    """Encapsulates an Amazon Comprehend custom classifier."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client
        self.classifier_arn = None


    def start_job(
        self,
        job_name,
        input_bucket,
        input_key,
        input_format,
        output_bucket,
        output_key,
        data_access_role_arn,
    ):
        """
        Starts a classification job. The classifier must be trained or the job
        will fail. Input is read from the specified Amazon S3 input bucket and
        written to the specified output bucket. Output data is stored in a tar
        archive compressed in gzip format. The job runs asynchronously, so you
 can
        call `describe_document_classification_job` to get job status until it
        returns a status of SUCCEEDED.

        :param job_name: The name of the job.
        :param input_bucket: The Amazon S3 bucket that contains input data.
        :param input_key: The prefix used to find input data in the input
                          bucket. If multiple objects have the same prefix, all
                          of them are used.
        :param input_format: The format of the input data, either one document
 per
                             file or one document per line.
        :param output_bucket: The Amazon S3 bucket where output data is written.
        :param output_key: The prefix prepended to the output data.
        :param data_access_role_arn: The Amazon Resource Name (ARN) of a role
 that
```

```
                                    grants Comprehend permission to read from
the
                                    input bucket and write to the output bucket.
        :return: Information about the job, including the job ID.
        """
        try:
            response = self.comprehend_client.start_document_classification_job(
                DocumentClassifierArn=self.classifier_arn,
                JobName=job_name,
                InputDataConfig={
                    "S3Uri": f"s3://{input_bucket}/{input_key}",
                    "InputFormat": input_format.value,
                },
                OutputDataConfig={"S3Uri": f"s3://{output_bucket}/{output_key}"},
                DataAccessRoleArn=data_access_role_arn,
            )
            logger.info(
                "Document classification job %s is %s.", job_name,
response["JobStatus"]
            )
        except ClientError:
            logger.exception("Couldn't start classification job %s.", job_name)
            raise
        else:
            return response
```

- For API details, see StartDocumentClassificationJob in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Use `StartTopicsDetectionJob` with an AWS SDK or CLI

The following code examples show how to use `StartTopicsDetectionJob`.

Action examples are code excerpts from larger programs and must be run in context. You can see this action in context in the following code example:

- Run a topic modeling job on sample data

.NET

AWS SDK for .NET

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
    using System;
    using System.Threading.Tasks;
    using Amazon.Comprehend;
    using Amazon.Comprehend.Model;

    /// <summary>
    /// This example scans the documents in an Amazon Simple Storage Service
    /// (Amazon S3) bucket and analyzes it for topics. The results are stored
    /// in another bucket and then the resulting job properties are displayed
    /// on the screen. This example was created using the AWS SDK for .NEt
    /// version 3.7 and .NET Core version 5.0.
    /// </summary>
    public static class TopicModeling
    {
        /// <summary>
        /// This methos calls a topic detection job by calling the Amazon
        /// Comprehend StartTopicsDetectionJobRequest.
        /// </summary>
        public static async Task Main()
        {
            var comprehendClient = new AmazonComprehendClient();

            string inputS3Uri = "s3://input bucket/input path";
            InputFormat inputDocFormat = InputFormat.ONE_DOC_PER_FILE;
            string outputS3Uri = "s3://output bucket/output path";
            string dataAccessRoleArn = "arn:aws:iam::account ID:role/data access
role";
            int numberOfTopics = 10;
```

```
            var startTopicsDetectionJobRequest = new
StartTopicsDetectionJobRequest()
            {
                InputDataConfig = new InputDataConfig()
                {
                    S3Uri = inputS3Uri,
                    InputFormat = inputDocFormat,
                },
                OutputDataConfig = new OutputDataConfig()
                {
                    S3Uri = outputS3Uri,
                },
                DataAccessRoleArn = dataAccessRoleArn,
                NumberOfTopics = numberOfTopics,
            };

            var startTopicsDetectionJobResponse = await
comprehendClient.StartTopicsDetectionJobAsync(startTopicsDetectionJobRequest);

            var jobId = startTopicsDetectionJobResponse.JobId;
            Console.WriteLine("JobId: " + jobId);

            var describeTopicsDetectionJobRequest = new
DescribeTopicsDetectionJobRequest()
            {
                JobId = jobId,
            };

            var describeTopicsDetectionJobResponse = await
comprehendClient.DescribeTopicsDetectionJobAsync(describeTopicsDetectionJobRequest);

PrintJobProperties(describeTopicsDetectionJobResponse.TopicsDetectionJobProperties);

            var listTopicsDetectionJobsResponse = await
comprehendClient.ListTopicsDetectionJobsAsync(new
ListTopicsDetectionJobsRequest());
            foreach (var props in
listTopicsDetectionJobsResponse.TopicsDetectionJobPropertiesList)
            {
                PrintJobProperties(props);
            }
        }

        /// <summary>
```

```
        /// This method is a helper method that displays the job properties
        /// from the call to StartTopicsDetectionJobRequest.
        /// </summary>
        /// <param name="props">A list of properties from the call to
        /// StartTopicsDetectionJobRequest.</param>
        private static void PrintJobProperties(TopicsDetectionJobProperties
 props)
        {
            Console.WriteLine($"JobId: {props.JobId}, JobName: {props.JobName},
 JobStatus: {props.JobStatus}");
            Console.WriteLine($"NumberOfTopics:
 {props.NumberOfTopics}\nInputS3Uri: {props.InputDataConfig.S3Uri}");
            Console.WriteLine($"InputFormat: {props.InputDataConfig.InputFormat},
 OutputS3Uri: {props.OutputDataConfig.S3Uri}");
        }
    }
```

- For API details, see StartTopicsDetectionJob in *AWS SDK for .NET API Reference*.

CLI

AWS CLI

To start a topics detection analysis job

The following `start-topics-detection-job` example starts an asynchronous topics detection job for all files located at the address specified by the `--input-data-config` tag. When the job is complete, the folder, `output`, is placed at the location specified by the `--ouput-data-config` tag. `output` contains topic-terms.csv and doc-topics.csv. The first output file, topic-terms.csv, is a list of topics in the collection. For each topic, the list includes, by default, the top terms by topic according to their weight. The second file, `doc-topics.csv`, lists the documents associated with a topic and the proportion of the document that is concerned with the topic.

```
aws comprehend start-topics-detection-job \
    --job-name example_topics_detection_job \
    --language-code en \
    --input-data-config "S3Uri=s3://DOC-EXAMPLE-BUCKET/" \
    --output-data-config "S3Uri=s3://DOC-EXAMPLE-DESTINATION-BUCKET/testfolder/"
 \
```

```
    --data-access-role-arn arn:aws:iam::111122223333:role/service-role/
AmazonComprehendServiceRole-example-role \
    --language-code en
```

Output:

```
{
    "JobId": "123456abcdeb0e11022f22a11EXAMPLE",
    "JobArn": "arn:aws:comprehend:us-west-2:111122223333:key-phrases-detection-
job/123456abcdeb0e11022f22a11EXAMPLE",
    "JobStatus": "SUBMITTED"
}
```

For more information, see Topic Modeling in the *Amazon Comprehend Developer Guide*.

- For API details, see StartTopicsDetectionJob in *AWS CLI Command Reference*.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

```
class ComprehendTopicModeler:
    """Encapsulates a Comprehend topic modeler."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def start_job(
        self,
        job_name,
        input_bucket,
        input_key,
```

```
        input_format,
        output_bucket,
        output_key,
        data_access_role_arn,
    ):
        """
        Starts a topic modeling job. Input is read from the specified Amazon S3
        input bucket and written to the specified output bucket. Output data is
stored
        in a tar archive compressed in gzip format. The job runs asynchronously,
so you
        can call `describe_topics_detection_job` to get job status until it
        returns a status of SUCCEEDED.

        :param job_name: The name of the job.
        :param input_bucket: An Amazon S3 bucket that contains job input.
        :param input_key: The prefix used to find input data in the input
                          bucket. If multiple objects have the same prefix,
all
                          of them are used.
        :param input_format: The format of the input data, either one document
per
                          file or one document per line.
        :param output_bucket: The Amazon S3 bucket where output data is written.
        :param output_key: The prefix prepended to the output data.
        :param data_access_role_arn: The Amazon Resource Name (ARN) of a role
that
                                     grants Comprehend permission to read from
the
                                     input bucket and write to the output bucket.
        :return: Information about the job, including the job ID.
        """
        try:
            response = self.comprehend_client.start_topics_detection_job(
                JobName=job_name,
                DataAccessRoleArn=data_access_role_arn,
                InputDataConfig={
                    "S3Uri": f"s3://{input_bucket}/{input_key}",
                    "InputFormat": input_format.value,
                },
                OutputDataConfig={"S3Uri": f"s3://{output_bucket}/{output_key}"},
            )
            logger.info("Started topic modeling job %s.", response["JobId"])
        except ClientError:
```

```
            logger.exception("Couldn't start topic modeling job.")
            raise
        else:
            return response
```

- For API details, see StartTopicsDetectionJob in *AWS SDK for Python (Boto3) API Reference*.

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Scenarios for Amazon Comprehend using AWS SDKs

The following code examples show you how to implement common scenarios in Amazon Comprehend with AWS SDKs. These scenarios show you how to accomplish specific tasks by calling multiple functions within Amazon Comprehend. Each scenario includes a link to GitHub, where you can find instructions on how to set up and run the code.

Examples

- Detect document elements with Amazon Comprehend and an AWS SDK
- Run an Amazon Comprehend topic modeling job on sample data using an AWS SDK
- Train a custom Amazon Comprehend classifier and classify documents using an AWS SDK

Detect document elements with Amazon Comprehend and an AWS SDK

The following code example shows how to:

- Detect languages, entities, and key phrases in a document.
- Detect personally identifiable information (PII) in a document.
- Detect the sentiment of a document.
- Detect syntax elements in a document.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create a class that wraps Amazon Comprehend actions.

```
import logging
from pprint import pprint
import boto3
from botocore.exceptions import ClientError

logger = logging.getLogger(__name__)

class ComprehendDetect:
    """Encapsulates Comprehend detection functions."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def detect_languages(self, text):
        """
        Detects languages used in a document.

        :param text: The document to inspect.
        :return: The list of languages along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_dominant_language(Text=text)
            languages = response["Languages"]
            logger.info("Detected %s languages.", len(languages))
        except ClientError:
            logger.exception("Couldn't detect languages.")
            raise
```

```
        else:
            return languages


    def detect_entities(self, text, language_code):
        """
        Detects entities in a document. Entities can be things like people and
places
        or other common terms.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of entities along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_entities(
                Text=text, LanguageCode=language_code
            )
            entities = response["Entities"]
            logger.info("Detected %s entities.", len(entities))
        except ClientError:
            logger.exception("Couldn't detect entities.")
            raise
        else:
            return entities


    def detect_key_phrases(self, text, language_code):
        """
        Detects key phrases in a document. A key phrase is typically a noun and
its
        modifiers.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of key phrases along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_key_phrases(
                Text=text, LanguageCode=language_code
            )
            phrases = response["KeyPhrases"]
            logger.info("Detected %s phrases.", len(phrases))
        except ClientError:
```

```
            logger.exception("Couldn't detect phrases.")
            raise
        else:
            return phrases


    def detect_pii(self, text, language_code):
        """
        Detects personally identifiable information (PII) in a document. PII can
be
        things like names, account numbers, or addresses.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of PII entities along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_pii_entities(
                Text=text, LanguageCode=language_code
            )
            entities = response["Entities"]
            logger.info("Detected %s PII entities.", len(entities))
        except ClientError:
            logger.exception("Couldn't detect PII entities.")
            raise
        else:
            return entities


    def detect_sentiment(self, text, language_code):
        """
        Detects the overall sentiment expressed in a document. Sentiment can
        be positive, negative, neutral, or a mixture.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The sentiments along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_sentiment(
                Text=text, LanguageCode=language_code
            )
            logger.info("Detected primary sentiment %s.", response["Sentiment"])
        except ClientError:
```

```
            logger.exception("Couldn't detect sentiment.")
            raise
        else:
            return response


    def detect_syntax(self, text, language_code):
        """
        Detects syntactical elements of a document. Syntax tokens are portions of
        text along with their use as parts of speech, such as nouns, verbs, and
        interjections.

        :param text: The document to inspect.
        :param language_code: The language of the document.
        :return: The list of syntax tokens along with their confidence scores.
        """
        try:
            response = self.comprehend_client.detect_syntax(
                Text=text, LanguageCode=language_code
            )
            tokens = response["SyntaxTokens"]
            logger.info("Detected %s syntax tokens.", len(tokens))
        except ClientError:
            logger.exception("Couldn't detect syntax.")
            raise
        else:
            return tokens
```

Call functions on the wrapper class to detect entities, phrases, and more in a document.

```
def usage_demo():
    print("-" * 88)
    print("Welcome to the Amazon Comprehend detection demo!")
    print("-" * 88)

    logging.basicConfig(level=logging.INFO, format="%(levelname)s: %(message)s")

    comp_detect = ComprehendDetect(boto3.client("comprehend"))
    with open("detect_sample.txt") as sample_file:
        sample_text = sample_file.read()
```

```
    demo_size = 3

    print("Sample text used for this demo:")
    print("-" * 88)
    print(sample_text)
    print("-" * 88)

    print("Detecting languages.")
    languages = comp_detect.detect_languages(sample_text)
    pprint(languages)
    lang_code = languages[0]["LanguageCode"]

    print("Detecting entities.")
    entities = comp_detect.detect_entities(sample_text, lang_code)
    print(f"The first {demo_size} are:")
    pprint(entities[:demo_size])

    print("Detecting key phrases.")
    phrases = comp_detect.detect_key_phrases(sample_text, lang_code)
    print(f"The first {demo_size} are:")
    pprint(phrases[:demo_size])

    print("Detecting personally identifiable information (PII).")
    pii_entities = comp_detect.detect_pii(sample_text, lang_code)
    print(f"The first {demo_size} are:")
    pprint(pii_entities[:demo_size])

    print("Detecting sentiment.")
    sentiment = comp_detect.detect_sentiment(sample_text, lang_code)
    print(f"Sentiment: {sentiment['Sentiment']}")
    print("SentimentScore:")
    pprint(sentiment["SentimentScore"])

    print("Detecting syntax elements.")
    syntax_tokens = comp_detect.detect_syntax(sample_text, lang_code)
    print(f"The first {demo_size} are:")
    pprint(syntax_tokens[:demo_size])

    print("Thanks for watching!")
    print("-" * 88)
```

- For API details, see the following topics in *AWS SDK for Python (Boto3) API Reference*.
  - DetectDominantLanguage
  - DetectEntities
  - DetectKeyPhrases
  - DetectPiiEntities
  - DetectSentiment
  - DetectSyntax

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Run an Amazon Comprehend topic modeling job on sample data using an AWS SDK

The following code example shows how to:

- Run an Amazon Comprehend topic modeling job on sample data.
- Get information about the job.
- Extract job output data from Amazon S3.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create a wrapper class to call Amazon Comprehend topic modeling actions.

```
class ComprehendTopicModeler:
```

```
    """Encapsulates a Comprehend topic modeler."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client


    def start_job(
        self,
        job_name,
        input_bucket,
        input_key,
        input_format,
        output_bucket,
        output_key,
        data_access_role_arn,
    ):
        """
        Starts a topic modeling job. Input is read from the specified Amazon S3
        input bucket and written to the specified output bucket. Output data is
stored
        in a tar archive compressed in gzip format. The job runs asynchronously,
so you
        can call `describe_topics_detection_job` to get job status until it
        returns a status of SUCCEEDED.

        :param job_name: The name of the job.
        :param input_bucket: An Amazon S3 bucket that contains job input.
        :param input_key: The prefix used to find input data in the input
                          bucket. If multiple objects have the same prefix,
all
                          of them are used.
        :param input_format: The format of the input data, either one document
per
                          file or one document per line.
        :param output_bucket: The Amazon S3 bucket where output data is written.
        :param output_key: The prefix prepended to the output data.
        :param data_access_role_arn: The Amazon Resource Name (ARN) of a role
that
                                     grants Comprehend permission to read from
the
                                     input bucket and write to the output bucket.
```

```
        :return: Information about the job, including the job ID.
        """
        try:
            response = self.comprehend_client.start_topics_detection_job(
                JobName=job_name,
                DataAccessRoleArn=data_access_role_arn,
                InputDataConfig={
                    "S3Uri": f"s3://{input_bucket}/{input_key}",
                    "InputFormat": input_format.value,
                },
                OutputDataConfig={"S3Uri": f"s3://{output_bucket}/{output_key}"},
            )
            logger.info("Started topic modeling job %s.", response["JobId"])
        except ClientError:
            logger.exception("Couldn't start topic modeling job.")
            raise
        else:
            return response


    def describe_job(self, job_id):
        """
        Gets metadata about a topic modeling job.

        :param job_id: The ID of the job to look up.
        :return: Metadata about the job.
        """
        try:
            response = self.comprehend_client.describe_topics_detection_job(
                JobId=job_id
            )
            job = response["TopicsDetectionJobProperties"]
            logger.info("Got topic detection job %s.", job_id)
        except ClientError:
            logger.exception("Couldn't get topic detection job %s.", job_id)
            raise
        else:
            return job


    def list_jobs(self):
        """
        Lists topic modeling jobs for the current account.
```

```
        :return: The list of jobs.
        """
        try:
            response = self.comprehend_client.list_topics_detection_jobs()
            jobs = response["TopicsDetectionJobPropertiesList"]
            logger.info("Got %s topic detection jobs.", len(jobs))
        except ClientError:
            logger.exception("Couldn't get topic detection jobs.")
            raise
        else:
            return jobs
```

Use the wrapper class to run a topic modeling job and get job data.

```
def usage_demo():
    print("-" * 88)
    print("Welcome to the Amazon Comprehend topic modeling demo!")
    print("-" * 88)

    logging.basicConfig(level=logging.INFO, format="%(levelname)s: %(message)s")

    input_prefix = "input/"
    output_prefix = "output/"
    demo_resources = ComprehendDemoResources(
        boto3.resource("s3"), boto3.resource("iam")
    )
    topic_modeler = ComprehendTopicModeler(boto3.client("comprehend"))

    print("Setting up storage and security resources needed for the demo.")
    demo_resources.setup("comprehend-topic-modeler-demo")
    print("Copying sample data from public bucket into input bucket.")
    demo_resources.bucket.copy(
        {"Bucket": "public-sample-us-west-2", "Key": "TopicModeling/Sample.txt"},
        f"{input_prefix}sample.txt",
    )

    print("Starting topic modeling job on sample data.")
    job_info = topic_modeler.start_job(
        "demo-topic-modeling-job",
        demo_resources.bucket.name,
```

```
        input_prefix,
        JobInputFormat.per_line,
        demo_resources.bucket.name,
        output_prefix,
        demo_resources.data_access_role.arn,
    )

    print(
        f"Waiting for job {job_info['JobId']} to complete. This typically takes "
        f"20 - 30 minutes."
    )
    job_waiter = JobCompleteWaiter(topic_modeler.comprehend_client)
    job_waiter.wait(job_info["JobId"])

    job = topic_modeler.describe_job(job_info["JobId"])
    print(f"Job {job['JobId']} complete:")
    pprint(job)

    print(
        f"Getting job output data from the output Amazon S3 bucket: "
        f"{job['OutputDataConfig']['S3Uri']}."
    )
    job_output = demo_resources.extract_job_output(job)
    lines = 10
    print(f"First {lines} lines of document topics output:")
    pprint(job_output["doc-topics.csv"]["data"][:lines])
    print(f"First {lines} lines of terms output:")
    pprint(job_output["topic-terms.csv"]["data"][:lines])

    print("Cleaning up resources created for the demo.")
    demo_resources.cleanup()

    print("Thanks for watching!")
    print("-" * 88)
```

- For API details, see the following topics in *AWS SDK for Python (Boto3) API Reference*.
  - DescribeTopicsDetectionJob
  - ListTopicsDetectionJobs
  - StartTopicsDetectionJob

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Train a custom Amazon Comprehend classifier and classify documents using an AWS SDK

The following code example shows how to:

- Create an Amazon Comprehend multi-label classifier.
- Train the classifier on sample data.
- Run a classification job on a second set of data.
- Extract the job output data from Amazon S3.

Python

SDK for Python (Boto3)

> **Note**
>
> There's more on GitHub. Find the complete example and learn how to set up and run in the AWS Code Examples Repository.

Create a wrapper class to call Amazon Comprehend document classifier actions.

```
class ComprehendClassifier:
    """Encapsulates an Amazon Comprehend custom classifier."""

    def __init__(self, comprehend_client):
        """
        :param comprehend_client: A Boto3 Comprehend client.
        """
        self.comprehend_client = comprehend_client
        self.classifier_arn = None


    def create(
        self,
```

```
        name,
        language_code,
        training_bucket,
        training_key,
        data_access_role_arn,
        mode,
    ):
        """
        Creates a custom classifier. After the classifier is created, it
immediately
        starts training on the data found in the specified Amazon S3 bucket.
Training
        can take 30 minutes or longer. The `describe_document_classifier`
function
        can be used to get training status and returns a status of TRAINED when
the
        classifier is ready to use.

        :param name: The name of the classifier.
        :param language_code: The language the classifier can operate on.
        :param training_bucket: The Amazon S3 bucket that contains the training
data.
        :param training_key: The prefix used to find training data in the
training
                             bucket. If multiple objects have the same prefix,
all
                             of them are used.
        :param data_access_role_arn: The Amazon Resource Name (ARN) of a role
that
                                     grants Comprehend permission to read from
the
                                     training bucket.
        :return: The ARN of the newly created classifier.
        """
        try:
            response = self.comprehend_client.create_document_classifier(
                DocumentClassifierName=name,
                LanguageCode=language_code,
                InputDataConfig={"S3Uri": f"s3://{training_bucket}/
{training_key}"},
                DataAccessRoleArn=data_access_role_arn,
                Mode=mode.value,
            )
            self.classifier_arn = response["DocumentClassifierArn"]
```

```
            logger.info("Started classifier creation. Arn is: %s.",
self.classifier_arn)
        except ClientError:
            logger.exception("Couldn't create classifier %s.", name)
            raise
        else:
            return self.classifier_arn


    def describe(self, classifier_arn=None):
        """
        Gets metadata about a custom classifier, including its current status.

        :param classifier_arn: The ARN of the classifier to look up.
        :return: Metadata about the classifier.
        """
        if classifier_arn is not None:
            self.classifier_arn = classifier_arn
        try:
            response = self.comprehend_client.describe_document_classifier(
                DocumentClassifierArn=self.classifier_arn
            )
            classifier = response["DocumentClassifierProperties"]
            logger.info("Got classifier %s.", self.classifier_arn)
        except ClientError:
            logger.exception("Couldn't get classifier %s.", self.classifier_arn)
            raise
        else:
            return classifier


    def list(self):
        """
        Lists custom classifiers for the current account.

        :return: The list of classifiers.
        """
        try:
            response = self.comprehend_client.list_document_classifiers()
            classifiers = response["DocumentClassifierPropertiesList"]
            logger.info("Got %s classifiers.", len(classifiers))
        except ClientError:
            logger.exception(
                "Couldn't get classifiers.",
```

```
            )
            raise
        else:
            return classifiers


    def delete(self):
        """
        Deletes the classifier.
        """
        try:
            self.comprehend_client.delete_document_classifier(
                DocumentClassifierArn=self.classifier_arn
            )
            logger.info("Deleted classifier %s.", self.classifier_arn)
            self.classifier_arn = None
        except ClientError:
            logger.exception("Couldn't deleted classifier %s.",
self.classifier_arn)
            raise


    def start_job(
        self,
        job_name,
        input_bucket,
        input_key,
        input_format,
        output_bucket,
        output_key,
        data_access_role_arn,
    ):
        """
        Starts a classification job. The classifier must be trained or the job
        will fail. Input is read from the specified Amazon S3 input bucket and
        written to the specified output bucket. Output data is stored in a tar
        archive compressed in gzip format. The job runs asynchronously, so you
can
        call `describe_document_classification_job` to get job status until it
        returns a status of SUCCEEDED.

        :param job_name: The name of the job.
        :param input_bucket: The Amazon S3 bucket that contains input data.
        :param input_key: The prefix used to find input data in the input
```

```
                            bucket. If multiple objects have the same prefix, all
                            of them are used.
        :param input_format: The format of the input data, either one document
per
                               file or one document per line.
        :param output_bucket: The Amazon S3 bucket where output data is written.
        :param output_key: The prefix prepended to the output data.
        :param data_access_role_arn: The Amazon Resource Name (ARN) of a role
that
                                     grants Comprehend permission to read from
the
                                     input bucket and write to the output bucket.
        :return: Information about the job, including the job ID.
        """
        try:
            response = self.comprehend_client.start_document_classification_job(
                DocumentClassifierArn=self.classifier_arn,
                JobName=job_name,
                InputDataConfig={
                    "S3Uri": f"s3://{input_bucket}/{input_key}",
                    "InputFormat": input_format.value,
                },
                OutputDataConfig={"S3Uri": f"s3://{output_bucket}/{output_key}"},
                DataAccessRoleArn=data_access_role_arn,
            )
            logger.info(
                "Document classification job %s is %s.", job_name,
response["JobStatus"]
            )
        except ClientError:
            logger.exception("Couldn't start classification job %s.", job_name)
            raise
        else:
            return response


    def describe_job(self, job_id):
        """
        Gets metadata about a classification job.

        :param job_id: The ID of the job to look up.
        :return: Metadata about the job.
        """
        try:
```

```
            response =
 self.comprehend_client.describe_document_classification_job(
                JobId=job_id
            )
            job = response["DocumentClassificationJobProperties"]
            logger.info("Got classification job %s.", job["JobName"])
        except ClientError:
            logger.exception("Couldn't get classification job %s.", job_id)
            raise
        else:
            return job


    def list_jobs(self):
        """
        Lists the classification jobs for the current account.

        :return: The list of jobs.
        """
        try:
            response = self.comprehend_client.list_document_classification_jobs()
            jobs = response["DocumentClassificationJobPropertiesList"]
            logger.info("Got %s document classification jobs.", len(jobs))
        except ClientError:
            logger.exception(
                "Couldn't get document classification jobs.",
            )
            raise
        else:
            return jobs
```

Create a class to help run the scenario.

```
class ClassifierDemo:
    """
    Encapsulates functions used to run the demonstration.
    """

    def __init__(self, demo_resources):
        """
```

```
        :param demo_resources: A ComprehendDemoResources class that manages
 resources
                               for the demonstration.
        """
        self.demo_resources = demo_resources
        self.training_prefix = "training/"
        self.input_prefix = "input/"
        self.input_format = JobInputFormat.per_line
        self.output_prefix = "output/"

    def setup(self):
        """Creates AWS resources used by the demo."""
        self.demo_resources.setup("comprehend-classifier-demo")

    def cleanup(self):
        """Deletes AWS resources used by the demo."""
        self.demo_resources.cleanup()

    @staticmethod
    def _sanitize_text(text):
        """Removes characters that cause errors for the document parser."""
        return text.replace("\r", " ").replace("\n", " ").replace(",", ";")

    @staticmethod
    def _get_issues(query, issue_count):
        """
        Gets issues from GitHub using the specified query parameters.

        :param query: The query string used to request issues from the GitHub
 API.
        :param issue_count: The number of issues to retrieve.
        :return: The list of issues retrieved from GitHub.
        """
        issues = []
        logger.info("Requesting issues from %s?%s.", GITHUB_SEARCH_URL, query)
        response = requests.get(f"{GITHUB_SEARCH_URL}?
{query}&per_page={issue_count}")
        if response.status_code == 200:
            issue_page = response.json()["items"]
            logger.info("Got %s issues.", len(issue_page))
            issues = [
                {
                    "title": ClassifierDemo._sanitize_text(issue["title"]),
                    "body": ClassifierDemo._sanitize_text(issue["body"]),
```

```
                    "labels": {label["name"] for label in issue["labels"]},
                }
                for issue in issue_page
            ]
        else:
            logger.error(
                "GitHub returned error code %s with message %s.",
                response.status_code,
                response.json(),
            )
        logger.info("Found %s issues.", len(issues))
        return issues

    def get_training_issues(self, training_labels):
        """
        Gets issues used for training the custom classifier. Training issues are
        closed issues from the Boto3 repo that have known labels. Comprehend
        requires a minimum of ten training issues per label.

        :param training_labels: The issue labels to use for training.
        :return: The set of issues used for training.
        """
        issues = []
        per_label_count = 15
        for label in training_labels:
            issues += self._get_issues(
                f"q=type:issue+repo:boto/boto3+state:closed+label:{label}",
                per_label_count,
            )
            for issue in issues:
                issue["labels"] = issue["labels"].intersection(training_labels)
        return issues

    def get_input_issues(self, training_labels):
        """
        Gets input issues from GitHub. For demonstration purposes, input issues
        are open issues from the Boto3 repo with known labels, though in practice
        any issue could be submitted to the classifier for labeling.

        :param training_labels: The set of labels to query for.
        :return: The set of issues used for input.
        """
        issues = []
        per_label_count = 5
```

```
        for label in training_labels:
            issues += self._get_issues(
                f"q=type:issue+repo:boto/boto3+state:open+label:{label}",
                per_label_count,
            )
        return issues

    def upload_issue_data(self, issues, training=False):
        """
        Uploads issue data to an Amazon S3 bucket, either for training or for
input.
        The data is first put into the format expected by Comprehend. For
training,
        the set of pipe-delimited labels is prepended to each document. For
        input, labels are not sent.

        :param issues: The set of issues to upload to Amazon S3.
        :param training: Indicates whether the issue data is used for training or
                         input.
        """
        try:
            obj_key = (
                self.training_prefix if training else self.input_prefix
            ) + "issues.txt"
            if training:
                issue_strings = [
                    f"{'|'.join(issue['labels'])},{issue['title']}
{issue['body']}"
                    for issue in issues
                ]
            else:
                issue_strings = [
                    f"{issue['title']} {issue['body']}" for issue in issues
                ]
            issue_bytes = BytesIO("\n".join(issue_strings).encode("utf-8"))
            self.demo_resources.bucket.upload_fileobj(issue_bytes, obj_key)
            logger.info(
                "Uploaded data as %s to bucket %s.",
                obj_key,
                self.demo_resources.bucket.name,
            )
        except ClientError:
            logger.exception(
```

```
                "Couldn't upload data to bucket %s.",
self.demo_resources.bucket.name
            )
            raise

    def extract_job_output(self, job):
        """Extracts job output from Amazon S3."""
        return self.demo_resources.extract_job_output(job)

    @staticmethod
    def reconcile_job_output(input_issues, output_dict):
        """
        Reconciles job output with the list of input issues. Because the input
issues
        have known labels, these can be compared with the labels added by the
        classifier to judge the accuracy of the output.

        :param input_issues: The list of issues used as input.
        :param output_dict: The dictionary of data that is output by the
classifier.
        :return: The list of reconciled input and output data.
        """
        reconciled = []
        for archive in output_dict.values():
            for line in archive["data"]:
                in_line = int(line["Line"])
                in_labels = input_issues[in_line]["labels"]
                out_labels = {
                    label["Name"]
                    for label in line["Labels"]
                    if float(label["Score"]) > 0.3
                }
                reconciled.append(
                    f"{line['File']}, line {in_line} has labels {in_labels}.\n"
                    f"\tClassifier assigned {out_labels}."
                )
        logger.info("Reconciled input and output labels.")
        return reconciled
```

Train a classifier on a set of GitHub issues with known labels, then send a second set of GitHub issues to the classifier so that they can be labeled.

```
def usage_demo():
    print("-" * 88)
    print("Welcome to the Amazon Comprehend custom document classifier demo!")
    print("-" * 88)

    logging.basicConfig(level=logging.INFO, format="%(levelname)s: %(message)s")

    comp_demo = ClassifierDemo(
        ComprehendDemoResources(boto3.resource("s3"), boto3.resource("iam"))
    )
    comp_classifier = ComprehendClassifier(boto3.client("comprehend"))
    classifier_trained_waiter = ClassifierTrainedWaiter(
        comp_classifier.comprehend_client
    )
    training_labels = {"bug", "feature-request", "dynamodb", "s3"}

    print("Setting up storage and security resources needed for the demo.")
    comp_demo.setup()

    print("Getting training data from GitHub and uploading it to Amazon S3.")
    training_issues = comp_demo.get_training_issues(training_labels)
    comp_demo.upload_issue_data(training_issues, True)

    classifier_name = "doc-example-classifier"
    print(f"Creating document classifier {classifier_name}.")
    comp_classifier.create(
        classifier_name,
        "en",
        comp_demo.demo_resources.bucket.name,
        comp_demo.training_prefix,
        comp_demo.demo_resources.data_access_role.arn,
        ClassifierMode.multi_label,
    )
    print(
        f"Waiting until {classifier_name} is trained. This typically takes "
        f"30-40 minutes."
    )
    classifier_trained_waiter.wait(comp_classifier.classifier_arn)

    print(f"Classifier {classifier_name} is trained:")
```

```
    pprint(comp_classifier.describe())

    print("Getting input data from GitHub and uploading it to Amazon S3.")
    input_issues = comp_demo.get_input_issues(training_labels)
    comp_demo.upload_issue_data(input_issues)

    print("Starting classification job on input data.")
    job_info = comp_classifier.start_job(
        "issue_classification_job",
        comp_demo.demo_resources.bucket.name,
        comp_demo.input_prefix,
        comp_demo.input_format,
        comp_demo.demo_resources.bucket.name,
        comp_demo.output_prefix,
        comp_demo.demo_resources.data_access_role.arn,
    )
    print(f"Waiting for job {job_info['JobId']} to complete.")
    job_waiter = JobCompleteWaiter(comp_classifier.comprehend_client)
    job_waiter.wait(job_info["JobId"])

    job = comp_classifier.describe_job(job_info["JobId"])
    print(f"Job {job['JobId']} complete:")
    pprint(job)

    print(
        f"Getting job output data from Amazon S3: "
        f"{job['OutputDataConfig']['S3Uri']}."
    )
    job_output = comp_demo.extract_job_output(job)
    print("Job output:")
    pprint(job_output)

    print("Reconciling job output with labels from GitHub:")
    reconciled_output = comp_demo.reconcile_job_output(input_issues, job_output)
    print(*reconciled_output, sep="\n")

    answer = input(f"Do you want to delete the classifier {classifier_name} (y/
n)? ")
    if answer.lower() == "y":
        print(f"Deleting {classifier_name}.")
        comp_classifier.delete()

    print("Cleaning up resources created for the demo.")
    comp_demo.cleanup()
```

```
    print("Thanks for watching!")
    print("-" * 88)
```

- For API details, see the following topics in *AWS SDK for Python (Boto3) API Reference*.
  - CreateDocumentClassifier
  - DeleteDocumentClassifier
  - DescribeDocumentClassificationJob
  - DescribeDocumentClassifier
  - ListDocumentClassificationJobs
  - ListDocumentClassifiers
  - StartDocumentClassificationJob

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Cross-service examples for Amazon Comprehend using AWS SDKs

The following sample applications use AWS SDKs to combine Amazon Comprehend with other AWS services. Each example includes a link to GitHub, where you can find instructions on how to set up and run the application.

Examples

- Build an Amazon Transcribe streaming app
- Create an Amazon Lex chatbot to engage your website visitors
- Create a web application that sends and retrieves messages by using Amazon SQS
- Create an application that analyzes customer feedback and synthesizes audio
- Detect entities in text extracted from an image using an AWS SDK

Build an Amazon Transcribe streaming app

The following code example shows how to build an app that records, transcribes, and translates live audio in real-time, and emails the results.

JavaScript

SDK for JavaScript (v3)

Shows how to use Amazon Transcribe to build an app that records, transcribes, and translates live audio in real-time, and emails the results using Amazon Simple Email Service (Amazon SES).

For complete source code and instructions on how to set up and run, see the full example on GitHub.

Services used in this example

- Amazon Comprehend
- Amazon SES
- Amazon Transcribe
- Amazon Translate

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Create an Amazon Lex chatbot to engage your website visitors

The following code examples show how to create a chatbot to engage your website visitors.

Java

SDK for Java 2.x

Shows how to use the Amazon Lex API to create a Chatbot within a web application to engage your web site visitors.

For complete source code and instructions on how to set up and run, see the full example on GitHub.

Services used in this example

- Amazon Comprehend
- Amazon Lex
- Amazon Translate

JavaScript

SDK for JavaScript (v3)

Shows how to use the Amazon Lex API to create a Chatbot within a web application to engage your web site visitors.

For complete source code and instructions on how to set up and run, see the full example Building an Amazon Lex chatbot in the AWS SDK for JavaScript developer guide.

Services used in this example

- Amazon Comprehend
- Amazon Lex
- Amazon Translate

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Create a web application that sends and retrieves messages by using Amazon SQS

The following code examples show how to create a messaging application by using Amazon SQS.

Java

SDK for Java 2.x

Shows how to use the Amazon SQS API to develop a Spring REST API that sends and retrieves messages.

For complete source code and instructions on how to set up and run, see the full example on GitHub.

Services used in this example

- Amazon Comprehend
- Amazon SQS

Kotlin

SDK for Kotlin

Shows how to use the Amazon SQS API to develop a Spring REST API that sends and retrieves messages.

For complete source code and instructions on how to set up and run, see the full example on GitHub.

Services used in this example

- Amazon Comprehend
- Amazon SQS

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Create an application that analyzes customer feedback and synthesizes audio

The following code examples show how to create an application that analyzes customer comment cards, translates them from their original language, determines their sentiment, and generates an audio file from the translated text.

.NET

AWS SDK for .NET

This example application analyzes and stores customer feedback cards. Specifically, it fulfills the need of a fictitious hotel in New York City. The hotel receives feedback from guests in various languages in the form of physical comment cards. That feedback is uploaded into the app through a web client. After an image of a comment card is uploaded, the following steps occur:

- Text is extracted from the image using Amazon Textract.
- Amazon Comprehend determines the sentiment of the extracted text and its language.
- The extracted text is translated to English using Amazon Translate.
- Amazon Polly synthesizes an audio file from the extracted text.

The full app can be deployed with the AWS CDK. For source code and deployment instructions, see the project in GitHub.

Services used in this example

- Amazon Comprehend
- Lambda
- Amazon Polly
- Amazon Textract
- Amazon Translate

Java

SDK for Java 2.x

This example application analyzes and stores customer feedback cards. Specifically, it fulfills the need of a fictitious hotel in New York City. The hotel receives feedback from guests in various languages in the form of physical comment cards. That feedback is uploaded into the app through a web client. After an image of a comment card is uploaded, the following steps occur:

- Text is extracted from the image using Amazon Textract.
- Amazon Comprehend determines the sentiment of the extracted text and its language.
- The extracted text is translated to English using Amazon Translate.
- Amazon Polly synthesizes an audio file from the extracted text.

The full app can be deployed with the AWS CDK. For source code and deployment instructions, see the project in GitHub.

Services used in this example

- Amazon Comprehend
- Lambda

- Amazon Polly
- Amazon Textract
- Amazon Translate

JavaScript

SDK for JavaScript (v3)

This example application analyzes and stores customer feedback cards. Specifically, it fulfills the need of a fictitious hotel in New York City. The hotel receives feedback from guests in various languages in the form of physical comment cards. That feedback is uploaded into the app through a web client. After an image of a comment card is uploaded, the following steps occur:

- Text is extracted from the image using Amazon Textract.
- Amazon Comprehend determines the sentiment of the extracted text and its language.
- The extracted text is translated to English using Amazon Translate.
- Amazon Polly synthesizes an audio file from the extracted text.

The full app can be deployed with the AWS CDK. For source code and deployment instructions, see the project in GitHub. The following excerpts show how the AWS SDK for JavaScript is used inside of Lambda functions.

```
import {
  ComprehendClient,
  DetectDominantLanguageCommand,
  DetectSentimentCommand,
} from "@aws-sdk/client-comprehend";

/**
 * Determine the language and sentiment of the extracted text.
 *
 * @param {{ source_text: string}} extractTextOutput
 */
export const handler = async (extractTextOutput) => {
  const comprehendClient = new ComprehendClient({});

  const detectDominantLanguageCommand = new DetectDominantLanguageCommand({
    Text: extractTextOutput.source_text,
  });
```

```
  // The source language is required for sentiment analysis and
  // translation in the next step.
  const { Languages } = await comprehendClient.send(
    detectDominantLanguageCommand,
  );

  const languageCode = Languages[0].LanguageCode;

  const detectSentimentCommand = new DetectSentimentCommand({
    Text: extractTextOutput.source_text,
    LanguageCode: languageCode,
  });

  const { Sentiment } = await comprehendClient.send(detectSentimentCommand);

  return {
    sentiment: Sentiment,
    language_code: languageCode,
  };
};
```

```
import {
  DetectDocumentTextCommand,
  TextractClient,
} from "@aws-sdk/client-textract";

/**
 * Fetch the S3 object from the event and analyze it using Amazon Textract.
 *
 * @param {import("@types/aws-lambda").EventBridgeEvent<"Object Created">}
 eventBridgeS3Event
 */
export const handler = async (eventBridgeS3Event) => {
  const textractClient = new TextractClient();

  const detectDocumentTextCommand = new DetectDocumentTextCommand({
    Document: {
      S3Object: {
        Bucket: eventBridgeS3Event.bucket,
        Name: eventBridgeS3Event.object,
      },
    },
```

```
  });

  // Textract returns a list of blocks. A block can be a line, a page, word, etc.
  // Each block also contains geometry of the detected text.
  // For more information on the Block type, see https://docs.aws.amazon.com/
textract/latest/dg/API_Block.html.
  const { Blocks } = await textractClient.send(detectDocumentTextCommand);

  // For the purpose of this example, we are only interested in words.
  const extractedWords = Blocks.filter((b) => b.BlockType === "WORD").map(
    (b) => b.Text,
  );

  return extractedWords.join(" ");
};
```

```
import { PollyClient, SynthesizeSpeechCommand } from "@aws-sdk/client-polly";
import { S3Client } from "@aws-sdk/client-s3";
import { Upload } from "@aws-sdk/lib-storage";

/**
 * Synthesize an audio file from text.
 *
 * @param {{ bucket: string, translated_text: string, object: string}}
 sourceDestinationConfig
 */
export const handler = async (sourceDestinationConfig) => {
  const pollyClient = new PollyClient({});

  const synthesizeSpeechCommand = new SynthesizeSpeechCommand({
    Engine: "neural",
    Text: sourceDestinationConfig.translated_text,
    VoiceId: "Ruth",
    OutputFormat: "mp3",
  });

  const { AudioStream } = await pollyClient.send(synthesizeSpeechCommand);

  const audioKey = `${sourceDestinationConfig.object}.mp3`;

  // Store the audio file in S3.
  const s3Client = new S3Client();
  const upload = new Upload({
```

```
    client: s3Client,
    params: {
      Bucket: sourceDestinationConfig.bucket,
      Key: audioKey,
      Body: AudioStream,
      ContentType: "audio/mp3",
    },
  });

  await upload.done();
  return audioKey;
};
```

```
import {
  TranslateClient,
  TranslateTextCommand,
} from "@aws-sdk/client-translate";

/**
 * Translate the extracted text to English.
 *
 * @param {{ extracted_text: string, source_language_code: string}}
 textAndSourceLanguage
 */
export const handler = async (textAndSourceLanguage) => {
  const translateClient = new TranslateClient({});

  const translateCommand = new TranslateTextCommand({
    SourceLanguageCode: textAndSourceLanguage.source_language_code,
    TargetLanguageCode: "en",
    Text: textAndSourceLanguage.extracted_text,
  });

  const { TranslatedText } = await translateClient.send(translateCommand);

  return { translated_text: TranslatedText };
};
```

Services used in this example

- Amazon Comprehend
- Lambda

- Amazon Polly
- Amazon Textract
- Amazon Translate

Ruby

SDK for Ruby

This example application analyzes and stores customer feedback cards. Specifically, it fulfills the need of a fictitious hotel in New York City. The hotel receives feedback from guests in various languages in the form of physical comment cards. That feedback is uploaded into the app through a web client. After an image of a comment card is uploaded, the following steps occur:

- Text is extracted from the image using Amazon Textract.
- Amazon Comprehend determines the sentiment of the extracted text and its language.
- The extracted text is translated to English using Amazon Translate.
- Amazon Polly synthesizes an audio file from the extracted text.

The full app can be deployed with the AWS CDK. For source code and deployment instructions, see the project in GitHub.

Services used in this example

- Amazon Comprehend
- Lambda
- Amazon Polly
- Amazon Textract
- Amazon Translate

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Detect entities in text extracted from an image using an AWS SDK

The following code example shows how to use Amazon Comprehend to detect entities in text extracted by Amazon Textract from an image that is stored in Amazon S3.

Python

SDK for Python (Boto3)

Shows how to use the AWS SDK for Python (Boto3) in a Jupyter notebook to detect entities in text that is extracted from an image. This example uses Amazon Textract to extract text from an image stored in Amazon Simple Storage Service (Amazon S3) and Amazon Comprehend to detect entities in the extracted text.

This example is a Jupyter notebook and must be run in an environment that can host notebooks. For instructions on how to run the example using Amazon SageMaker, see the directions in TextractAndComprehendNotebook.ipynb.

For complete source code and instructions on how to set up and run, see the full example on GitHub.

Services used in this example

- Amazon Comprehend
- Amazon S3
- Amazon Textract

For a complete list of AWS SDK developer guides and code examples, see Using Amazon Comprehend with an AWS SDK. This topic also includes information about getting started and details about previous SDK versions.

Security in Amazon Comprehend

Cloud security at AWS is the highest priority. As an AWS customer, you benefit from a data center and network architecture that is built to meet the requirements of the most security-sensitive organizations.

Security is a shared responsibility between AWS and you. The shared responsibility model describes this as security *of* the cloud and security *in* the cloud:

- **Security of the cloud** – AWS is responsible for protecting the infrastructure that runs AWS services in the AWS Cloud. AWS also provides you with services that you can use securely. Third-party auditors regularly test and verify the effectiveness of our security as part of the AWS Compliance Programs. To learn about the compliance programs that apply to Amazon Comprehend, see AWS Services in Scope by Compliance Program.
- **Security in the cloud** – Your responsibility is determined by the AWS service that you use. You are also responsible for other factors including the sensitivity of your data, your company's requirements, and applicable laws and regulations.

This documentation helps you understand how to apply the shared responsibility model when using Amazon Comprehend. The following topics show you how to configure Amazon Comprehend to meet your security and compliance objectives. You also learn how to use other AWS services that help you to monitor and secure your Amazon Comprehend resources.

Topics

Data protection in Amazon Comprehend

The AWS shared responsibility model applies to data protection in Amazon Comprehend. As described in this model, AWS is responsible for protecting the global infrastructure that runs all

of the AWS Cloud. You are responsible for maintaining control over your content that is hosted on this infrastructure. You are also responsible for the security configuration and management tasks for the AWS services that you use. For more information about data privacy, see the Data Privacy FAQ. For information about data protection in Europe, see the AWS Shared Responsibility Model and GDPR blog post on the *AWS Security Blog*.

For data protection purposes, we recommend that you protect AWS account credentials and set up individual users with AWS IAM Identity Center or AWS Identity and Access Management (IAM). That way, each user is given only the permissions necessary to fulfill their job duties. We also recommend that you secure your data in the following ways:

- Use multi-factor authentication (MFA) with each account.
- Use SSL/TLS to communicate with AWS resources. We require TLS 1.2 and recommend TLS 1.3.
- Set up API and user activity logging with AWS CloudTrail.
- Use AWS encryption solutions, along with all default security controls within AWS services.
- Use advanced managed security services such as Amazon Macie, which assists in discovering and securing sensitive data that is stored in Amazon S3.
- If you require FIPS 140-2 validated cryptographic modules when accessing AWS through a command line interface or an API, use a FIPS endpoint. For more information about the available FIPS endpoints, see Federal Information Processing Standard (FIPS) 140-2.

We strongly recommend that you never put confidential or sensitive information, such as your customers' email addresses, into tags or free-form text fields such as a **Name** field. This includes when you work with Amazon Comprehend or other AWS services using the console, API, AWS CLI, or AWS SDKs. Any data that you enter into tags or free-form text fields used for names may be used for billing or diagnostic logs. If you provide a URL to an external server, we strongly recommend that you do not include credentials information in the URL to validate your request to that server.

Topics

- KMS encryption in Amazon Comprehend
- Cross-service confused deputy prevention
- Protect jobs by using an Amazon Virtual Private Cloud
- Amazon Comprehend and interface VPC endpoints (AWS PrivateLink)

KMS encryption in Amazon Comprehend

Amazon Comprehend works with AWS Key Management Service (AWS KMS) to provide enhanced encryption for your data. Amazon S3 already enables you to encrypt your input documents when creating a text analysis, topic modeling, or custom Amazon Comprehend job. Integration with AWS KMS enables you to encrypt the data in the storage volume for Start* and Create* jobs, and it encrypts the output results of Start* jobs using your own KMS key.

For the AWS Management Console, Amazon Comprehend encrypts custom models with its own KMS key. For the AWS CLI, Amazon Comprehend can encrypt custom models using either its own KMS key or a provided customer managed key (CMK).

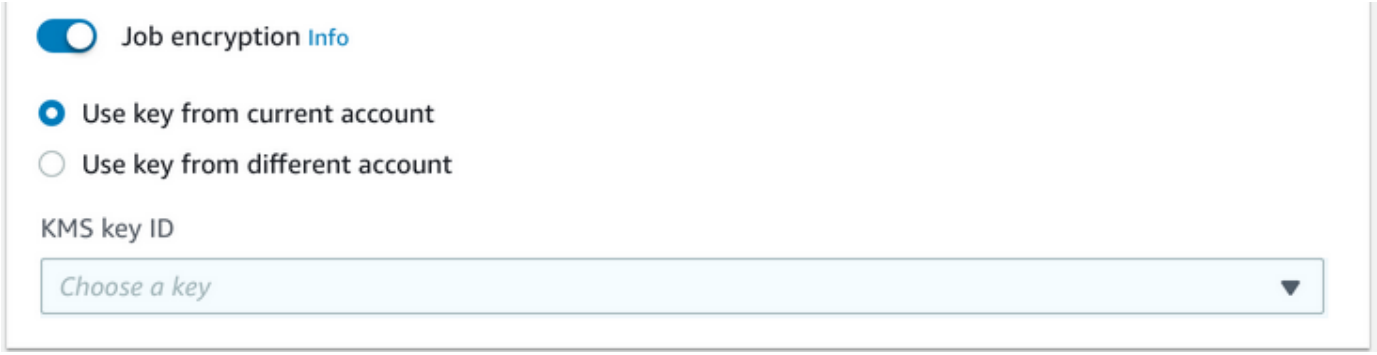
KMS encryption using the AWS Management Console

Two encryption options are available when using the console:

- Volume encryption
- Output result encryption

To enable volume encryption

1. Under **Job Settings**, choose the **Job encryption** option.



2. Choose whether the KMS customer-managed key (CMK) is from the account you're currently using or from a different account. If you want to use a key from the current account, choose the key alias from **KMS key ID**. If you're using a key from a different account, you must enter the key's ARN.

To enable output result encryption

1. Under **Output Settings**, choose the **Encryption** option.

2. Choose whether the customer-managed key (CMK) is from the account you're currently using or from a different account. If you want to use a key from the current account, choose the key ID from **KMS key ID**. If you're using a key from a different account, you must enter the key's ARN.

If you have previously setup encryption using SSE-KMS on the your S3 input documents, this can provide you with additional security. However, if you do this, the IAM role used must have `kms:Decrypt` permission for the KMS key with which the input documents are encrypted. For more information, see Permissions required to use KMS encryption.

KMS encryption with API operations

All Amazon Comprehend `Start*` and `Create*` API operations support KMS encrypted input documents. `Describe*` and `List*` API operations return the `KmsKeyId` in `OutputDataConfig` if the original job had `KmsKeyId` provided as an input. If it was not provided as input, it isn't returned.

This can be seen in the following AWS CLI example using the StartEntitiesDetectionJob operation:

```
aws comprehend start-entities-detection-job \
     --region region \
     --data-access-role-arn "data access role arn" \
     --entity-recognizer-arn "entity recognizer arn" \
     --input-data-config "S3Uri=s3://Bucket Name/Bucket Path" \
     --job-name job name \
     --language-code en \
     --output-data-config "KmsKeyId=Output S3 KMS key ID" "S3Uri=s3://Bucket
 Name/Bucket Path/" \
     --volumekmskeyid "Volume KMS key ID"
```

> **Note**
>
> This example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Customer Managed Key (CMK) encryption with API operations

Amazon Comprehend custom model API operations, `CreateEntityRecognizer`, `CreateDocumentClassifier`, and `CreateEndpoint`, support encryption using customer managed keys via the AWS CLI.

You need an IAM policy to allow a principal to use or manage customer managed keys. These keys are specified in the `Resource` element of the policy statement. As best practice, limit customer managed keys to only those that the principals must use in your policy statement.

The following AWS CLI example creates a custom entity recognizer with model encryption using the CreateEntityRecognizer operation:

```
aws comprehend create-entity-recognizer \
     --recognizer-name name \
     --data-access-role-arn data access role arn \
     --language-code en \
     --model-kms-key-id Model KMS Key ID \
     --input-data-config file:///path/input-data-config.json
```

> **Note**
>
> This example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

Cross-service confused deputy prevention

The confused deputy problem is a security issue where an entity that doesn't have permission to perform an action can coerce a more-privileged entity to perform the action. In AWS, cross-service impersonation can result in the confused deputy problem. Cross-service impersonation can occur

when one service (the *calling service*) calls another service (the *called service*). The calling service can be manipulated to use its permissions to act on another customer's resources in a way it should not otherwise have permission to access. To prevent this, AWS provides tools that help you protect your data for all services with service principals that have been given access to resources in your account.

We recommend using the `aws:SourceArn` and `aws:SourceAccount` global condition context keys in resource policies to limit the permissions that Amazon Comprehend gives another service to the resource. If you use both global condition context keys, the `aws:SourceAccount` value and the account in the `aws:SourceArn` value must use the same account ID when used in the same policy statement.

The most effective way to protect against the confused deputy problem is to use the `aws:SourceArn` global condition context key with the full ARN of the resource. If you don't know the full ARN of the resource or if you are specifying multiple resources, use the `aws:SourceArn` global context condition key with wildcards (*) for the unknown portions of the ARN. For example, `arn:aws:`*`servicename`*`::`*`123456789012`*`:*`.

Using source account

The following example shows how you can use the `aws:SourceAccount` global condition context key in Amazon Comprehend.

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Sid": "ConfusedDeputyPreventionExamplePolicy",
    "Effect": "Allow",
    "Principal": {
          "Service": "comprehend.amazonaws.com"
    },
    "Action": "sts:AssumeRole",
    "Condition": {
        "StringEquals": {
             "aws:SourceAccount":"111122223333"
        }
    }
  }
}
```

Trust policy for endpoints of encrypted models

You need to create a trust policy to create or update an endpoint for an encrypted model. Set the `aws:SourceAccount` value to your account ID. If you use the `ArnEquals` condition, set the `aws:SourceArn` value to the ARN of the endpoint.

```
{
 "Version": "2012-10-17",
 "Statement": [
    {
        "Sid": "",
        "Effect": "Allow",
        "Principal": {
            "Service": "comprehend.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
            "StringEquals": {
               "aws:SourceAccount": "111122223333"
            },
            "ArnEquals": {
               "aws:SourceArn": "arn:aws:comprehend:us-west-2:111122223333:document-
classifier-endpoint/endpoint-name"
            }
        }
    }
  ]
}
```

Create custom model

You need to create a trust policy to create a custom model. Set the `aws:SourceAccount` value to your account ID. If you use the `ArnEquals` condition, set the `aws:SourceArn` value to the ARN of the custom model version.

```
{
 "Version": "2012-10-17",
  "Statement": [
      {
          "Sid": "",
          "Effect": "Allow",
          "Principal": {
```

```
                "Service": "comprehend.amazonaws.com"
            },
            "Action": "sts:AssumeRole",
            "Condition": {
                "StringEquals": {
                    "aws:SourceAccount": "111122223333"
                },
                "ArnEquals": {
                    "aws:SourceArn": "arn:aws:comprehend:us-west-2:111122223333:
                                document-classifier/smallest-classifier-test/
version/version-name"
                }
            }
        }
    ]
}
```

Protect jobs by using an Amazon Virtual Private Cloud

Amazon Comprehend uses a variety of security measures to ensure the safety of your data with our job containers where it's stored while being used by Amazon Comprehend. However, job containers access AWS resources—such as the Amazon S3 buckets where you store data and model artifacts—over the internet.

To control access to your data, we recommend that you create a *virtual private cloud* (VPC) and configure it so that the data and containers aren't accessible over the internet. For information about creating and configuring a VPC, see Getting Started With Amazon VPC in the *Amazon VPC User Guide*. Using a VPC helps to protect your data because you can configure your VPC so that it is not connected to the internet. Using a VPC also allows you to monitor all network traffic in and out of our job containers by using VPC flow logs. For more information, see VPC Flow Logs in the *Amazon VPC User Guide*.

You specify your VPC configuration when you create a job, by specifying the subnets and security groups. When you specify the subnets and security groups, Amazon Comprehend creates *elastic network interfaces* (ENIs) that are associated with your security groups in one of the subnets. ENIs allow our job containers to connect to resources in your VPC. For information about ENIs, see Elastic Network Interfaces in the *Amazon VPC User Guide*.

> **Note**
>
> For jobs, you can only configure subnets with a default tenancy VPC in which your instance runs on shared hardware. For more information on the tenancy attribute for VPCs, see Dedicated Instances in the *Amazon EC2 User Guide*.

Configure a job for Amazon VPC access

To specify subnets and security groups in your VPC, use the `VpcConfig` request parameter of the applicable API, or provide this information when you create a job in the Amazon Comprehend console. Amazon Comprehend uses this information to create ENIs and attach them to our job containers. The ENIs provide our job containers with a network connection within your VPC that is not connected to the internet.

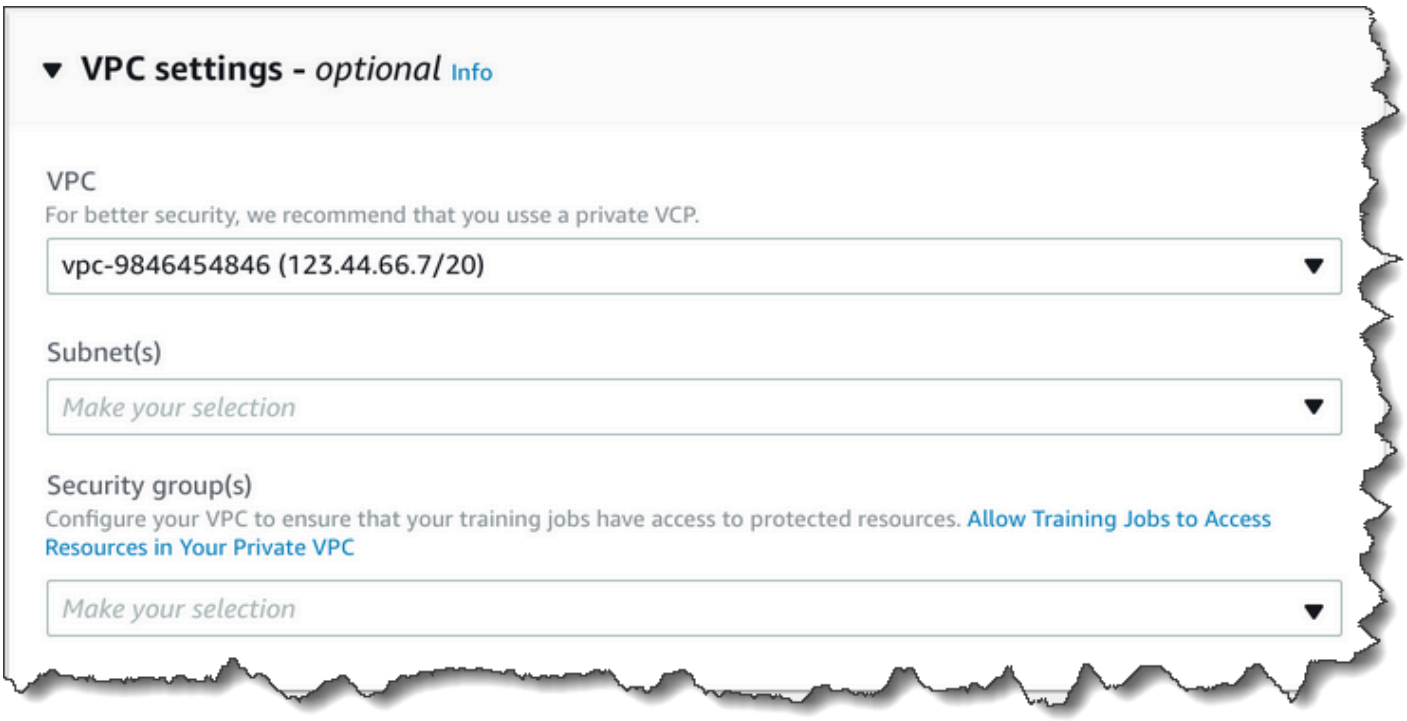
The following APIs contain the `VpcConfig` request parameter:

- `Create*` APIs: `CreateDocumentClassifier`, `CreateEntityRecognizer`
- `Start*` APIs: `StartDocumentClassificationJob`, `StartDominantLanguageDetectionJob`, `StartEntitiesDetectionJob`, `StartKeyPhrasesDetectionJob`, `StartSentimentDetectionJob`, `StartTargetedSentimentDetectionJob`, `StartTopicsDetectionJob`

The following is an example of the VpcConfig parameter that you include in your API call:

```
"VpcConfig": {
      "SecurityGroupIds": [
          " sg-0123456789abcdef0"
          ],
      "Subnets": [
          "subnet-0123456789abcdef0",
          "subnet-0123456789abcdef1",
          "subnet-0123456789abcdef2"
          ]
      }
```

To configure a VPC from the Amazon Comprehend console, choose the configuration details from the optional **VPC Settings** section when creating the job.



Configure your VPC for Amazon Comprehend jobs

When configuring the VPC for your Amazon Comprehend jobs, use the following guidelines. For information about setting up a VPC, see Working with VPCs and Subnets in the *Amazon VPC User Guide*.

Ensure That Subnets Have Enough IP Addresses

Your VPC subnets should have at least two private IP addresses for each instance in a job. For more information, see VPC and Subnet Sizing for IPv4 in the *Amazon VPC User Guide*.

Create an Amazon S3 VPC Endpoint

If you configure your VPC so that job containers don't have access to the internet, they can't connect to the Amazon S3 buckets that contain your data unless you create a VPC endpoint that allows access. By creating a VPC endpoint, you allow the job containers to access your data during training and analysis jobs.

When you create the VPC endpoint, configure these values:

- Select the service category as **AWS Services**
- Specify the service as `com.amazonaws.region.s3`
- Select **Gateway** as the VPC Endpoint type

If you're using AWS CloudFormation to create the VPC endpoint, follow the AWS CloudFormation VPCEndpoint documentation. The following example shows the **VPCEndpoint** configuration in a AWS CloudFormation template.

```
VpcEndpoint:
  Type: AWS::EC2::VPCEndpoint
  Properties:
    PolicyDocument:
      Version: '2012-10-17'
      Statement:
        - Action:
            - s3:GetObject
            - s3:PutObject
            - s3:ListBucket
            - s3:GetBucketLocation
            - s3:DeleteObject
            - s3:ListMultipartUploadParts
            - s3:AbortMultipartUpload
          Effect: Allow
          Resource:
            - "*"
          Principal: "*"
    RouteTableIds:
      - Ref: RouteTable
    ServiceName:
      Fn::Join:
        - ''
        - - com.amazonaws.
          - Ref: AWS::Region
          - ".s3"
    VpcId:
      Ref: VPC
```

We recommend that you also create a custom policy that allows only requests from your VPC to access to your S3 buckets. For more information, see Endpoints for Amazon S3 in the *Amazon VPC User Guide*.

The following policy allows access to S3 buckets. Edit this policy to allow access only the resources that your job needs.

```
{
    "Version": "2008-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": "*",
            "Action": [
                "s3:GetObject",
                "s3:PutObject",
                "s3:ListBucket",
                "s3:GetBucketLocation",
                "s3:DeleteObject",
                "s3:ListMultipartUploadParts",
                "s3:AbortMultipartUpload"
            ],
            "Resource": "*"
        }
    ]
}
```

Use default DNS settings for your endpoint route table, so that standard Amazon S3 URLs (for example, `http://s3-aws-region.amazonaws.com/DOC-EXAMPLE-BUCKET`) resolve. If you don't use default DNS settings, ensure that the URLs that you use to specify the locations of the data in your jobs resolve by configuring the endpoint route tables. For information about VPC endpoint route tables, see Routing for Gateway Endpoints in the *Amazon VPC User Guide*.

The default endpoint policy allows users to install packages from the Amazon Linux and Amazon Linux 2 repositories on our jobs container. If you don't want users to install packages from that repository, create a custom endpoint policy that explicitly denies access to the Amazon Linux and Amazon Linux 2 repositories. Comprehend itself doesn't need any such packages, so there won't be any functionality impact. The following is an example of a policy that denies access to these repositories:

```
{
    "Statement": [
      {
        "Sid": "AmazonLinuxAMIRepositoryAccess",
        "Principal": "*",
        "Action": [
            "s3:GetObject"
        ],
```

```
        "Effect": "Deny",
        "Resource": [
            "arn:aws:s3:::packages.*.amazonaws.com/*",
            "arn:aws:s3:::repo.*.amazonaws.com/*"
        ]
      }
    ]
}

{
    "Statement": [
        { "Sid": "AmazonLinux2AMIRepositoryAccess",
          "Principal": "*",
          "Action": [
              "s3:GetObject"
              ],
          "Effect": "Deny",
          "Resource": [
              "arn:aws:s3:::amazonlinux.*.amazonaws.com/*"
              ]
         }
    ]
}
```

Permissions for the `DataAccessRole`

When you use a VPC with your analysis job, the `DataAccessRole` used for the `Create*` and `Start*` operations must also have permissions to the VPC from which the input documents and the output bucket are accessed.

The following policy provides the access needed to the `DataAccessRole` used for the `Create*` and `Start*` operations.

```
{
    "Version": "2008-10-17",
    "Statement": [
        {
            "Effect": "Allow",
              "Action": [
                "ec2:CreateNetworkInterface",
                "ec2:CreateNetworkInterfacePermission",
                "ec2:DeleteNetworkInterface",
                "ec2:DeleteNetworkInterfacePermission",
```

```
                "ec2:DescribeNetworkInterfaces",
                "ec2:DescribeVpcs",
                "ec2:DescribeDhcpOptions",
                "ec2:DescribeSubnets",
                "ec2:DescribeSecurityGroups"
            ],
            "Resource": "*"
        }
    ]
}
```

Configure the VPC security group

With distributed jobs, you must allow communication between the different job containers in the same job. To do that, configure a rule for your security group that allows inbound connections between members of the same security group. For information, see Security Group Rules in the *Amazon VPC User Guide*.

Connect to resources outside your VPC

If you configure your VPC so that it doesn't have internet access, jobs that use that VPC do not have access to resources outside your VPC. If your jobs need access to resources outside your VPC, provide access with one of the following options:

- If your job needs access to an AWS service that supports interface VPC endpoints, create an endpoint to connect to that service. For a list of services that support interface endpoints, see VPC Endpoints in the *Amazon VPC User Guide*. For information about creating an interface VPC endpoint, see Interface VPC Endpoints (AWS PrivateLink) in the *Amazon VPC User Guide*.
- If your job needs access to an AWS service that doesn't support interface VPC endpoints or to a resource outside of AWS, create a NAT gateway and configure your security groups to allow outbound connections. For information about setting up a NAT gateway for your VPC, see Scenario 2: VPC with Public and Private Subnets (NAT) in the *Amazon VPC User Guide*.

Amazon Comprehend and interface VPC endpoints (AWS PrivateLink)

You can establish a private connection between your VPC and Amazon Comprehend by creating an *interface VPC endpoint*. Interface endpoints are powered by AWS PrivateLink, a technology that enables you to privately access Amazon Comprehend APIs without an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. Instances in your VPC don't need

public IP addresses to communicate with Amazon Comprehend APIs. Traffic between your VPC and Amazon Comprehend does not leave the Amazon network.

Each interface endpoint is represented by one or more Elastic network interfaces in your subnets.

For more information, see Interface VPC endpoints (AWS PrivateLink) in the *Amazon VPC User Guide*.

Considerations for Amazon Comprehend VPC endpoints

Before you set up an interface VPC endpoint for Amazon Comprehend, ensure that you review Interface endpoint properties and limitations in the *Amazon VPC User Guide*.

Amazon Comprehend endpoints are not available in all availability zones in a region. When you create the endpoint, use the following command to list the availability zones.

```
aws ec2 describe-vpc-endpoint-services \
  --service-names com.amazonaws.us-west-2.comprehend
```

Amazon Comprehend supports making calls to all of its API actions from your VPC.

Creating an interface VPC endpoint for Amazon Comprehend

You can create a VPC endpoint for the Amazon Comprehend service using either the Amazon VPC console or the AWS Command Line Interface (AWS CLI). For more information, see Creating an interface endpoint in the *Amazon VPC User Guide*.

Create a VPC endpoint for Amazon Comprehend using the following service name:

- com.amazonaws.*region*.comprehend

If you enable private DNS for the endpoint, you can make API requests to Amazon Comprehend using its default DNS name for the Region, for example, *comprehend.us-east-1.amazonaws.com*.

For more information, see Accessing a service through an interface endpoint in the *Amazon VPC User Guide*.

Creating a VPC endpoint policy for Amazon Comprehend

You can attach an endpoint policy to your VPC endpoint that controls access to Amazon Comprehend. The policy specifies the following information:

- The principal that can perform actions.
- The actions that can be performed.
- The resources on which actions can be performed.

For more information, see Controlling access to services with VPC endpoints in the *Amazon VPC User Guide*.

Example: VPC endpoint policy for Amazon Comprehend actions

The following is an example of an endpoint policy for Amazon Comprehend. When attached to an endpoint, this policy grants access to the Amazon Comprehend `DetectEntities` action for all principals on all resources.

```
{
   "Statement":[
      {
         "Principal":"*",
         "Effect":"Allow",
         "Action":[
            "comprehend:DetectEntities"
         ],
         "Resource":"*"
      }
   ]
}
```

Identity and Access Management for Amazon Comprehend

AWS Identity and Access Management (IAM) is an AWS service that helps an administrator securely control access to AWS resources. IAM administrators control who can be *authenticated* (signed in) and *authorized* (have permissions) to use Amazon Comprehend resources. IAM is an AWS service that you can use with no additional charge.

Topics

- Audience
- Authenticating with identities
- Managing access using policies

- How Amazon Comprehend works with IAM
- Identity-based policy examples for Amazon Comprehend
- AWS managed policies for Amazon Comprehend
- Troubleshooting Amazon Comprehend identity and access

Audience

How you use AWS Identity and Access Management (IAM) differs, depending on the work that you do in Amazon Comprehend.

Service user – If you use the Amazon Comprehend service to do your job, then your administrator provides you with the credentials and permissions that you need. As you use more Amazon Comprehend features to do your work, you might need additional permissions. Understanding how access is managed can help you request the right permissions from your administrator. If you cannot access a feature in Amazon Comprehend, see Troubleshooting Amazon Comprehend identity and access.

Service administrator – If you're in charge of Amazon Comprehend resources at your company, you probably have full access to Amazon Comprehend. It's your job to determine which Amazon Comprehend features and resources your service users should access. You must then submit requests to your IAM administrator to change the permissions of your service users. Review the information on this page to understand the basic concepts of IAM. To learn more about how your company can use IAM with Amazon Comprehend, see How Amazon Comprehend works with IAM.

IAM administrator – If you're an IAM administrator, you might want to learn details about how you can write policies to manage access to Amazon Comprehend. To view example Amazon Comprehend identity-based policies that you can use in IAM, see Identity-based policy examples for Amazon Comprehend.

Authenticating with identities

Authentication is how you sign in to AWS using your identity credentials. You must be *authenticated* (signed in to AWS) as the AWS account root user, as an IAM user, or by assuming an IAM role.

You can sign in to AWS as a federated identity by using credentials provided through an identity source. AWS IAM Identity Center (IAM Identity Center) users, your company's single sign-on

authentication, and your Google or Facebook credentials are examples of federated identities. When you sign in as a federated identity, your administrator previously set up identity federation using IAM roles. When you access AWS by using federation, you are indirectly assuming a role.

Depending on the type of user you are, you can sign in to the AWS Management Console or the AWS access portal. For more information about signing in to AWS, see How to sign in to your AWS account in the *AWS Sign-In User Guide*.

If you access AWS programmatically, AWS provides a software development kit (SDK) and a command line interface (CLI) to cryptographically sign your requests by using your credentials. If you don't use AWS tools, you must sign requests yourself. For more information about using the recommended method to sign requests yourself, see Signing AWS API requests in the *IAM User Guide*.

Regardless of the authentication method that you use, you might be required to provide additional security information. For example, AWS recommends that you use multi-factor authentication (MFA) to increase the security of your account. To learn more, see Multi-factor authentication in the *AWS IAM Identity Center User Guide* and Using multi-factor authentication (MFA) in AWS in the *IAM User Guide*.

AWS account root user

When you create an AWS account, you begin with one sign-in identity that has complete access to all AWS services and resources in the account. This identity is called the AWS account *root user* and is accessed by signing in with the email address and password that you used to create the account. We strongly recommend that you don't use the root user for your everyday tasks. Safeguard your root user credentials and use them to perform the tasks that only the root user can perform. For the complete list of tasks that require you to sign in as the root user, see Tasks that require root user credentials in the *IAM User Guide*.

Federated identity

As a best practice, require human users, including users that require administrator access, to use federation with an identity provider to access AWS services by using temporary credentials.

A *federated identity* is a user from your enterprise user directory, a web identity provider, the AWS Directory Service, the Identity Center directory, or any user that accesses AWS services by using credentials provided through an identity source. When federated identities access AWS accounts, they assume roles, and the roles provide temporary credentials.

For centralized access management, we recommend that you use AWS IAM Identity Center. You can create users and groups in IAM Identity Center, or you can connect and synchronize to a set of users and groups in your own identity source for use across all your AWS accounts and applications. For information about IAM Identity Center, see What is IAM Identity Center? in the *AWS IAM Identity Center User Guide*.

IAM users and groups

An *IAM user* is an identity within your AWS account that has specific permissions for a single person or application. Where possible, we recommend relying on temporary credentials instead of creating IAM users who have long-term credentials such as passwords and access keys. However, if you have specific use cases that require long-term credentials with IAM users, we recommend that you rotate access keys. For more information, see Rotate access keys regularly for use cases that require long-term credentials in the *IAM User Guide*.

An *IAM group* is an identity that specifies a collection of IAM users. You can't sign in as a group. You can use groups to specify permissions for multiple users at a time. Groups make permissions easier to manage for large sets of users. For example, you could have a group named *IAMAdmins* and give that group permissions to administer IAM resources.

Users are different from roles. A user is uniquely associated with one person or application, but a role is intended to be assumable by anyone who needs it. Users have permanent long-term credentials, but roles provide temporary credentials. To learn more, see When to create an IAM user (instead of a role) in the *IAM User Guide*.

IAM roles

An *IAM role* is an identity within your AWS account that has specific permissions. It is similar to an IAM user, but is not associated with a specific person. You can temporarily assume an IAM role in the AWS Management Console by switching roles. You can assume a role by calling an AWS CLI or AWS API operation or by using a custom URL. For more information about methods for using roles, see Using IAM roles in the *IAM User Guide*.

IAM roles with temporary credentials are useful in the following situations:

- **Federated user access** – To assign permissions to a federated identity, you create a role and define permissions for the role. When a federated identity authenticates, the identity is associated with the role and is granted the permissions that are defined by the role. For information about roles for federation, see Creating a role for a third-party Identity Provider

in the *IAM User Guide*. If you use IAM Identity Center, you configure a permission set. To control what your identities can access after they authenticate, IAM Identity Center correlates the permission set to a role in IAM. For information about permissions sets, see Permission sets in the *AWS IAM Identity Center User Guide*.

- **Temporary IAM user permissions** – An IAM user or role can assume an IAM role to temporarily take on different permissions for a specific task.
- **Cross-account access** – You can use an IAM role to allow someone (a trusted principal) in a different account to access resources in your account. Roles are the primary way to grant cross-account access. However, with some AWS services, you can attach a policy directly to a resource (instead of using a role as a proxy). To learn the difference between roles and resource-based policies for cross-account access, see Cross account resource access in IAM in the *IAM User Guide*.
- **Cross-service access** – Some AWS services use features in other AWS services. For example, when you make a call in a service, it's common for that service to run applications in Amazon EC2 or store objects in Amazon S3. A service might do this using the calling principal's permissions, using a service role, or using a service-linked role.
  - **Forward access sessions (FAS)** – When you use an IAM user or role to perform actions in AWS, you are considered a principal. When you use some services, you might perform an action that then initiates another action in a different service. FAS uses the permissions of the principal calling an AWS service, combined with the requesting AWS service to make requests to downstream services. FAS requests are only made when a service receives a request that requires interactions with other AWS services or resources to complete. In this case, you must have permissions to perform both actions. For policy details when making FAS requests, see Forward access sessions.
  - **Service role** – A service role is an IAM role that a service assumes to perform actions on your behalf. An IAM administrator can create, modify, and delete a service role from within IAM. For more information, see Creating a role to delegate permissions to an AWS service in the *IAM User Guide*.
  - **Service-linked role** – A service-linked role is a type of service role that is linked to an AWS service. The service can assume the role to perform an action on your behalf. Service-linked roles appear in your AWS account and are owned by the service. An IAM administrator can view, but not edit the permissions for service-linked roles.
- **Applications running on Amazon EC2** – You can use an IAM role to manage temporary credentials for applications that are running on an EC2 instance and making AWS CLI or AWS API requests. This is preferable to storing access keys within the EC2 instance. To assign an AWS role to an EC2 instance and make it available to all of its applications, you create an instance profile

that is attached to the instance. An instance profile contains the role and enables programs that are running on the EC2 instance to get temporary credentials. For more information, see Using an IAM role to grant permissions to applications running on Amazon EC2 instances in the *IAM User Guide*.

To learn whether to use IAM roles or IAM users, see When to create an IAM role (instead of a user) in the *IAM User Guide*.

Managing access using policies

You control access in AWS by creating policies and attaching them to AWS identities or resources. A policy is an object in AWS that, when associated with an identity or resource, defines their permissions. AWS evaluates these policies when a principal (user, root user, or role session) makes a request. Permissions in the policies determine whether the request is allowed or denied. Most policies are stored in AWS as JSON documents. For more information about the structure and contents of JSON policy documents, see Overview of JSON policies in the *IAM User Guide*.

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

By default, users and roles have no permissions. To grant users permission to perform actions on the resources that they need, an IAM administrator can create IAM policies. The administrator can then add the IAM policies to roles, and users can assume the roles.

IAM policies define permissions for an action regardless of the method that you use to perform the operation. For example, suppose that you have a policy that allows the `iam:GetRole` action. A user with that policy can get role information from the AWS Management Console, the AWS CLI, or the AWS API.

Identity-based policies

Identity-based policies are JSON permissions policy documents that you can attach to an identity, such as an IAM user, group of users, or role. These policies control what actions users and roles can perform, on which resources, and under what conditions. To learn how to create an identity-based policy, see Creating IAM policies in the *IAM User Guide*.

Identity-based policies can be further categorized as *inline policies* or *managed policies*. Inline policies are embedded directly into a single user, group, or role. Managed policies are standalone policies that you can attach to multiple users, groups, and roles in your AWS account. Managed

policies include AWS managed policies and customer managed policies. To learn how to choose between a managed policy or an inline policy, see Choosing between managed policies and inline policies in the *IAM User Guide*.

Resource-based policies

Resource-based policies are JSON policy documents that you attach to a resource. Examples of resource-based policies are IAM *role trust policies* and Amazon S3 *bucket policies*. In services that support resource-based policies, service administrators can use them to control access to a specific resource. For the resource where the policy is attached, the policy defines what actions a specified principal can perform on that resource and under what conditions. You must specify a principal in a resource-based policy. Principals can include accounts, users, roles, federated users, or AWS services.

Resource-based policies are inline policies that are located in that service. You can't use AWS managed policies from IAM in a resource-based policy.

Access control lists (ACLs)

Access control lists (ACLs) control which principals (account members, users, or roles) have permissions to access a resource. ACLs are similar to resource-based policies, although they do not use the JSON policy document format.

Amazon S3, AWS WAF, and Amazon VPC are examples of services that support ACLs. To learn more about ACLs, see Access control list (ACL) overview in the *Amazon Simple Storage Service Developer Guide*.

Other policy types

AWS supports additional, less-common policy types. These policy types can set the maximum permissions granted to you by the more common policy types.

- **Permissions boundaries** – A permissions boundary is an advanced feature in which you set the maximum permissions that an identity-based policy can grant to an IAM entity (IAM user or role). You can set a permissions boundary for an entity. The resulting permissions are the intersection of an entity's identity-based policies and its permissions boundaries. Resource-based policies that specify the user or role in the `Principal` field are not limited by the permissions boundary. An explicit deny in any of these policies overrides the allow. For more information about permissions boundaries, see Permissions boundaries for IAM entities in the *IAM User Guide*.

- **Service control policies (SCPs)** – SCPs are JSON policies that specify the maximum permissions for an organization or organizational unit (OU) in AWS Organizations. AWS Organizations is a service for grouping and centrally managing multiple AWS accounts that your business owns. If you enable all features in an organization, then you can apply service control policies (SCPs) to any or all of your accounts. The SCP limits permissions for entities in member accounts, including each AWS account root user. For more information about Organizations and SCPs, see How SCPs work in the *AWS Organizations User Guide*.
- **Session policies** – Session policies are advanced policies that you pass as a parameter when you programmatically create a temporary session for a role or federated user. The resulting session's permissions are the intersection of the user or role's identity-based policies and the session policies. Permissions can also come from a resource-based policy. An explicit deny in any of these policies overrides the allow. For more information, see Session policies in the *IAM User Guide*.

Multiple policy types

When multiple types of policies apply to a request, the resulting permissions are more complicated to understand. To learn how AWS determines whether to allow a request when multiple policy types are involved, see Policy evaluation logic in the *IAM User Guide*.

How Amazon Comprehend works with IAM

Before you use IAM to manage access to Amazon Comprehend, learn what IAM features are available to use with Amazon Comprehend.

IAM features you can use with Amazon Comprehend

IAM feature	Amazon Comprehend support
Identity-based policies	Yes
Resource-based policies	Yes
Policy actions	Yes
Policy resources	Yes
Policy condition keys (service-specific)	Yes

IAM feature	Amazon Comprehend support
ACLs	No
ABAC (tags in policies)	Partial
Temporary credentials	Yes
Forward access sessions (FAS)	Yes
Service roles	Yes
Service-linked roles	No

To get a high-level view of how Amazon Comprehend and other AWS services work with most IAM features, see AWS services that work with IAM in the *IAM User Guide*.

Identity-based policies for Amazon Comprehend

Supports identity-based policies	Yes

Identity-based policies are JSON permissions policy documents that you can attach to an identity, such as an IAM user, group of users, or role. These policies control what actions users and roles can perform, on which resources, and under what conditions. To learn how to create an identity-based policy, see Creating IAM policies in the *IAM User Guide*.

With IAM identity-based policies, you can specify allowed or denied actions and resources as well as the conditions under which actions are allowed or denied. You can't specify the principal in an identity-based policy because it applies to the user or role to which it is attached. To learn about all of the elements that you can use in a JSON policy, see IAM JSON policy elements reference in the *IAM User Guide*.

Identity-based policy examples for Amazon Comprehend

To view examples of Amazon Comprehend identity-based policies, see Identity-based policy examples for Amazon Comprehend.

Resource-based policies within Amazon Comprehend

Supports resource-based policies	Yes

Resource-based policies are JSON policy documents that you attach to a resource. Examples of resource-based policies are IAM *role trust policies* and Amazon S3 *bucket policies*. In services that support resource-based policies, service administrators can use them to control access to a specific resource. For the resource where the policy is attached, the policy defines what actions a specified principal can perform on that resource and under what conditions. You must specify a principal in a resource-based policy. Principals can include accounts, users, roles, federated users, or AWS services.

To enable cross-account access, you can specify an entire account or IAM entities in another account as the principal in a resource-based policy. Adding a cross-account principal to a resource-based policy is only half of establishing the trust relationship. When the principal and the resource are in different AWS accounts, an IAM administrator in the trusted account must also grant the principal entity (user or role) permission to access the resource. They grant permission by attaching an identity-based policy to the entity. However, if a resource-based policy grants access to a principal in the same account, no additional identity-based policy is required. For more information, see Cross account resource access in IAM in the *IAM User Guide*.

The Amazon Comprehend service supports only one type of resource-based policy (a *custom model policy*), which is attached to a custom model. This policy defines other accounts that can use the custom model.

To learn how to attach a resource-based policy to a custom model, see Resource-based policies for custom models.

Policy actions for Amazon Comprehend

Supports policy actions	Yes

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

The `Action` element of a JSON policy describes the actions that you can use to allow or deny access in a policy. Policy actions usually have the same name as the associated AWS API operation.

There are some exceptions, such as *permission-only actions* that don't have a matching API operation. There are also some operations that require multiple actions in a policy. These additional actions are called *dependent actions*.

Include actions in a policy to grant permissions to perform the associated operation.

To see a list of Amazon Comprehend actions, see Actions defined by Amazon Comprehend in the *Service Authorization Reference*.

Policy actions in Amazon Comprehend use the following prefix before the action:

```
comprehend
```

To specify multiple actions in a single statement, separate them with commas.

```
"Action": [
      "comprehend:DetectSentiment",
      "comprehend:ClassifyDocument"
         ]
```

You can specify multiple actions using wildcards (*). For example, to specify all actions that begin with the word `Describe`, include the following action:

```
"Action": "comprehend:Describe*"
```

Don't use wildcards to specify all of the actions for a service. Use the best practice of granting least privilege when you specify the permissions in a policy.

To view examples of Amazon Comprehend identity-based policies, see Identity-based policy examples for Amazon Comprehend.

Policy resources for Amazon Comprehend

Supports policy resources	Yes

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

The `Resource` JSON policy element specifies the object or objects to which the action applies. Statements must include either a `Resource` or a `NotResource` element. As a best practice, specify a resource using its Amazon Resource Name (ARN). You can do this for actions that support a specific resource type, known as *resource-level permissions*.

For actions that don't support resource-level permissions, such as listing operations, use a wildcard (*) to indicate that the statement applies to all resources.

```
"Resource": "*"
```

To see a list of Amazon Comprehend resource types and their ARNs, see Resources defined by Amazon Comprehend in the *Service Authorization Reference*. To learn with which actions you can specify the ARN of each resource, see Actions defined by Amazon Comprehend.

Policy condition keys for Amazon Comprehend

Supports service-specific policy condition keys	Yes

Administrators can use AWS JSON policies to specify who has access to what. That is, which **principal** can perform **actions** on what **resources**, and under what **conditions**.

The `Condition` element (or `Condition` *block*) lets you specify conditions in which a statement is in effect. The `Condition` element is optional. You can create conditional expressions that use condition operators, such as equals or less than, to match the condition in the policy with values in the request.

If you specify multiple `Condition` elements in a statement, or multiple keys in a single `Condition` element, AWS evaluates them using a logical `AND` operation. If you specify multiple values for a single condition key, AWS evaluates the condition using a logical `OR` operation. All of the conditions must be met before the statement's permissions are granted.

You can also use placeholder variables when you specify conditions. For example, you can grant an IAM user permission to access a resource only if it is tagged with their IAM user name. For more information, see IAM policy elements: variables and tags in the *IAM User Guide*.

AWS supports global condition keys and service-specific condition keys. To see all AWS global condition keys, see AWS global condition context keys in the *IAM User Guide*.

To see a list of Amazon Comprehend condition keys, see Condition keys for Amazon Comprehend in the *Service Authorization Reference*. To learn with which actions and resources you can use a condition key, see Actions defined by Amazon Comprehend.

To view examples of Amazon Comprehend identity-based policies, see Identity-based policy examples for Amazon Comprehend.

ACLs in Amazon Comprehend

Supports ACLs	No

Access control lists (ACLs) control which principals (account members, users, or roles) have permissions to access a resource. ACLs are similar to resource-based policies, although they do not use the JSON policy document format.

ABAC with Amazon Comprehend

Supports ABAC (tags in policies)	Partial

Attribute-based access control (ABAC) is an authorization strategy that defines permissions based on attributes. In AWS, these attributes are called *tags*. You can attach tags to IAM entities (users or roles) and to many AWS resources. Tagging entities and resources is the first step of ABAC. Then you design ABAC policies to allow operations when the principal's tag matches the tag on the resource that they are trying to access.

ABAC is helpful in environments that are growing rapidly and helps with situations where policy management becomes cumbersome.

To control access based on tags, you provide tag information in the condition element of a policy using the `aws:ResourceTag/key-name`, `aws:RequestTag/key-name`, or `aws:TagKeys` condition keys.

If a service supports all three condition keys for every resource type, then the value is **Yes** for the service. If a service supports all three condition keys for only some resource types, then the value is **Partial**.

For more information about ABAC, see What is ABAC? in the *IAM User Guide*. To view a tutorial with steps for setting up ABAC, see Use attribute-based access control (ABAC) in the *IAM User Guide*.

For more information about tagging Amazon Comprehend resources, see Tagging your resources.

Using temporary credentials with Amazon Comprehend

Supports temporary credentials	Yes

Some AWS services don't work when you sign in using temporary credentials. For additional information, including which AWS services work with temporary credentials, see AWS services that work with IAM in the *IAM User Guide*.

You are using temporary credentials if you sign in to the AWS Management Console using any method except a user name and password. For example, when you access AWS using your company's single sign-on (SSO) link, that process automatically creates temporary credentials. You also automatically create temporary credentials when you sign in to the console as a user and then switch roles. For more information about switching roles, see Switching to a role (console) in the *IAM User Guide*.

You can manually create temporary credentials using the AWS CLI or AWS API. You can then use those temporary credentials to access AWS. AWS recommends that you dynamically generate temporary credentials instead of using long-term access keys. For more information, see Temporary security credentials in IAM.

Forward access sessions for Amazon Comprehend

Supports forward access sessions (FAS)	Yes

When you use an IAM user or role to perform actions in AWS, you are considered a principal. When you use some services, you might perform an action that then initiates another action in a different service. FAS uses the permissions of the principal calling an AWS service, combined with the requesting AWS service to make requests to downstream services. FAS requests are only made when a service receives a request that requires interactions with other AWS services or resources to complete. In this case, you must have permissions to perform both actions. For policy details when making FAS requests, see Forward access sessions.

Service roles for Amazon Comprehend

Supports service roles	Yes

A service role is an IAM role that a service assumes to perform actions on your behalf. An IAM administrator can create, modify, and delete a service role from within IAM. For more information, see Creating a role to delegate permissions to an AWS service in the *IAM User Guide*.

> **Warning**
>
> Changing the permissions for a service role might break Amazon Comprehend functionality. Edit service roles only when Amazon Comprehend provides guidance to do so.

To use the Amazon Comprehend asynchronous operations, you must grant Amazon Comprehend access to the Amazon S3 bucket that contains your document collection. You do this by creating a data access role in your account with a trust policy to trust the Amazon Comprehend service principal.

For a policy example, see Role-based permissions required for asynchronous operations

Service-linked roles for Amazon Comprehend

Supports service-linked roles	No

A service-linked role is a type of service role that is linked to an AWS service. The service can assume the role to perform an action on your behalf. Service-linked roles appear in your AWS account and are owned by the service. An IAM administrator can view, but not edit the permissions for service-linked roles.

For details about creating or managing service-linked roles, see AWS services that work with IAM. Find a service in the table that includes a `Yes` in the **Service-linked role** column. Choose the **Yes** link to view the service-linked role documentation for that service.

Identity-based policy examples for Amazon Comprehend

By default, users and roles don't have permission to create or modify Amazon Comprehend resources. They also can't perform tasks by using the AWS Management Console, AWS Command Line Interface (AWS CLI), or AWS API. To grant users permission to perform actions on the resources that they need, an IAM administrator can create IAM policies. The administrator can then add the IAM policies to roles, and users can assume the roles.

To learn how to create an IAM identity-based policy by using these example JSON policy documents, see Creating IAM policies in the *IAM User Guide*.

For details about actions and resource types defined by Amazon Comprehend, including the format of the ARNs for each of the resource types, see Actions, resources, and condition keys for Amazon Comprehend in the *Service Authorization Reference*.

Topics

Policy best practices

Identity-based policies determine whether someone can create, access, or delete Amazon Comprehend resources in your account. These actions can incur costs for your AWS account. When you create or edit identity-based policies, follow these guidelines and recommendations:

- **Get started with AWS managed policies and move toward least-privilege permissions** – To get started granting permissions to your users and workloads, use the *AWS managed policies* that grant permissions for many common use cases. They are available in your AWS account. We

recommend that you reduce permissions further by defining AWS customer managed policies that are specific to your use cases. For more information, see AWS managed policies or AWS managed policies for job functions in the *IAM User Guide*.

- **Apply least-privilege permissions** – When you set permissions with IAM policies, grant only the permissions required to perform a task. You do this by defining the actions that can be taken on specific resources under specific conditions, also known as *least-privilege permissions*. For more information about using IAM to apply permissions, see Policies and permissions in IAM in the *IAM User Guide*.
- **Use conditions in IAM policies to further restrict access** – You can add a condition to your policies to limit access to actions and resources. For example, you can write a policy condition to specify that all requests must be sent using SSL. You can also use conditions to grant access to service actions if they are used through a specific AWS service, such as AWS CloudFormation. For more information, see IAM JSON policy elements: Condition in the *IAM User Guide*.
- **Use IAM Access Analyzer to validate your IAM policies to ensure secure and functional permissions** – IAM Access Analyzer validates new and existing policies so that the policies adhere to the IAM policy language (JSON) and IAM best practices. IAM Access Analyzer provides more than 100 policy checks and actionable recommendations to help you author secure and functional policies. For more information, see IAM Access Analyzer policy validation in the *IAM User Guide*.
- **Require multi-factor authentication (MFA)** – If you have a scenario that requires IAM users or a root user in your AWS account, turn on MFA for additional security. To require MFA when API operations are called, add MFA conditions to your policies. For more information, see Configuring MFA-protected API access in the *IAM User Guide*.

For more information about best practices in IAM, see Security best practices in IAM in the *IAM User Guide*.

Using the Amazon Comprehend console

To access the Amazon Comprehend console, you must have a minimum set of permissions. These permissions must allow you to list and view details about the Amazon Comprehend resources in your AWS account. If you create an identity-based policy that is more restrictive than the minimum required permissions, the console won't function as intended for entities (users or roles) with that policy.

You don't need to allow minimum console permissions for users that are making calls only to the AWS CLI or the AWS API. Instead, allow access to only the actions that match the API operation that they're trying to perform.

For minimum Amazon Comprehend console permissions, you can attach the *ComprehendReadOnly* AWS managed policy to the entities. For more information, see Adding permissions to a user in the *IAM User Guide*.

To use the Amazon Comprehend console, you also need permissions for the actions shown in the following policy:

```
{
  "Version": "2012-10-17",
  "Statement": [
  {
      "Action": [
          "iam:ListRoles",
          "iam:GetRole",
          "s3:ListAllMyBuckets",
          "s3:ListBucket",
          "s3:GetBucketLocation"
      ],
      "Effect": "Allow",
      "Resource": "*"
  }
  ]
  }
```

The Amazon Comprehend console needs these additional permissions for the following reasons:

- `iam` permissions to list the available IAM roles for your account.
- `s3` permissions to access the Amazon S3 buckets and objects that contain the data for topic modeling.

When you create an asynchronous batch job or a topic modeling job using the console, you have the option to have the console create an IAM role for your job. To create an IAM role, users must be granted the following additional permissions to create IAM roles and policies, and to attach policies to roles:

```
{
```

```
  "Version": "2012-10-17",
  "Statement":
    [
      {
        "Action":
          [
          "iam:CreateRole",
          "iam:CreatePolicy",
          "iam:AttachRolePolicy"
          ],
        "Effect": "Allow",
        "Resource": "*"
      },
      {
        "Action":
          [
          "iam:PassRole"
          ],
        "Effect": "Allow",
        "Resource": "arn:aws:iam::*:role/*Comprehend*"
      }
    ]
  }
```

The Amazon Comprehend console needs these additional permissions for the following reasons:

- `iam` permissions to create roles and policies and to attach roles and policies. The `iam:PassRole` action enables the console to pass the role to Amazon Comprehend.

Allow users to view their own permissions

This example shows how you might create a policy that allows IAM users to view the inline and managed policies that are attached to their user identity. This policy includes permissions to complete this action on the console or programmatically using the AWS CLI or AWS API.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "ViewOwnUserInfo",
            "Effect": "Allow",
            "Action": [
```

```
                "iam:GetUserPolicy",
                "iam:ListGroupsForUser",
                "iam:ListAttachedUserPolicies",
                "iam:ListUserPolicies",
                "iam:GetUser"
            ],
            "Resource": ["arn:aws:iam::*:user/${aws:username}"]
        },
        {
            "Sid": "NavigateInConsole",
            "Effect": "Allow",
            "Action": [
                "iam:GetGroupPolicy",
                "iam:GetPolicyVersion",
                "iam:GetPolicy",
                "iam:ListAttachedGroupPolicies",
                "iam:ListGroupPolicies",
                "iam:ListPolicyVersions",
                "iam:ListPolicies",
                "iam:ListUsers"
            ],
            "Resource": "*"
        }
    ]
}
```

Permissions required to perform document analysis actions

The following example policy grants permissions to use the Amazon Comprehend document analysis actions:

```
{
 "Version": "2012-10-17",
 "Statement": [{
    "Sid": "AllowDetectActions",
    "Effect": "Allow",
    "Action": [
              "comprehend:DetectEntities",
              "comprehend:DetectKeyPhrases",
              "comprehend:DetectDominantLanguage",
              "comprehend:DetectSentiment",
              "comprehend:DetectTargetedSentiment",
              "comprehend:DetectSyntax",
```

```
                "textract:DetectDocumentText",
                "textract:AnalyzeDocument"
            ],
    "Resource": "*"
    }
 ]
}
```

The policy has one statement that grants permission to use the `DetectEntities`, `DetectKeyPhrases`, `DetectDominantLanguage`, `DetectTargetedSentiment`, `DetectSentiment`, and `DetectSyntax` actions. The policy statement also grants permissions to use two Amazon Textract API methods. Amazon Comprehend calls these methods to extract text from image files and scanned PDF documents. You can remove these permissions for users that never run custom inference for these types of input files.

A user with this policy would not be able to perform batch actions or asynchronous actions in your account.

The policy doesn't specify the `Principal` element because you don't specify the principal who gets the permission in an identity-based policy. When you attach a policy to a user, the user is the implicit principal. When you attach a permissions policy to an IAM role, the principal identified in the role's trust policy gets the permissions.

For a table showing all the Amazon Comprehend API actions and the resources that they apply to, see Actions, resources, and condition keys for Amazon Comprehend in the *Service Authorization Reference*.

Permissions required to use KMS encryption

To fully use Amazon Key Management Service (KMS) for data and job encryption in an asynchronous job, you need to grant permissions for the actions shown in the following policy:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
        "Action": [
            "kms:CreateGrant"
        ],
```

```
            "Effect": "Allow",
            "Resource": "*"
        },
        {
            "Action": [
                "kms:Decrypt",
                "kms:GenerateDatakey"
            ],
            "Effect": "Allow",
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "kms:ViaService": [
                        "s3.region.amazonaws.com"
                    ]
                }
            }
        }
    ]
}
```

When you create an asychronous job with Amazon Comprehend you use input data stored on Amazon S3. With S3, you have the option to encrypt your stored data, which is encrypted by S3, not by Amazon Comprehend. We can decrypt and read that encrypted input data if you provide `kms:Decrypt` permission for the key with which the original input data was encrypted to the data access role used by the Amazon Comprehend job.

You also have the option of using KMS customer-managed keys (CMK) to encrypt the output results on S3, as well as the storage volume used during job processing. When you do this, you can use the same KMS key for both types of encryption, but this is not necessary. Separate fields are available when creating the job to specify the keys for output encryption and volume encryption and you can even use a KMS key from a different account.

When using KMS encryption, `kms:CreateGrant` permission is required for volume encryption and `kms:GenerateDataKey` permission is needed for output data encryption. For reading encrypted input (as when the input data is already encrypted by Amazon S3), `kms:Decrypt` permission is required. The IAM role needs to give these permissions as needed. However, if the key is from a different account than is currently being used, the KMS key policy for that kms key must also give these permissions to the data access role for the job.

AWS managed (predefined) policies for Amazon Comprehend

AWS addresses many common use cases by providing standalone IAM policies that are created and administered by AWS. These AWS managed policies grant necessary permissions for common use cases so that you can avoid having to investigate what permissions are needed. For more information, see AWS managed policies in the *IAM User Guide*.

The following AWS managed policies, which you can attach to users in your account, are specific to Amazon Comprehend:

- **ComprehendFullAccess** – Grants full access to Amazon Comprehend resources including running topic modeling jobs. Includes permission to list and get IAM roles.
- **ComprehendReadOnly** – Grants permission to run all Amazon Comprehend actions except `StartDominantLanguageDetectionJob`, `StartEntitiesDetectionJob`, `StartKeyPhrasesDetectionJob`, `StartSentimentDetectionJob`, `StartTargetedSentimentDetectionJob`, and `StartTopicsDetectionJob`.

You need to apply the following additional policy to any user that will use Amazon Comprehend:

```
{
    "Version": "2012-10-17",
    "Statement":
      [
        {
          "Action":
            [
              "iam:PassRole"
            ],
          "Effect": "Allow",
          "Resource": "arn:aws:iam::*:role/*Comprehend*"
        }
      ]
  }
```

You can review the managed permissions policies by signing in to the IAM console and searching for specific policies there.

These policies work when you are using AWS SDKs or the AWS CLI.

You can also create your own custom IAM policies to allow permissions for Amazon Comprehend actions and resources. You can attach these custom policies to the users, groups or roles that require those permissions.

Role-based permissions required for asynchronous operations

To use the Amazon Comprehend asynchronous operations, you must grant Amazon Comprehend access to the Amazon S3 bucket that contains your document collection. You do this by creating a data access role in your account with a trust policy to trust the Amazon Comprehend service principal. For more information about creating a role, see Creating a role to delegate permissions to an AWS service in the *AWS Identity and Access Management User Guide*.

The following shows an example trust policy for the role that you create. To help with confused deputy prevention, you restrict the scope of the permission by using one or more global condition context keys. Set the `aws:SourceAccount` value to your account ID. If you use the `ArnEquals` condition, set the `aws:SourceArn` value to the ARN of the job. Use a wildcard for the job number in the ARN, because Amazon Comprehend generates this number as part of job creation.

```
{
    "Version": "2012-10-17",
    "Statement": [
      {
        "Effect": "Allow",
        "Principal": {
          "Service": "comprehend.amazonaws.com"
        },
        "Action": "sts:AssumeRole",
        "Condition": {
          "StringEquals": {
            "aws:SourceAccount": "111122223333"
          },
          "ArnEquals": {
            "aws:SourceArn": "arn:aws:comprehend:us-west-2:111122223333:pii-entities-
detection-job/*"
          }
        }
      }
    ]
 }
```

After you create the role, create an access policy for that role. This should grant the Amazon S3 `GetObject` and `ListBucket` permissions to the Amazon S3 bucket that contains your input data, and the Amazon S3 `PutObject` permission to your Amazon S3 output data bucket.

Permissions to allow all Amazon Comprehend actions

After you sign up for AWS, you create an administrator user to manage your account, including creating users and managing their permissions.

You might choose to create a user who has permissions for all Amazon Comprehend actions (think of this user as a service-specific administrator) for working with Amazon Comprehend. You can attach the following permissions policy to this user.

```
{
  "Version": "2012-10-17",
  "Statement":
    [
      {
        "Sid": "AllowAllComprehendActions",
        "Effect": "Allow",
        "Action":
        [
            "comprehend:*",
            "iam:ListRoles",
            "iam:GetRole",
            "s3:ListAllMyBuckets",
            "s3:ListBucket",
            "s3:GetBucketLocation",
            "iam:CreateRole",
            "iam:CreatePolicy",
            "iam:AttachRolePolicy",
            "kms:CreateGrant",
            "kms:Decrypt",
            "kms:GenerateDatakey"
      ],
      "Resource": "*"
    },
    {
      "Action":
        [
          "iam:PassRole"
        ],
      "Effect": "Allow",
```

```
      "Resource": "arn:aws:iam::*:role/*Comprehend*"
    }
  ]
}
```

These permissions can be modified with regard to encryption in the following ways:

- To enable Amazon Comprehend to analyze documents stored in an encrypted S3 bucket, the IAM role must have the `kms:Decrypt` permission.
- To enable Amazon Comprehend to encrypt documents stored on a storage volume attached to the compute instance that processes the analysis job, the IAM role must have the `kms:CreateGrant` permission.
- To enable Amazon Comprehend to encrypt the output results in their S3 bucket, the IAM role must have the `kms:GenerateDataKey` permission.

Permissions to allow topic modeling actions

The following permissions policy grants user permissions to perform the Amazon Comprehend topic modeling operations.

```
{
  "Version": "2012-10-17",
  "Statement": [{
  "Sid": "AllowTopicModelingActions",
  "Effect": "Allow",
  "Action": [
            "comprehend:DescribeTopicsDetectionJob",
            "comprehend:ListTopicsDetectionJobs",
            "comprehend:StartTopicsDetectionJob",
         ],
        "Resource": "*"
        ]
    }
  ]
  }
```

Permissions required for a custom asynchronous analysis job

> **Important**
>
> If you have an IAM policy which restricts model access, you won't be able to complete an inference job with a custom model. Your IAM policy should be updated to having a wildcard resource for a custom async analysis job.

If you are using the StartDocumentClassificationJob and StartEntitiesDetectionJob APIs, you need to update your IAM policy unless you are currently using wildcards as resources. If you are using a StartEntitiesDetectionJob using a pretrained model this does not impact you and you don't need to make any changes.

The following example policy contains an **outdated** reference.

```
{
    "Action": [
        "comprehend:StartDocumentClassificationJob",
        "comprehend:StartEntitiesDetectionJob",
    ],
    "Resource": [
        "arn:aws:comprehend:us-east-1:123456789012:document-classifier/myClassifier",
        "arn:aws:comprehend:us-east-1:123456789012:entity-recognizer/myRecognizer"
    ],
    "Effect": "Allow"
}
```

This is the **updated** policy you need to use to sucessfully run StartDocumentClassificationJob and StartEntitiesDetectionJob.

```
{
    "Action": [
        "comprehend:StartDocumentClassificationJob",
        "comprehend:StartEntitiesDetectionJob",
    ],
    "Resource": [
        "arn:aws:comprehend:us-east-1:123456789012:document-classifier/myClassifier",
        "arn:aws:comprehend:us-east-1:123456789012:document-classification-job/*",
        "arn:aws:comprehend:us-east-1:123456789012:entity-recognizer/myRecognizer",
        "arn:aws:comprehend:us-east-1:123456789012:entities-detection-job/*"
```

```
    ],
    "Effect": "Allow"
}
```

AWS managed policies for Amazon Comprehend

To add permissions to users, groups, and roles, it is easier to use AWS managed policies than to write policies yourself. It takes time and expertise to create IAM customer managed policies that provide your team with only the permissions they need. To get started quickly, you can use our AWS managed policies. These policies cover common use cases and are available in your AWS account. For more information about AWS managed policies, see AWS managed policies in the *IAM User Guide*.

AWS services maintain and update AWS managed policies. You can't change the permissions in AWS managed policies. Services occasionally add additional permissions to an AWS managed policy to support new features. This type of update affects all identities (users, groups, and roles) where the policy is attached. Services are most likely to update an AWS managed policy when a new feature is launched or when new operations become available. Services do not remove permissions from an AWS managed policy, so policy updates won't break your existing permissions.

Additionally, AWS supports managed policies for job functions that span multiple services. For example, the **ReadOnlyAccess** AWS managed policy provides read-only access to all AWS services and resources. When a service launches a new feature, AWS adds read-only permissions for new operations and resources. For a list and descriptions of job function policies, see AWS managed policies for job functions in the *IAM User Guide*.

AWS managed policy: ComprehendFullAccess

This policy grants full access to Amazon Comprehend resources including running topic modeling jobs. This policy also grants list and get permissions for Amazon S3 buckets and IAM roles.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
```

```
                "comprehend:*",
                "iam:GetRole",
                "iam:ListRoles",
                "s3:GetBucketLocation",
                "s3:ListAllMyBuckets",
                "s3:ListBucket",
            ],
            "Resource": "*"
        }
    ]
}
```

AWS managed policy: ComprehendReadOnly

This policy grants read-only permissions to run all Amazon Comprehend actions **except** the following:

- `StartDominantLanguageDetectionJob`
- `StartEntitiesDetectionJob`
- `StartKeyPhrasesDetectionJob`
- `StartSentimentDetectionJob`
- `StartTargetedSentimentDetectionJob`
- `StartTopicsDetectionJob`

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Action": [
                "comprehend:BatchDetectDominantLanguage",
                "comprehend:BatchDetectEntities",
                "comprehend:BatchDetectKeyPhrases",
                "comprehend:BatchDetectSentiment",
                "comprehend:BatchDetectSyntax",
                "comprehend:ClassifyDocument",
                "comprehend:ContainsPiiEntities",
                "comprehend:DescribeDocumentClassificationJob",
                "comprehend:DescribeDocumentClassifier",
                "comprehend:DescribeDominantLanguageDetectionJob",
                "comprehend:DescribeEndpoint",
```

```
                "comprehend:DescribeEntitiesDetectionJob",
                "comprehend:DescribeEntityRecognizer",
                "comprehend:DescribeKeyPhrasesDetectionJob",
                "comprehend:DescribePiiEntitiesDetectionJob",
                "comprehend:DescribeResourcePolicy",
                "comprehend:DescribeSentimentDetectionJob",
                "comprehend:DescribeTargetedSentimentDetectionJob",
                "comprehend:DescribeTopicsDetectionJob",
                "comprehend:DetectDominantLanguage",
                "comprehend:DetectEntities",
                "comprehend:DetectKeyPhrases",
                "comprehend:DetectPiiEntities",
                "comprehend:DetectSentiment",
                "comprehend:DetectSyntax",
                "comprehend:ListDocumentClassificationJobs",
                "comprehend:ListDocumentClassifiers",
                "comprehend:ListDocumentClassifierSummaries",
                "comprehend:ListDominantLanguageDetectionJobs",
                "comprehend:ListEndpoints",
                "comprehend:ListEntitiesDetectionJobs",
                "comprehend:ListEntityRecognizers",
                "comprehend:ListEntityRecognizerSummaries",
                "comprehend:ListKeyPhrasesDetectionJobs",
                "comprehend:ListPiiEntitiesDetectionJobs",
                "comprehend:ListSentimentDetectionJobs",
                "comprehend:ListTargetedSentimentDetectionJobs",
                "comprehend:ListTagsForResource",
                "comprehend:ListTopicsDetectionJobs"
            ],
            "Effect": "Allow",
            "Resource": "*"
        }
    ]
}
```

Amazon Comprehend updates to AWS managed policies

View details about updates to AWS managed policies for Amazon Comprehend since this service began tracking these changes. For automatic alerts about changes to this page, subscribe to the RSS feed on the Amazon Comprehend Document history page.

Change	Description	Date
ComprehendReadOnly – Update to an existing policy	Amazon Comprehend now allows the `comprehend:DescribeTargetedSentimentDetectionJob` and `comprehend:ListTargetedSentimentDetectionJobs` actions in the ComprehendReadOnly policy	Mar 30, 2022
ComprehendReadOnly – Update to an existing policy	Amazon Comprehend now allows the `comprehend:DescribeResourcePolicy` action in the ComprehendReadOnly policy	Feb 2, 2022
ComprehendReadOnly – Update to an existing policy	Amazon Comprehend now allows the `ListDocumentClassifierSummaries` and `ListEntityRecognizerSummaries` actions in the ComprehendReadOnly policy	September 21, 2021
ComprehendReadOnly – Update to an existing policy	Amazon Comprehend now allows the `ContainsPIIEntities` action in the ComprehendReadOnly policy	March 26, 2021
Amazon Comprehend started tracking changes	Amazon Comprehend started tracking changes for its AWS managed policies.	March 1, 2021

Troubleshooting Amazon Comprehend identity and access

Use the following information to help you diagnose and fix common issues that you might encounter when working with Amazon Comprehend and IAM.

Topics

I am not authorized to perform an action in Amazon Comprehend

If you receive an error that you're not authorized to perform an action, your policies must be updated to allow you to perform the action.

The following example error occurs when the `mateojackson` IAM user tries to use the console to view details about a fictional *`my-example-widget`* resource but does not have the fictional `comprehend:`*`GetWidget`* permissions.

```
User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform:
 comprehend:GetWidget on resource: my-example-widget
```

In this case, Mateo's policy must be updated to allow him to access the *`my-example-widget`* resource using the `comprehend:`*`GetWidget`* action.

If you need help, contact your AWS administrator. Your administrator is the person who provided you with your sign-in credentials.

I am not authorized to perform iam:PassRole

If you receive an error that you're not authorized to perform the `iam:PassRole` action, your policies must be updated to allow you to pass a role to Amazon Comprehend.

Some AWS services allow you to pass an existing role to that service instead of creating a new service role or service-linked role. To do this, you must have permissions to pass the role to the service.

The following example error occurs when an IAM user named `marymajor` tries to use the console to perform an action in Amazon Comprehend. However, the action requires the service to have

permissions that are granted by a service role. Mary does not have permissions to pass the role to the service.

```
User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform:
 iam:PassRole
```

In this case, Mary's policies must be updated to allow her to perform the `iam:PassRole` action.

If you need help, contact your AWS administrator. Your administrator is the person who provided you with your sign-in credentials.

I want to allow people outside of my AWS account to access my Amazon Comprehend resources

You can create a role that users in other accounts or people outside of your organization can use to access your resources. You can specify who is trusted to assume the role. For services that support resource-based policies or access control lists (ACLs), you can use those policies to grant people access to your resources.

To learn more, consult the following:

- To learn whether Amazon Comprehend supports these features, see How Amazon Comprehend works with IAM.
- To learn how to provide access to your resources across AWS accounts that you own, see Providing access to an IAM user in another AWS account that you own in the *IAM User Guide*.
- To learn how to provide access to your resources to third-party AWS accounts, see Providing access to AWS accounts owned by third parties in the *IAM User Guide*.
- To learn how to provide access through identity federation, see Providing access to externally authenticated users (identity federation) in the *IAM User Guide*.
- To learn the difference between using roles and resource-based policies for cross-account access, see Cross account resource access in IAM in the *IAM User Guide*.

Logging Amazon Comprehend API calls with AWS CloudTrail

Amazon Comprehend is integrated with AWS CloudTrail, a service that provides a record of actions taken by a user, role, or an AWS service in Amazon Comprehend. CloudTrail captures API calls for Amazon Comprehend as events. The calls captured include calls from the Amazon Comprehend

console and code calls to the Amazon Comprehend API operations. If you create a trail, you can enable continuous delivery of CloudTrail events to an Amazon S3 bucket, including events for Amazon Comprehend. If you don't configure a trail, you can still view the most recent events in the CloudTrail console in **Event history**. Using the information collected by CloudTrail, you can determine the request that was made to Amazon Comprehend, the IP address from which the request was made, who made the request, when it was made, and additional details.

To learn more about CloudTrail, including how to configure and enable it, see the AWS CloudTrail User Guide.

Amazon Comprehend information in CloudTrail

CloudTrail is enabled on your AWS account when you create the account. When supported event activity occurs in Amazon Comprehend, that activity is recorded in a CloudTrail event along with other AWS service events in **Event history**. You can view, search, and download recent events in your AWS account. For more information, see Viewing events with CloudTrail event history.

For an ongoing record of events in your AWS account, including events for Amazon Comprehend, create a trail. A *trail* enables CloudTrail to deliver log files to an Amazon S3 bucket. By default, when you create a trail in the console, the trail applies to all AWS Regions. The trail logs events from all Regions in the AWS partition and delivers the log files to the Amazon S3 bucket that you specify. Additionally, you can configure other AWS services to further analyze and act upon the event data collected in CloudTrail logs. For more information, see the following:

- Overview for creating a trail
- CloudTrail supported services and integrations
- Configuring Amazon SNS notifications for CloudTrail
- Receiving CloudTrail log files from multiple regions and Receiving CloudTrail log files from multiple accounts

Amazon Comprehend supports logging the following actions as events in CloudTrail log files:

- BatchDetectDominantLanguage
- BatchDetectEntities
- BatchDetectKeyPhrases
- BatchDetectSentiment
- BatchDetectSyntax

- ClassifyDocument
- CreateDocumentClassifier
- CreateEndpoint
- CreateEntityRecognizer
- DeleteDocumentClassifier
- DeleteEndpoint
- DeleteEntityRecognizer
- DescribeDocumentClassificationJob
- DescribeDocumentClassifier
- DescribeDominantLanguageDetectionJob
- DescribeEndpoint
- DescribeEntitiesDetectionJob
- DescribeEntityRecognizer
- DescribeKeyPhrasesDetectionJob
- DescribePiiEntitiesDetectionJob
- DescribeSentimentDetectionJob
- DescribeTargetedSentimentDetectionJob
- DescribeTopicsDetectionJob
- DetectDominantLanguage
- DetectEntities
- DetectKeyPhrases
- DetectPiiEntities
- DetectSentiment
- DetectSyntax
- ListDocumentClassificationJobs
- ListDocumentClassifiers
- ListDominantLanguageDetectionJobs
- ListEndpoints
- ListEntitiesDetectionJobs

- ListEntityRecognizers
- ListKeyPhrasesDetectionJobs
- ListPiiEntitiesDetectionJobs
- ListSentimentDetectionJobs
- ListTargetedSentimentDetectionJobs
- ListTagsForResource
- ListTopicsDetectionJobs
- StartDocumentClassificationJob
- StartDominantLanguageDetectionJob
- StartEntitiesDetectionJob
- StartKeyPhrasesDetectionJob
- StartPiiEntitiesDetectionJob
- StartSentimentDetectionJob
- StartTargetedSentimentDetectionJob
- StartTopicsDetectionJob
- StopDominantLanguageDetectionJob
- StopEntitiesDetectionJob
- StopKeyPhrasesDetectionJob
- StopPiiEntitiesDetectionJob
- StopSentimentDetectionJob
- StopTargetedSentimentDetectionJob
- StopTrainingDocumentClassifier
- StopTrainingEntityRecognizer
- TagResource
- UntagResource
- UpdateEndpoint

Every event or log entry contains information about who generated the request. The identity information helps you determine the following:

- Whether the request was made with the root user credentials.

- Whether the request was made with temporary security credentials for a role or federated user.
- Whether the request was made by another AWS service.

For more information, see the CloudTrail userIdentity element.

Example: Amazon Comprehend log file entries

A trail is a configuration that enables delivery of events as log files to an Amazon S3 bucket that you specify. CloudTrail log files contain one or more log entries. An event represents a single request from any source and includes information about the requested action, the date and time of the action, request parameters, and so on. CloudTrail log files aren't an ordered stack trace of the public API calls, so they don't appear in any specific order.

The following example shows a CloudTrail log entry that demonstrates the `ClassifyDocument` action.


```
{
    "eventVersion": "1.08",
    "userIdentity": {
        "type": "IAMUser",
        "principalId": "AROAICFHPEXAMPLE",
        "arn": "arn:aws:iam::12345678910:user/myadmin2",
        "accountId": "12345678910",
        "accessKeyId": "ASIA3VZEXAMPLE",
        "sessionContext": {
            "sessionIssuer": {},
            "webIdFederationData": {},
            "attributes": {
                "creationDate": "2023-10-19T14:22:09Z",
                "mfaAuthenticated": "false"
            }
        }
    },
    "eventTime": "2023-10-19T17:31:20Z",
    "eventSource": "comprehend.amazonaws.com",
    "eventName": "ClassifyDocument",
    "awsRegion": "us-east-2",
    "sourceIPAddress": "3.21.185.237",
    "userAgent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15; rv:109.0)
 Gecko/20100101 Firefox/115.0",
    "requestParameters": null,
```

```
    "responseElements": null,
    "requestID": "fd916e66-caac-46c9-a1fc-81a0ef33e61b",
    "eventID": "535ca22b-b3a3-4c13-b2c5-bf51ab082794",
    "readOnly": false,
    "resources": [
        {
            "accountId": "12345678910",
            "type": "AWS::Comprehend::DocumentClassifierEndpoint",
            "ARN": "arn:aws:comprehend:us-east-2:12345678910:document-classifier-
endpoint/endpointExample"
        }
    ],
    "eventType": "AwsApiCall",
    "recipientAccountId": "12345678910"
}
```

Compliance validation for Amazon Comprehend

Third-party auditors assess the security and compliance of Amazon Comprehend as part of multiple AWS compliance programs. These include PCI, FedRAMP, HIPAA, and others. You can download third-party audit reports using AWS Artifact. For more information, see Downloading Reports in AWS Artifact.

Your compliance responsibility when using Amazon Comprehend is determined by the sensitivity of your data, your company's compliance objectives, and applicable laws and regulations. AWS provides the following resources to help with compliance:

- Security and Compliance Quick Start Guides – These deployment guides discuss architectural considerations and provide steps for deploying security- and compliance-focused baseline environments on AWS.
- Architecting for HIPAA Security and Compliance Whitepaper – This whitepaper describes how companies can use AWS to create HIPAA-compliant applications.
- AWS Compliance Resources – This collection of workbooks and guides might apply to your industry and location.
- AWS Config – This AWS service assesses how well your resource configurations comply with internal practices, industry guidelines, and regulations.
- AWS Security Hub – This AWS service provides a comprehensive view of your security state within AWS that helps you check your compliance with security industry standards and best practices.

For a list of AWS services in scope of specific compliance programs, see AWS Services in Scope by Compliance Program. For general information, see AWS Compliance Programs.

Resilience in Amazon Comprehend

The AWS global infrastructure is built around AWS Region s and Availability Zones. AWS Region s provide multiple physically separated and isolated Availability Zones, which are connected with low-latency, high-throughput, and highly redundant networking. With Availability Zones, you can design and operate applications and databases that automatically fail over between Availability Zones without interruption. Availability Zones are more highly available, fault tolerant, and scalable than traditional single or multiple data center infrastructures.

For more information about AWS Region s and Availability Zones, see AWS Global Infrastructure.

Infrastructure security in Amazon Comprehend

As a managed service, Amazon Comprehend is protected by the AWS global network security. For information about AWS security services and how AWS protects infrastructure, see AWS Cloud Security. To design your AWS environment using the best practices for infrastructure security, see Infrastructure Protection in *Security Pillar AWS Well-Architected Framework*.

You use AWS published API calls to access Amazon Comprehend through the network. Clients must support the following:

- Transport Layer Security (TLS). We require TLS 1.2 and recommend TLS 1.3.
- Cipher suites with perfect forward secrecy (PFS) such as DHE (Ephemeral Diffie-Hellman) or ECDHE (Elliptic Curve Ephemeral Diffie-Hellman). Most modern systems such as Java 7 and later support these modes.

Additionally, requests must be signed by using an access key ID and a secret access key that is associated with an IAM principal. Or you can use the AWS Security Token Service (AWS STS) to generate temporary security credentials to sign requests.

Guidelines and quotas

Unless otherwise specified, the Amazon Comprehend quotas are per region. You can request an increase to adjustable quotas if needed for your applications. For information about quotas and to request a quota increase, see AWS Service Quotas.

Topics

Supported Regions

Amazon Comprehend is available in the following AWS Regions:

- US East (Ohio)
- US East (N. Virginia)
- US West (Oregon)
- Asia Pacific (Mumbai)
- Asia Pacific (Seoul)
- Asia Pacific (Singapore)
- Asia Pacific (Sydney)
- Asia Pacific (Tokyo)
- Canada (Central)
- Europe (Frankfurt)
- Europe (Ireland)
- Europe (London)
- AWS GovCloud (US-West)

By default, Amazon Comprehend provides all API operations in each of the supported regions. For exceptions, see Document processing.

For information about API endpoints, see Amazon Comprehend Regions and Endpoints in the *Amazon Web Services General Reference*.

To review current quotas in a region, or to request quota increases for adjustable quotas, open the Service Quotas console.

Quotas for built-in models

Amazon Comprehend provides built-in models for you to analyze UTF-8 text documents. Amazon Comprehend provides synchronous and asynchronous operations that use the built-in models.

Topics

Real-time (synchronous) analysis

This section describes quotas related to real-time analysis using the built-in models.

Topics

Single document operations

The Amazon Comprehend API provides operations that take a single document as input. The following quotas apply to these operations.

General quotas for single document operations

The following quotas apply to real-time analysis for detecting entities, key-phrases, or dominant language. For entity detection, these quotas apply to detection with the built-in models. For custom entity detection, see the quotas in Custom entity recognition .

Description	Quota/Guideline
Maximum document size	100 KB

Operation-specific quotas for single document operations

The following quotas apply to real-time analysis for detecting sentiment, targeted sentiment, and syntax.

Description	Quota/Guideline
Maximum document size	5 KB

Multiple document operations

The Amazon Comprehend API provides batch operations that process multiple documents with a single API request. The following quotas apply to the batch operations.

Description	Quota/Guideline
Maximum document size	5 KB
Maximum documents per request	25

For more information about using the batch document operations, see Multiple document synchronous processing.

Request throttling for real-time (synchronous) requests

Amazon Comprehend applies dynamic throttling to synchronous requests. If system processing bandwidth is available, Amazon Comprehend gradually increases the number of your requests that it processes. To control your application's usage of the synchronous API operations, we recommend that you turn on billing alerts or implement rate-limiting in your application.

Asynchronous analysis

This section describes quotas related to asynchronous analysis using the built-in models.

Asynchronous API operations each support a maximum of 10 active jobs. To view the quotas for each API operation, see the Service Quotas table in Amazon Comprehend endpoints and quotas in the *Amazon Web Services General Reference*.

For adjustable quotas, you can request a quota increase using the Service Quotas console.

Topics

General quotas for asynchronous operations

You can run asynchronous analysis jobs using the console or any of the API `Start*` operations. For information about when to use asynchronous operations, see Asynchronous batch processing. The following quotas apply to most of the API `Start*` operations for built-in models. For the exceptions, see Operation-specific quotas for asynchronous jobs.

Description	Quota/Guideline
Maximum size of each document in jobs that detect entities, key phrases, PII, and languages	1 MB
Maximum total size of all files in a request	5 GB
Minimum total size of all files in a request	500 bytes
Maximum number of files, one document per file	1,000,000
Maximum total number of lines, one document per line	1,000,000

Operation-specific quotas for asynchronous jobs

This section describes quotas for specific asynchronous operations. If a quota isn't specified in the following tables, the general quota value applies.

Topics

- Sentiment
- Targeted sentiment
- Events
- Topic modeling

Sentiment

Asynchronous sentiment jobs, which you create with the StartSentimentDetectionJob operation, have the following quotas.

Description	Quota/Guideline
Maximum size of each input document	5 KB

Targeted sentiment

Asynchronous targeted sentiment jobs, which you create with the StartTargetedSentimentDetectionJob operation, have the following quotas.

Description	Quota/Guideline
Supported document formats	UTF-8
Maximum size of each document in a job	10 KB
Maximum size of all documents in a job	300 MB
Maximum number of files, one document per file	30,000
Maximum total number of lines, one document per line (for all files in a request)	30,000

Events

Asynchronous events detection jobs, which you create with the StartEventsDetectionJob operation, have the following quotas.

Description	Quotas
Character encoding	UTF-8
Total size of all files in a job	50 MB
Maximum size of each document in a job	10 KB
Maximum number of files, one document per file	5,000
Maximum total number of lines, one document per line (for all files in request)	5,000

Topic modeling

Asynchronous topic modeling jobs, which you create with the StartTopicsDetectionJob operation, have the following quotas.

Description	Quota/Guideline
Character encoding	UTF-8
Maximum number of topics to return	100
Maximum file size for one file, one document per file	100 MB

For more information, see Topic modeling

Request throttling for asynchronous requests

Each asynchronous API operation supports a maximum number of requests per second (per region, per account), and also a maximum of 10 active jobs. To view the quotas for each API operation, see the Service Quotas table in Amazon Comprehend endpoints and quotas in the *Amazon Web Services General Reference*.

For adjustable quotas, you can request a quota increase using the Service Quotas console.

Quotas for custom models

You can use Amazon Comprehend to build your own custom models for custom classification and custom entity recognition. This section provides the guidelines and quotas related to training and using custom models. For more information about custom models, see Amazon Comprehend Custom.

Topics

General quotas

Amazon Comprehend sets general size quotas for each type of input document that you can analyze with custom models. For real-time analysis quotas, see Maximum document sizes for real-time analysis. For asynchronous analysis quotas, see Inputs for asynchronous custom analysis.

Each asynchronous API operation supports a maximum number of requests per second (per region, per account), and also a maximum of 10 active jobs. To view the quotas for each API operation, see the Service Quotas table in Amazon Comprehend endpoints and quotas in the *Amazon Web Services General Reference*.

For adjustable quotas, you can request a quota increase using the Service Quotas console.

Quotas for endpoints

You create an endpoint to run real-time analysis with a custom model. For information about endpoints, see Managing Amazon Comprehend endpoints.

The following quotas apply to the endpoints. For information about how to request a quota increase, see AWS Service Quotas.

Description	Quota/Guideline
Maximum number of active endpoints per Region for each account	20

Description	Quota/Guideline
Maximum number of inference units per Region for each account	200
Maximum number of inference units per endpoint per region	50
Maximum throughput per inference unit (characters)	100/second
Maximum throughput per inference unit (documents)	2/second

Document classification

This section describes the guidelines and quotas for the following document classification operations:

- Classifier training jobs that you start with the CreateDocumentClassifier operation.
- Asynchronous document classification jobs that you start with the StartDocumentClassificationJob operation.
- Synchronous document classification requests that use the ClassifyDocument operation.

General quotas for document classification

The following table describes general quotas related to training custom classifiers.

Description	Quota/Guideline
Maximum length of class name	5,000 characters
Number of classes (multi-class mode)	2–1,000
Number of classes (multi-label mode)	2–100
Annotations format	
Minimum number of annotations per class (multi-class mode)	10
Minimum number of annotations per class (multi-label mode)	10

Description	Quota/Guideline
Minimum number of annotations (multi-label mode)	50
CSV file format	
Minimum number of training documents per class (multi-class mode)	50
Minimum number of training documents per class (multi-label mode)	10
Minimum number of training documents (multi-label mode)	50

Classification for plain text documents

You create and train a plain-text model using plain-text input documents. Amazon Comprehend provides real-time and asynchronous operations to classify plain text documents using a plain-text model.

Training

The following table describes quotas related to training a custom classifier with plain text documents.

Description	Quota/Guideline
Total size of all files in training job	5 GB
Maximum number of augmented manifest files for training a custom classifier	5
Maximum number of attribute names for each augmented manifest file	5
Maximum length of attribute name	63 characters

Real-time (synchronous) analysis

The following table describes quotas related to real-time classification of plain text documents.

Description	Quota/Guideline
Maximum number of documents per synchronous request	1
Maximum text document size (UTF-8 encoded)	10 KB

Asynchronous analysis

The following table describes quotas related to asynchronous classification of plain text documents.

Description	Quota/Guideline
Total size of all files in asynchronous job	5 GB
Maximum file size for one file, one document per file	10 MB
Maximum number of files, one document per file	1,000,000
Maximum total number of lines, one document per line (for all files in request)	1,000,000

Classification for semi-structured documents

This section describes the guidelines and quotas for document classification of semi-structured documents. To classify semi-structured documents, use a native document model that you trained with native input documents.

Training a native document model with semi-structured docs

The following table describes quotas related to training a custom classifier with semi-structured documents, such as PDF documents, Word documents, and image files.

Description	Quota/Guideline
Maximum number of pages across all documents	10,000
Maximum annotations file size (all CSV file sizes combined)	5 MB
Document corpus size (training and test documents)	10 GB
File sizes for training and testing files	
Image file size (JPG, PNG, TIFF).	1 byte–10 MB. TIFF files: one page maximum.
Page size for PDF documents	1 byte–10 MB
Page size for Word documents	1 byte–10 MB
Amazon Textract API output JSON size	1 byte–1 MB

Real-time (synchronous) analysis

This section describes quotas related to real-time classification of semi-structured documents.

The following table shows the maximum file sizes for input documents. For all input document types, the input file maximum is one page, with no more than 10,000 characters.

File type	Maximum size (API)	Maximum size (console)
UTF-8 text documents	10 KB	10 KB
PDF documents	10 MB	5 MB
Word documents	10 MB	5 MB
Image files	10 MB	5 MB
Amazon Textract API output size	1 MB	n/a

Asynchronous analysis

The following table describes quotas related to asynchronous classification of semi-structured documents.

Description	Quota/Guideline
Maximum number of pages across all input documents for a job	25,000
Document corpus size	25 GB
Image file size (JPG, PNG, or TIFF)	1 byte–10 MB. TIFF files: one page maximum.
Page size for PDF documents	1 byte–10 MB
Page size for Word documents	1 byte–10 MB
Textract API output JSON size	1 byte–1 MB.

Custom entity recognition

This section describes the guidelines and quotas for the following operations for custom entity recognition:

- Entity recognizer training jobs started with the CreateEntityRecognizer operation.
- Asynchronous entity recognition jobs started with the StartEntitiesDetectionJob operation.
- Synchronous entity recognition requests using the DetectEntities operation.

Custom entity recognition for plain text documents

Amazon Comprehend provides async and sync operations to analyze plain text documents with a custom entity recognizer.

Training

This section describes quotas related to training a custom entity recognizer to analyze plain text documents. To train the model, you can provide an entity list or a set of annotated text documents.

The following table describes quotas related to training the model with an entity list.

Description	Quota/Guideline
Number of entities per model	1–25
Document size (UTF-8)	1–5,000 byte
Number of items in entity list	1–1 million
Length of individual entry (post-strip) in entry list	1–5,000
Entity list corpus size (all docs in plaintext combined)	5 KB –200 MB

The following table describes quotas related to training the model with annotated text documents.

Description	Quota/Guideline
Number of entities per model/custom entity recognizer	1–25
Document size (UTF-8)	1–5,000 byte
Number of documents (see Plain-text annotations)	3–200,000
Document corpus size (all docs in plaintext combined)	5 KB - 200 MB
Minimum number of annotations per entity	25

Real-time (synchronous) analysis

The following table describes quotas related to real-time analysis of plain text documents.

Description	Quota/Guideline
Maximum number of documents per synchronous request	1
Maximum text document size (UTF-8 encoded)	5 KB

Asynchronous analysis

The following table describes quotas related to asynchronous entity recognition of plain text documents.

Description	Quota/Guideline
Document size (UTF-8)	1 byte–1 MB
Maximum number of files, one document per file	1,000,000
Maximum total number of lines, one document per line (for all files in request)	1,000,000
Document corpus size (all docs in plaintext combined)	1 byte–5 GB

Custom entity recognition for semi-structured documents

Amazon Comprehend provides async and sync operations to analyze semi-structured documents with a custom entity recognizer. You must train the model using annotated PDF documents.

Training

The following table describes quotas related to training a custom entity recognizer (CreateEntityRecognizer) to analyze semi-structured documents.

Description	Quota/Guideline
Number of entities per model/custom entity recognizer	1–25
Maximum annotation file size (UTF-8 JSON)	5 MB

Description	Quota/Guideline
Number of documents	250–10,000
Document corpus size (all docs in plaintext combined)	5 KB–1 GB
Minimum number of annotations per entity	100
Maximum number of augmented manifest files for training a custom entity recognizer	5
Maximum number of attribute names for each augmented manifest file	5
Maximum length of attribute name	63 characters

Real-time (synchronous) analysis

This section describes quotas related to real-time analysis of semi-structured documents.

The following table shows the maximum file sizes for input documents. For all input document types, the input file maximum is one page, with no more than 10,000 characters.

File type	Maximum size (API)	Maximum size (console)
UTF-8 text documents	10 KB	10 KB
PDF documents	10 MB	5 MB
Word documents	10 MB	5 MB
Image files	10 MB	5 MB
Textract output files	1 MB	n/a

Asynchronous analysis

This section describes quotas for asynchronous analysis of semi-structured documents.

Description	Quota/Guideline
Image size (JPG or PNG)	1 byte–10 MB
Image size (TIFF)	1 byte–10 MB. Maximum one page.
Document size (PDF)	1 byte–50 MB
Document size (Docx)	1 byte–5 MB
Document size (UTF-8)	1 byte–1 MB
Maximum number of files, one document per file (one document per line not allowed for image files or PDF/Word documents)	500
Maximum number of pages for a PDF or Docx file	100
Document corpus size after text extraction (plaintext, all files combined)	1 byte–5 GB

For more information about limits for images, see Hard Limits in Amazon Textract

Quotas for flywheels

Use flywheels to manage training and tracking of custom model versions for custom classification and custom entity recognition. For more information about Flywheels, see Flywheels.

General quotas for flywheels

The follow quotas apply to flywheels and flywheel iterations.

Description	Quota/Guideline
Maximum number of flywheels	50
Maximum number of flywheels in CREATING state	10

Description	Quota/Guideline
Maximum number of training datasets per flywheel	50
Maximum number of test datasets per flywheel	50
Maximum number of datasets with INGESTING status	10
Maximum number of in-progress flywheel iterations per account	10

Dataset quotas for custom classification models

When you ingest a dataset for a flywheel associated with a custom classification model, the following quotas apply.

Description	Quota/Guideline
Minimum number of training documents per class (multi-label mode)	50
Maximum number of training documents	1,000,000
Minimum dataset size	500 bytes
Maximum dataset size	5 GB
Maximum file size for one file, one document per file	10 MB

Dataset quotas for custom entity recognition models

When you ingest a dataset for a flywheel associated with a custom entity recognition model, the following quotas apply.

Description	Quota/Guideline
Maximum document size	5 KB

Description	Quota/Guideline
Minimum number of training documents	3
Maximum number of training documents	200,000
Minimum number of annotations per entity	25
Maximum dataset size	200 MB

Tutorials and other resources

Tutorials and other resources for Amazon Comprehend.

Topics

Tutorial: Analyzing insights from customer reviews with Amazon Comprehend

This tutorial explains how to use Amazon Comprehend with Amazon Simple Storage Service, AWS Glue, Amazon Athena, and Amazon QuickSight to gain valuable insights into your documents. Amazon Comprehend can extract sentiment (the mood of a document) and entities (names of people, organizations, events, dates, products, places, quantities, and titles) from unstructured text.

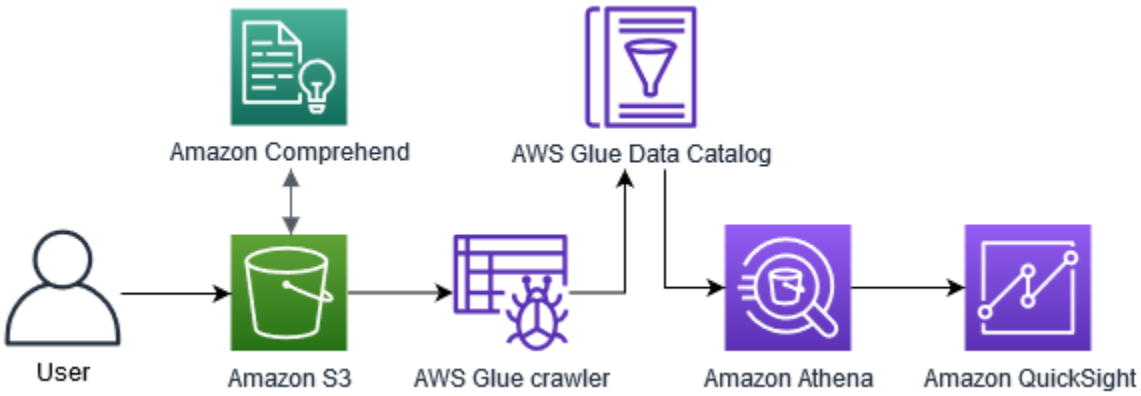
For example, you can get actionable insights from customer reviews. In this tutorial, you analyze a sample dataset of customer reviews about a novel. You use Amazon Comprehend sentiment analysis to determine whether customers feel positive or negative about the novel. You also use Amazon Comprehend entities analysis to discover mentions of important entities, such as related novels or authors. After following this tutorial, you might discover that over 50% of the reviews are positive. You might also discover that customers are comparing authors and expressing interest in other classic novels.

In this tutorial, you accomplish the following:

- Store a sample dataset of reviews in Amazon Simple Storage Service (Amazon S3). Amazon Simple Storage Service is an object storage service.
- Use Amazon Comprehend to analyze the sentiment and entities in the review documents.
- Use an AWS Glue crawler to store the results of the analysis in a database. AWS Glue is an extract, transform, and load (ETL) service that lets you catalog and clean your data for analytics.
- Run Amazon Athena queries to clean your data. Amazon Athena is a serverless interactive query service.

- Create visualizations with your data in Amazon QuickSight. Amazon QuickSight is a serverless business intelligence tool for extracting insights from your data.

The following diagram shows the workflow.



Estimated time to complete this tutorial: 1 hour

Estimated cost: Some of the actions in this tutorial incur charges on your AWS account. For information about the charges for each of these services, see the following pricing pages.

- Amazon S3 pricing
- Amazon Comprehend pricing
- AWS Glue pricing
- Amazon Athena pricing
- Amazon QuickSight pricing

Topics

Prerequisites

To complete this tutorial, you need the following:

- An AWS account. For information about setting up an AWS account, see Setting up.
- An IAM entity (user, group or role). To learn how to set up users and groups for your account, see the Getting started tutorial in the *IAM User Guide*.
- The following permissions policy attached to your user, group or role. The policy grants some of the permissions required to complete this tutorial. The next prerequisite describes the additional permissions you need.

```
{
  "Version": "2012-10-17",
  "Statement":
  [
    {
      "Sid": "VisualEditor0",
      "Effect": "Allow",
      "Action":
      [
        "comprehend:*",
        "ds:AuthorizeApplication",
        "ds:CheckAlias",
        "ds:CreateAlias",
        "ds:CreateIdentityPoolDirectory",
        "ds:DeleteDirectory",
        "ds:DescribeDirectories",
        "ds:DescribeTrusts",
        "ds:UnauthorizeApplication",
        "iam:AttachRolePolicy",
        "iam:CreatePolicy",
        "iam:CreatePolicyVersion",
        "iam:CreateRole",
        "iam:DeletePolicyVersion",
        "iam:DeleteRole",
        "iam:DetachRolePolicy",
        "iam:GetPolicy",
        "iam:GetPolicyVersion",
        "iam:GetRole",
        "iam:ListAccountAliases",
        "iam:ListAttachedRolePolicies",
        "iam:ListEntitiesForPolicy",
```

```
        "iam:ListPolicies",
        "iam:ListPolicyVersions",
        "iam:ListRoles",
        "quicksight:*",
        "s3:*",
        "tag:GetResources"
      ],
      "Resource": "*"
    },
    {
      "Action":
      [
        "iam:PassRole"
      ],
      "Effect": "Allow",
      "Resource":
      [
        "arn:aws:iam::*:role/*Comprehend*"
      ]
    }
  ]
}
```

Use the previous policy to create an IAM policy and attach it to your group or user. For information about creating an IAM policy, see Creating IAM policies in the *IAM User Guide*. For information about attaching an IAM policy, see Adding and removing IAM identity permissions in the *IAM User Guide*.

- Managed policies attached to your IAM group or user. In addition to the previous policy, you must also attach the following AWS managed policies to your group or user:
  - `AWSGlueConsoleFullAccess`
  - `AWSQuicksightAthenaAccess`

  These managed policies give you permission to use AWS Glue, Amazon Athena, and Amazon QuickSight. For information about attaching an IAM policy, see Adding and removing IAM identity permissions in the *IAM User Guide*.

Step 1: Adding documents to Amazon S3

Before starting the Amazon Comprehend analysis jobs, you need to store a sample dataset of customer reviews in Amazon Simple Storage Service (Amazon S3). Amazon S3 hosts your data in

containers called buckets. Amazon Comprehend can analyze documents stored in a bucket and it sends results of the analysis to a bucket. In this step, you create an S3 bucket, create input and output folders in the bucket, and upload a sample dataset to the bucket.

Topics

Prerequisites

Before you begin, review Tutorial: Analyzing insights from customer reviews with Amazon Comprehend and complete the prerequisites.

Download sample data

The following sample dataset contains Amazon reviews taken from the larger dataset "Amazon reviews - Full", which was published with the article "Character-level Convolutional Networks for Text Classification" (Xiang Zhang et al., 2015). Download the dataset to your computer.

To get the sample data

1. Download the zip file tutorial-reviews-data.zip to your computer.
2. Extract the zip file on your computer. There are two files. The file `THIRD_PARTY_LICENSES.txt` is the open source license for the dataset published by Xiang Zhang et al. The file `amazon-reviews.csv` is the dataset you analyze in the tutorial.

Create an Amazon S3 bucket

After downloading the sample dataset, create an Amazon S3 bucket to store your input and output data. You can create an S3 bucket using the Amazon S3 console or the AWS Command Line Interface (AWS CLI).

Create an Amazon S3 bucket (console)

In the Amazon S3 console, you create a bucket with a name that is unique in all of AWS.

To create an S3 bucket (console)

1. Sign in to the AWS Management Console and open the Amazon S3 console at https://console.aws.amazon.com/s3/.
2. In **Buckets**, choose **Create bucket**.
3. For **Bucket name**, enter a globally unique name that describes the bucket's purpose.
4. For **Region**, choose the AWS Region where you want to create the bucket. The Region you choose must support Amazon Comprehend. To reduce latency, choose the AWS Region closest to your geographic location that is supported by Amazon Comprehend. For a list of Regions that support Amazon Comprehend, see the Region table in the *Global Infrastructure Guide*.
5. Leave the default settings for **Object Ownership**, **Bucket settings for Block Public Access**, **Bucket Versioning**, and **Tags**.
6. For **Default encryption**, choose **Disable**.

> **Tip**
>
> While this tutorial does not use encryption, you might want to use encryption when analyzing important data. For end-to-end encryption, you can encrypt your data at rest in the bucket and also when you run analysis jobs. For more information about encryption with AWS, see What is AWS Key Management Service? in the *AWS Key Management Service Developer Guide*.

7. Review your bucket configurations and then choose **Create bucket**.

Create an Amazon S3 bucket (AWS CLI)

After opening the AWS CLI, you run the `create-bucket` command to create a bucket that will store the input and output data.

To create an Amazon S3 bucket (AWS CLI)

1. To create your bucket, run the following command in the AWS CLI. Replace DOC-EXAMPLE-BUCKET with a name for the bucket that is unique in all of AWS.

```
aws s3api create-bucket --bucket DOC-EXAMPLE-BUCKET
```

By default, the `create-bucket` command creates a bucket in the `us-east-1` AWS Region. To create a bucket in an AWS Region other than `us-east-1`, add the `LocationConstraint` parameter to specify your Region. For example, the following command creates a bucket in the `us-west-2` Region.

```
aws s3api create-bucket --bucket DOC-EXAMPLE-BUCKET
--region us-west-2 --create-bucket-configuration LocationConstraint=us-west-2
```

Note that only certain Regions support Amazon Comprehend. For a list of Regions that support Amazon Comprehend, see the Region table in the *Global Infrastructure Guide*.

2. To ensure that your bucket was created successfully, run the following command. The command lists all of the S3 buckets associated with your account.

```
aws s3 ls
```

(Console only) create folders

Next, create two folders in your S3 bucket. The first folder is for your input data. The second folder is where Amazon Comprehend sends the analysis results. If you use the Amazon S3 console, you have to manually create the folders. If you use the AWS CLI, you can create folders when you upload the sample dataset or run an analysis job. For that reason, we provide a procedure for creating folders only for console users. If you are using the AWS CLI, you will create folders in Upload the input data and in Step 3: Running analysis jobs on documents in Amazon S3.

To create folders in your S3 bucket (console)

1. Open the Amazon S3 console at https://console.aws.amazon.com/s3/.
2. In **Buckets**, choose your bucket from the list of buckets.
3. In the **Overview** tab, choose **Create folder**.
4. For the new folder name, enter `input`.
5. For the encryption settings, choose **None (Use bucket settings)**.
6. Choose **Save**.
7. Repeat steps 3 through 6 to create another folder for the output of the analysis jobs, but in step 4, enter the new folder name `output`.

Upload the input data

Now that you have a bucket, upload the sample dataset `amazon-reviews.csv`. You can upload data to S3 buckets with the Amazon S3 console or the AWS CLI.

Upload sample documents to a bucket (console)

In the Amazon S3 console, upload the sample dataset file to the input folder.

To upload the sample documents (console)

1. Open the Amazon S3 console at https://console.aws.amazon.com/s3/.
2. In **Buckets**, choose your bucket from the list of buckets.
3. Choose the `input` folder and then choose **Upload**.
4. Choose **Add files** and then choose the `amazon-reviews.csv` file on your computer.
5. Leave the other settings at their default values.
6. Choose **Upload**.

Upload sample documents to a bucket (AWS CLI)

Create an input folder in your S3 bucket and upload the dataset file to the new folder with the `cp` command.

To upload the sample documents (AWS CLI)

1. To upload the `amazon-reviews.csv` file to a new folder in your bucket, run the following AWS CLI command. Replace DOC-EXAMPLE-BUCKET with the name of your bucket. By adding the path `/input/` at the end, Amazon S3 automatically creates a new folder called `input` in your bucket and uploads the dataset file to that folder.

   ```
   aws s3 cp amazon-reviews.csv s3://DOC-EXAMPLE-BUCKET/input/
   ```

2. To ensure that your file was uploaded successfully, run the following command. The command lists the contents of your bucket's `input` folder.

   ```
   aws s3 ls s3://DOC-EXAMPLE-BUCKET/input/
   ```

Now, you have an S3 bucket with the `amazon-reviews.csv` file in a folder called `input`. If you used the console, you also have an `output` folder in the bucket. If you used the AWS CLI, you will create the output folder when running the Amazon Comprehend analysis jobs.

Step 2: (CLI only) creating an IAM role for Amazon Comprehend

This step is necessary only if you are using the AWS Command Line Interface (AWS CLI) to complete this tutorial. If you are using the Amazon Comprehend console to run the analysis jobs, skip to Step 3: Running analysis jobs on documents in Amazon S3.

To run analysis jobs, Amazon Comprehend requires access to the Amazon S3 bucket that contains the sample dataset and will contain the jobs' output. IAM roles allow you to control the permissions of AWS services or users. In this step, you create an IAM role for Amazon Comprehend. Then, you create and attach to this role a resource-based policy that grants Amazon Comprehend access to your S3 bucket. By the end of this step, Amazon Comprehend will have the necessary permissions to access your input data, store your output, and run sentiment and entities analysis jobs.

For more information about using IAM with Amazon Comprehend, see How Amazon Comprehend works with IAM.

Topics

Prerequisites

Before you begin, do the following:

- Complete Step 1: Adding documents to Amazon S3.
- Have a code or text editor to save JSON policies and keep track of your Amazon Resource Names (ARNs).

Create an IAM role

To access your Amazon Simple Storage Service (Amazon S3) bucket, Amazon Comprehend needs to assume an AWS Identity and Access Management (IAM) role. The IAM role declares Amazon

Comprehend as a trusted entity. After Amazon Comprehend assumes the role and becomes a trusted entity, you can grant bucket access permissions to Amazon Comprehend. In this step, you create a role that labels Amazon Comprehend as a trusted entity. You can create a role with the AWS CLI or the Amazon Comprehend console. To use the console, skip to Step 3: Running analysis jobs on documents in Amazon S3.

The Amazon Comprehend console lets you select roles where the role name contains 'Comprehend' and the trust policy includes **comprehend.amazonaws.com**. Configure your CLI-created roles to meet these criteria if you want the console to display them.

To create an IAM role for Amazon Comprehend (AWS CLI)

1. Save the following trust policy as a JSON document called `comprehend-trust-policy.json` in a code or text editor on your computer. This trust policy declares Amazon Comprehend as a trusted entity and allows it to assume an IAM role.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "comprehend.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

2. To create the IAM role, run the following AWS CLI command. The command creates an IAM role called `AmazonComprehendServiceRole-access-role` and attaches the trust policy to the role. Replace *`path/`* with your local computer's path to the JSON document.

```
aws iam create-role --role-name AmazonComprehendServiceRole-access-role
--assume-role-policy-document file://path/comprehend-trust-policy.json
```

> **Tip**
>
> If you get an Error parsing parameter message, the path to your JSON trust policy file is probably incorrect. Provide the relative path to the file based on your home directory.

3. Copy the Amazon Resource Name (ARN) and save it in a text editor. The ARN has a format similar to *arn:aws:iam::123456789012:role/AmazonComprehendServiceRole-access-role*. You need this ARN to run Amazon Comprehend analysis jobs.

Attach an IAM policy to the IAM role

To access your Amazon S3 bucket, Amazon Comprehend needs permissions to list, read, and write. To give Amazon Comprehend the required permissions, create and attach an IAM policy to your IAM role. The IAM policy allows Amazon Comprehend to retrieve the input data from your bucket and write analysis results to the bucket. After creating the policy, you attach it to your IAM role.

To create an IAM policy (AWS CLI)

1. Save the following policy locally as a JSON document called `comprehend-access-policy.json`. It grants Amazon Comprehend access to the specified S3 bucket.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Action": [
                "s3:GetObject"
            ],
            "Resource": [
                "arn:aws:s3:::DOC-EXAMPLE-BUCKET/*"
            ],
            "Effect": "Allow"
        },
        {
            "Action": [
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::DOC-EXAMPLE-BUCKET"
```

```
            ],
            "Effect": "Allow"
        },
        {
            "Action": [
                "s3:PutObject"
            ],
            "Resource": [
                "arn:aws:s3:::DOC-EXAMPLE-BUCKET/*"
            ],
            "Effect": "Allow"
        }
    ]
}
```

2. To create the S3 bucket access policy, run the following AWS CLI command. Replace *path/* with your local computer's path to the JSON document.

```
aws iam create-policy --policy-name comprehend-access-policy
--policy-document file://path/comprehend-access-policy.json
```

3. Copy the access policy ARN and save it in a text editor. The ARN has a format similar to *arn:aws:iam::123456789012:policy/comprehend-access-policy*. You need this ARN to attach your access policy to your IAM role.

To attach the IAM policy to your IAM role (AWS CLI)

- Run the following command. Replace *policy-arn* with the access policy ARN that you copied in the previous step.

```
aws iam attach-role-policy --policy-arn policy-arn
--role-name AmazonComprehendServiceRole-access-role
```

You now have an IAM role called `AmazonComprehendServiceRole-access-role` that has a trust policy for Amazon Comprehend and an access policy that grants Amazon Comprehend access to your S3 bucket. You also have the ARN for the IAM role copied to a text editor.

Step 3: Running analysis jobs on documents in Amazon S3

After storing the data in Amazon S3, you can begin running Amazon Comprehend analysis jobs. A *sentiment* analysis job determines the overall mood of a document (positive, negative, neutral, or mixed). An *entities* analysis job extracts the names of real-world objects from a document. These objects include people, places, titles, events, dates, quantities, products, and organizations. In this step, you run two Amazon Comprehend analysis jobs to extract the sentiment and entities from the sample dataset.

Topics

Prerequisites

Before you begin, do the following:

- Complete Step 1: Adding documents to Amazon S3.
- (Optional) If you are using the AWS CLI, complete Step 2: (CLI only) creating an IAM role for Amazon Comprehend and have your IAM role ARN ready.

Analyze sentiment and entities

The first job you run analyzes the sentiment of each customer review in the sample dataset. The second job extracts the entities in each customer review. You can perform Amazon Comprehend analysis jobs either using the Amazon Comprehend console or the AWS CLI.

> **Tip**
>
> Make sure that you are in an AWS Region that supports Amazon Comprehend. For more information, see the Region table in the *Global Infrastructure Guide*.

Analyze sentiments and entities (console)

When using the Amazon Comprehend console, you create one job at a time. You need to repeat the following steps in order to run both a sentiment and an entities analysis job. Note that for the first

job, you create an IAM role, but for the second job, you can reuse the first job's IAM role. You can reuse the IAM role as long as you use the same S3 bucket and folders.

To run sentiment and entities analysis jobs (console)

1. Ensure that you're in the same Region in which you created your Amazon Simple Storage Service (Amazon S3) bucket. If you're in another Region, in the navigation bar, choose the AWS Region where you created your S3 bucket from the **Region selector**.
2. Open the Amazon Comprehend console at https://console.aws.amazon.com/comprehend/
3. Choose **Launch Amazon Comprehend**.
4. In the navigation pane, choose **Analysis jobs**.
5. Choose **Create job**.
6. In the **Job settings** section, do the following:
    a. For **Name**, enter `reviews-sentiment-analysis`.
    b. For **Analysis type**, choose **Sentiment**.
    c. For **Language**, choose **English**.
    d. Leave the **Job encryption** setting as disabled.
7. In the **Input data** section, do the following:
    a. For **Data source**, choose **My documents**.
    b. For **S3 location**, choose **Browse S3** and then choose your bucket from the list of buckets.
    c. In your S3 bucket, for **Objects**, choose your `input` folder.
    d. In the `input` folder, choose the sample dataset `amazon-reviews.csv` and then choose **Choose**.
    e. For **Input format**, choose **One document per line**.
8. In the **Output data** section, do the following:
    a. For **S3 location**, choose **Browse S3** and then choose your bucket from the list of buckets.
    b. In your S3 bucket, for **Objects**, choose the `output` folder and then choose **Choose**.
    c. Leave **Encryption** turned off.
9. In the **Access permissions** section, do the following:
    a. For **IAM role**, choose **Create an IAM role**.
    b. For **Permissions to access**, choose **Input and Output S3 buckets**.

c. For **Name suffix**, enter `comprehend-access-role`. This role provides access to your Amazon S3 bucket.

10. Choose **Create job**.
11. Repeat steps 1-10 to create an entities analysis job. Make the following changes:

    a. In **Job settings**, for **Name**, enter `reviews-entities-analysis`.

    b. In **Job settings**, for **Analysis type**, choose **Entities**.

    c. In **Access permissions**, choose **Use an existing IAM role**. For **Role name**, choose `AmazonComprehendServiceRole-comprehend-access-role` (this is the same role you created for the sentiment job).

Analyze sentiments and entities (AWS CLI)

You use the `start-sentiment-detection-job` and the `start-entities-detection-job` commands to run sentiment and entities analysis jobs. After you run each command, the AWS CLI shows a JSON object with a `JobId` value that allows you to access details about the job, including the output S3 location.

To run sentiment and entities analysis jobs (AWS CLI)

1. Start a sentiment analysis job by running the following command in the AWS CLI. Replace *arn:aws:iam::123456789012:role/comprehend-access-role* with the IAM role ARN that you previously copied to a text editor. If your default AWS CLI Region differs from the Region in which you created your Amazon S3 bucket, include the `--region` parameter and replace *us-east-1* with the Region where your bucket resides.

   ```
   aws comprehend start-sentiment-detection-job
   --input-data-config S3Uri=s3://DOC-EXAMPLE-BUCKET/input/
   --output-data-config S3Uri=s3://DOC-EXAMPLE-BUCKET/output/
   --data-access-role-arn arn:aws:iam::123456789012:role/comprehend-access-role
   --job-name reviews-sentiment-analysis
   --language-code en
   [--region us-east-1]
   ```

2. After you submit the job, copy the `JobId` and save it to a text editor. You will need the `JobId` to find the output files from the analysis job.
3. Start an entities analysis job by running the following command.

```
aws comprehend start-entities-detection-job
--input-data-config S3Uri=s3://DOC-EXAMPLE-BUCKET/input/
--output-data-config S3Uri=s3://DOC-EXAMPLE-BUCKET/output/
--data-access-role-arn arn:aws:iam::123456789012:role/comprehend-access-role
--job-name reviews-entities-analysis
--language-code en
[--region us-east-1]
```

4. After you submit the job, copy the `JobId` and save it to a text editor.
5. Check the status of your jobs. You can view the progress of a job by tracking its `JobId`.

   To track the progress of your sentiment analysis job, run the following command. Replace *sentiment-job-id* with the `JobId` that you copied after running your sentiment analysis.

   ```
   aws comprehend describe-sentiment-detection-job
   --job-id sentiment-job-id
   ```

   To track your entities analysis job, run the following command. Replace *entities-job-id* with the `JobId` that you copied after running your entities analysis.

   ```
   aws comprehend describe-entities-detection-job
   --job-id entities-job-id
   ```

   It takes several minutes for the `JobStatus` to show as `COMPLETED`.

You have completed sentiment and entities analysis jobs. Both of the jobs should be completed before you move on to the next step. It can take several minutes for the jobs to finish.

Step 4: Preparing the Amazon Comprehend output for data visualization

To prepare the results of the sentiment and entities analysis jobs for creating data visualizations, you use AWS Glue and Amazon Athena. In this step, you extract the Amazon Comprehend results files. Then, you create an AWS Glue *crawler* that explores your data and automatically catalogs it in tables in the AWS Glue Data Catalog. After that, you access and transform these tables using Amazon Athena, a serverless and interactive query service. When you have finished this step, your Amazon Comprehend results are clean and ready for visualization.

For a PII entity detection job, the output file is plaintext, not a compressed archive. The output file name is the same as the input file, with `.out` appended at the end. You don't need the step of extracting the output file. Skip to Load the Data into an AWS Glue Data Catalog.

Topics

Prerequisites

Before you begin, complete Step 3: Running analysis jobs on documents in Amazon S3.

Download the Output

The Amazon Comprehend uses Gzip compression to compress output files and save them as a tar archive. The simplest way to extract the output files is to download the `output.tar.gz` archives locally.

In this step, you download the sentiment and entities output archives.

Download the Output Files (Console)

To find the output files for each job, return to the analysis job in the Amazon Comprehend console. The analysis job provides the S3 location for the output, where you can download the output file.

To download the output files (console)

1. In the Amazon Comprehend console, in the navigation pane, return to **Analysis jobs**.
2. Choose your sentiment analysis job `reviews-sentiment-analysis`.
3. Under **Output**, choose the link displayed next to **Output data location**. This redirects you to the `output.tar.gz` archive in your S3 bucket.

4. In the **Overview** tab, choose **Download**.
5. On your computer, rename the archive as `sentiment-output.tar.gz`. Since all of the output files have the same name, this helps you keep track of the sentiment and entities files.
6. Repeat steps 1-4 to find and download the output from your `reviews-entities-analysis` job. On your computer, rename the archive as `entities-output.tar.gz`.

Download the output files (AWS CLI)

To find the output files for each job, use the `JobId` from the analysis job to find the output's S3 location. Then, use the `cp` command to download the output file to your computer.

To download the output files (AWS CLI)

1. To list details about your sentiment analysis job, run the following command. Replace *`sentiment-job-id`* with the sentiment `JobId` that you saved.

   ```
   aws comprehend describe-sentiment-detection-job --job-id sentiment-job-id
   ```

   If you lost track of your `JobId`, you can run the following command to list all of your sentiment jobs and filter for your job by name.

   ```
   aws comprehend list-sentiment-detection-jobs
   --filter JobName="reviews-sentiment-analysis"
   ```

2. In the `OutputDataConfig` object, find the `S3Uri` value. The `S3Uri` value should be similar to the following format: *`s3://DOC-EXAMPLE-BUCKET/.../output/output.tar.gz`*. Copy this value to a text editor.
3. To download the sentiment output archive to your local directory, run the following command. Replace the S3 bucket path with the `S3Uri` you copied in the previous step. Replace *`path/`* with the folder path to your local directory. The name `sentiment-output.tar.gz` replaces the original archive name to help you keep track of the sentiment and entities files.

   ```
   aws s3 cp s3://DOC-EXAMPLE-BUCKET/.../output/output.tar.gz
   path/sentiment-output.tar.gz
   ```

4. To list details about your entities analysis job, run the following command.

   ```
   aws comprehend describe-entities-detection-job
   ```

```
--job-id entities-job-id
```

If you don't know your `JobId`, run the following command to list all of your entities jobs and filter for your job by name.

```
aws comprehend list-entities-detection-jobs
--filter JobName="reviews-entities-analysis"
```

5. From the `OutputDataConfig` object in your entities job description, copy the `S3Uri` value.
6. To download the entities output archive to your local directory, run the following command. Replace the S3 bucket path with the `S3Uri` you copied in the previous step. Replace *path/* with the folder path to your local directory. The name `entities-output.tar.gz` replaces the original archive name.

```
aws s3 cp s3://DOC-EXAMPLE-BUCKET/.../output/output.tar.gz
path/entities-output.tar.gz
```

Extract the output files

Before you can access the Amazon Comprehend results, unpack the sentiment and entities archives. You can use either your local file system or a terminal to unpack the archives.

Extract the output files (GUI file system)

If you use macOS, double-click the archive in your GUI file system to extract the output file from the archive.

If you use Windows, you can use a third-party tool such as 7-Zip to extract the output files in your GUI file system. In Windows, you must perform two steps to access the output file in the archive. First decompress the archive, and then extract the archive.

Rename the sentiment file as `sentiment-output` and the entities file as `entities-output` to distinguish between the output files.

Extract the output files (terminal)

If you use Linux or macOS, you can use your standard terminal. If you use Windows, you must have access to a Unix-style environment, such as Cygwin, to run tar commands.

To extract the sentiment output file from the sentiment archive, run the following command in your local terminal.

```
tar -xvf sentiment-output.tar.gz --transform 's,^,sentiment-,'
```

Note that the `--transform` parameter adds the prefix `sentiment-` to the output file inside of the archive, renaming the file as `sentiment-output`. This allows you to distinguish between the sentiment and entities output files and prevent overwriting.

To extract the entities output file from the entities archive, run the following command in your local terminal.

```
tar -xvf entities-output.tar.gz --transform 's,^,entities-,'
```

The `--transform` parameter adds the prefix `entities-` to the output file name.

> **Tip**
>
> To save storage costs in Amazon S3, you can compress the files again with Gzip before uploading them. It's important to decompress and unpack the original archives because AWS Glue can't automatically read data from a tar archive. However, AWS Glue can read from files in Gzip format.

Upload the extracted files

After extracting the files, upload them to your bucket. You must store the sentiment and entities output files in separate folders in order for AWS Glue to read the data properly. In your bucket, create a folder for the extracted sentiment results and a second folder for the extracted entities results. You can create folders either with the Amazon S3 console or the AWS CLI.

Upload the extracted files to Amazon S3 (console)

In your S3 bucket, create one folder for the extracted sentiment results file and one folder for the entities results file. Then, upload the extracted results files to their respective folders.

To upload the extracted files to Amazon S3 (console)

1. Open the Amazon S3 console at https://console.aws.amazon.com/s3/.

2. In **Buckets**, choose your bucket and then choose **Create folder**.
3. For the new folder name, enter `sentiment-results` and choose **Save**. This folder will contain the extracted sentiment output file.
4. In your bucket's **Overview** tab, from the list of bucket contents, choose the new folder `sentiment-results`. Choose **Upload**.
5. Choose **Add files**, choose the `sentiment-output` file from your local computer, and then choose **Next**.
6. Leave the options for **Manage users**, **Access for other AWS account**, and **Manage public permissions** as the defaults. Choose **Next**.
7. For **Storage class**, choose **Standard**. Leave the options for **Encryption**, **Metadata**, and **Tag** as the defaults. Choose **Next**.
8. Review the upload options and then choose **Upload**.
9. Repeat steps 1-8 to create a folder called `entities-results`, and upload the `entities-output` file to it.

Upload the extracted files to Amazon S3 (AWS CLI)

You can create a folder in your S3 bucket while uploading a file with the `cp` command.

To upload the extracted files to Amazon S3 (AWS CLI)

1. Create a sentiment folder and upload your sentiment file to it by running the following command. Replace *`path/`* with the local path to your extracted sentiment output file.

```
aws s3 cp path/sentiment-output s3://DOC-EXAMPLE-BUCKET/sentiment-results/
```

2. Create an entities output folder and upload your entities file to it by running the following command. Replace *`path/`* with the local path to your extracted entities output file.

```
aws s3 cp path/entities-output s3://DOC-EXAMPLE-BUCKET/entities-results/
```

Load the data into an AWS Glue Data Catalog

To get the results into a database, you can use an AWS Glue *crawler*. An AWS Glue *crawler* scans files and discovers the schema of the data. It then arranges the data in tables in an AWS Glue Data

Catalog (a serverless database). You can create a crawler with the AWS Glue console or the AWS CLI.

Load the data into an AWS Glue Data Catalog (console)

Create an AWS Glue crawler that scans your `sentiment-results` and `entities-results` folders separately. A new IAM role for AWS Glue gives the crawler permission to access your S3 bucket. You create this IAM role while setting up the crawler.

To load the data into an AWS Glue Data Catalog (console)

1. Ensure that you're in a Region which supports AWS Glue. If you're in another Region, in the navigation bar, choose a supported Region from the **Region selector**. For a list of Regions that support AWS Glue, see the Region Table in the *Global Infrastructure Guide*.
2. Open the AWS Glue console at https://console.aws.amazon.com/glue/.
3. In the navigation pane, choose **Crawlers** and then choose **Add crawler**.
4. For **Crawler name**, enter `comprehend-analysis-crawler` and then choose **Next**.
5. For **Crawler source type**, choose **Data stores** and then choose **Next**.
6. For **Add a data store**, do the following:
    a. For **Choose a data store**, choose **S3**.
    b. Leave **Connection** blank.
    c. For **Crawl data in**, choose **Specified path in my account**.
    d. For **Include path**, enter the full S3 path of the sentiment output folder: `s3://DOC-EXAMPLE-BUCKET/sentiment-results`.
    e. Choose **Next**.
7. For **Add another data store**, choose **Yes** and then choose **Next**. Repeat Step 6, but enter the full S3 path of the entities output folder: `s3://DOC-EXAMPLE-BUCKET/entities-results`.
8. For **Add another data store**, choose **No** and then choose **Next**.
9. For **Choose an IAM role**, do the following:
    a. Choose **Create an IAM role**.
    b. For **IAM role**, enter `glue-access-role` and then choose **Next**.
10. For **Create a schedule for this crawler**, choose **Run on demand** and choose **Next**.
11. For **Configure the crawler's output**, do the following:

a. For **Database**, choose **Add database**.

b. For **Database name**, enter `comprehend-results`. This database will store your Amazon Comprehend output tables.

c. Leave the other options on their default settings and choose **Next**.

12. Review the crawler information and then choose **Finish**.

13. In the Glue console, in **Crawlers**, choose `comprehend-analysis-crawler` and choose **Run crawler**. It can take a few minutes for the crawler to finish.

Load the data into an AWS Glue Data Catalog (AWS CLI)

Create an IAM role for AWS Glue that provides permission to access your S3 bucket. Then, create a database in the AWS Glue Data Catalog. Finally, create and run a crawler that loads your data into tables in the database.

To load the data into an AWS Glue Data Catalog (AWS CLI)

1. To create an IAM role for AWS Glue, do the following:

   a. Save the following trust policy as a JSON document called `glue-trust-policy.json` on your computer.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "glue.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

   b. To create an IAM role, run the following command. Replace *path/* with your local computer's path to the JSON document.

```
aws iam create-role --role-name glue-access-role
```

```
--assume-role-policy-document file://path/glue-trust-policy.json
```

c. When the AWS CLI lists the Amazon Resource Number (ARN) for the new role, copy and save it to a text editor.

d. Save the following IAM policy as a JSON document called `glue-access-policy.json` on your computer. The policy grants AWS Glue permission to crawl your results folders.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:PutObject"
            ],
            "Resource": [
                "arn:aws:s3:::DOC-EXAMPLE-BUCKET/sentiment-results*",
                "arn:aws:s3:::DOC-EXAMPLE-BUCKET/entities-results*"
            ]
        }
    ]
}
```

e. To create the IAM policy, run the following command. Replace *path/* with your local computer's path to the JSON document.

```
aws iam create-policy --policy-name glue-access-policy
--policy-document file://path/glue-access-policy.json
```

f. When the AWS CLI lists the access policy's ARN, copy and save it to a text editor.

g. Attach the new policy to the IAM role by running the following command. Replace *policy-arn* with the IAM policy ARN you copied in the previous step.

```
aws iam attach-role-policy --policy-arn policy-arn
--role-name glue-access-role
```

h. Attach the AWS managed policy `AWSGlueServiceRole` to your IAM role by running the following command.

```
aws iam attach-role-policy --policy-arn
```

```
arn:aws:iam::aws:policy/service-role/AWSGlueServiceRole
--role-name glue-access-role
```

2. Create an AWS Glue database by running the following command.

```
aws glue create-database
--database-input Name="comprehend-results"
```

3. Create a new AWS Glue crawler by running the following command. Replace *glue-iam-role-arn* with the ARN of your AWS Glue IAM role.

```
aws glue create-crawler
--name comprehend-analysis-crawler
--role glue-iam-role-arn
--targets S3Targets=[
{Path="s3://DOC-EXAMPLE-BUCKET/sentiment-results"},
{Path="s3://DOC-EXAMPLE-BUCKET/entities-results"}]
--database-name comprehend-results
```

4. Start the crawler by running the following command.

```
aws glue start-crawler --name comprehend-analysis-crawler
```

It can take a few minutes for the crawler to finish.

Prepare the data for analysis

Now you have a database populated with the Amazon Comprehend results. However, the results are nested. To unnest them, you run a few SQL statements in Amazon Athena. Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage and it has a pay-per-query pricing model. In this step, you create new tables of cleaned data that you can use for analysis and visualization. You use the Athena console to prepare the data.

To prepare the data

1. Open the Athena console at https://console.aws.amazon.com/athena/.
2. In the query editor, choose **Settings**, then choose **Manage**.

3. For **Location of query results**, enter `s3://DOC-EXAMPLE-BUCKET/query-results/`. This creates a new folder called `query-results` in your bucket that stores the output of the Amazon Athena queries you run. Choose **Save**.
4. In the query editor, choose **Editor**.
5. For **Database**, choose the AWS Glue database `comprehend-results` that you created.
6. In the **Tables** section, you should have two tables called `sentiment_results` and `entities_results`. Preview the tables to make sure that the crawler loaded the data. In each table's options (the three dots next to the table name), choose **Preview table**. A short query runs automatically. Check the **Results** pane to ensure that the tables contain data.

   > **Tip**
   >
   > If the tables don't have any data, try checking the folders in your S3 bucket. Make sure that there is one folder for entities results and one folder for sentiment results. Then, try running a new AWS Glue crawler.

7. To unnest the `sentiment_results` table, enter the following query in the **Query editor** and choose **Run**.

```
CREATE TABLE sentiment_results_final AS
SELECT file, line, sentiment,
sentimentscore.mixed AS mixed,
sentimentscore.negative AS negative,
sentimentscore.neutral AS neutral,
sentimentscore.positive AS positive
FROM sentiment_results
```

8. To begin unnesting the entities table, enter the following query in the **Query editor** and choose **Run**.

```
CREATE TABLE entities_results_1 AS
SELECT file, line, nested FROM entities_results
CROSS JOIN UNNEST(entities) as t(nested)
```

9. To finish unnesting the entities table, enter the following query in the **Query editor** and choose **Run query**.

```
CREATE TABLE entities_results_final AS
SELECT file, line,
```

```
nested.beginoffset AS beginoffset,
nested.endoffset AS endoffset,
nested.score AS score,
nested.text AS entity,
nested.type AS category
FROM entities_results_1
```

Your `sentiment_results_final` table should look like the following, with columns named **file**, **line**, **sentiment**, **mixed**, **negative**, **neutral**, and **positive**. The table should have one value per cell. The **sentiment** column describes the most likely overall sentiment of a particular review. The **mixed**, **negative**, **neutral**, and **positive** columns give scores for each type of sentiment.

Results

	file	line	sentiment	mixed	negative	neutral	positive
1	amazon-reviews.csv	6	MIXED	0.9862896203994751	0.0015502438182011247	1.6660270921420306E-4	0.0119935879483
2	amazon-reviews.csv	8	POSITIVE	0.0012987082591280341	0.01186690479516983	0.174478679895401	0.8123556375503
3	amazon-reviews.csv	11	POSITIVE	6.5368581090297084E-6	0.0013866390800103545	0.007405391428619623	0.9912014007568
4	amazon-reviews.csv	13	POSITIVE	4.7155481297522783E-4	0.24615342915058136	0.017713148146867752	0.7356618046760
5	amazon-reviews.csv	14	POSITIVE	1.5821871784282848E-5	0.06828905642032623	0.014075091108679771	0.9176200628280
6	amazon-reviews.csv	16	MIXED	0.9864791035652161	8.548551704734564E-4	1.0789262159960344E-4	0.0125581491738
7	amazon-reviews.csv	20	NEGATIVE	1.1621621524682269E-4	0.9815887212753296	0.004688907880336046	0.0136061981320
8	amazon-reviews.csv	21	POSITIVE	4.663573781726882E-5	0.009533549658954144	0.0015825830632820725	0.9888372421264
9	amazon-reviews.csv	23	POSITIVE	1.7699007003102452E-4	0.40269607305526733	0.0018250439316034317	0.5953019261360
10	amazon-reviews.csv	25	POSITIVE	1.8434448065818287E-6	1.15832663141191E-4	0.0010993879986926913	0.9987829327583

Your `entities_results_final` table should look like the following, with columns named **file**, **line**, **beginoffset**, **endoffset**, **score**, **entity**, and **category**. The table should have one value per cell. The **score** column indicates Amazon Comprehend's confidence in the **entity** it detected. The **category** indicates what kind of entity Comprehend detected.

Results

	file	line	beginoffset	endoffset	score	entity	category
1	amazon-reviews.csv	0	15	22	0.9885989378545348	English	OTHER
2	amazon-reviews.csv	2	24	28	0.9699371997593782	2 me	QUANTITY
3	amazon-reviews.csv	2	94	95	0.6523066984191679	2	QUANTITY
4	amazon-reviews.csv	2	125	126	0.713791396412543	2	QUANTITY
5	amazon-reviews.csv	4	30	36	0.9957169942979278	kindle	COMMERCIAL_ITEM
6	amazon-reviews.csv	5	1	10	0.9979111763962706	Hawthorne	PERSON
7	amazon-reviews.csv	5	135	142	0.5065408081314243	Puritan	OTHER
8	amazon-reviews.csv	5	143	148	0.7702269458801602	Salem	LOCATION
9	amazon-reviews.csv	5	211	229	0.999675563687763	The Scarlet Letter	TITLE
10	amazon-reviews.csv	5	233	236	0.8944631322676461	one	QUANTITY

Now that you have the Amazon Comprehend results loaded into tables, you can visualize and extract meaningful insights from the data.

Step 5: Visualizing Amazon Comprehend output in Amazon QuickSight

After storing the Amazon Comprehend results in tables, you can connect to and visualize the data with Amazon QuickSight. Amazon QuickSight is an AWS managed business intelligence (BI) tool for visualizing data. Amazon QuickSight makes it easy to connect to your data source and create powerful visuals. In this step, you connect Amazon QuickSight to your data, create visualizations that extract insights from the data, and publish a dashboard of visualizations.

Topics

Prerequisites

Before you begin, complete Step 4: Preparing the Amazon Comprehend output for data visualization.

Give Amazon QuickSight access

To import the data, Amazon QuickSight requires access to your Amazon Simple Storage Service (Amazon S3) bucket and Amazon Athena tables. To give Amazon QuickSight access to your data, you must be signed in as a QuickSight administrator and have access to edit the resource permissions. If you are unable to complete the following steps, review the IAM prerequisites from the overview page Tutorial: Analyzing insights from customer reviews with Amazon Comprehend.

To give Amazon QuickSight access to your data

1. Open the Amazon QuickSight console.
2. If this is the first time you have used Amazon QuickSight, the console prompts you to create a new administrator user by providing an email address. For **Email address**, enter the same email address as your AWS account. Choose **Continue**.
3. After signing in, choose your profile name in the navigation bar and choose **Manage QuickSight**. You must be signed in as an administrator to view the **Manage QuickSight** option.
4. Choose **Security and permissions**.
5. For **QuickSight access to AWS services**, choose **Add or remove**.
6. Choose **Amazon S3**.
7. From **Select Amazon S3 buckets**, choose your S3 bucket for both **S3 Bucket** and **Write permissions for Athena Workgroup**.
8. Choose **Finish**.
9. Choose **Update**.

Import the datasets

Before creating visualizations, you must add the sentiment and entities datasets to Amazon QuickSight. You do this with the Amazon QuickSight console. You import your unnested sentiment and unnested entities tables from Amazon Athena.

To import your datasets

1. Open the Amazon QuickSight console.
2. In the navigation bar, in **Datasets**, choose **New dataset**.
3. For **Create a Data Set**, choose **Athena**.
4. For **Data source name**, enter `reviews-sentiment-analysis` and choose **Create data source**.
5. For **Database**, choose the database `comprehend-results`.
6. For **Tables**, choose the sentiment table `sentiment_results_final` and then choose **Select**.
7. Choose **Import to SPICE for quicker analytics** and choose **Visualize**. SPICE is QuickSight's in-memory calculation engine that provides faster analyses than direct querying when creating visualizations.
8. Return to the Amazon QuickSight console and choose **Datasets**. Repeat steps 1-7 to create an entities dataset, but make the following changes:

    a. For **Data source name**, enter `reviews-entities-analysis`.

    b. For **Tables**, choose the entities table `entities_results_final`.

Create a sentiment visualization

Now that you can access your data in Amazon QuickSight, you can begin creating visualizations. You create a pie chart with the Amazon Comprehend sentiment data. The pie chart shows what proportion of the reviews are positive, neutral, mixed, and negative.

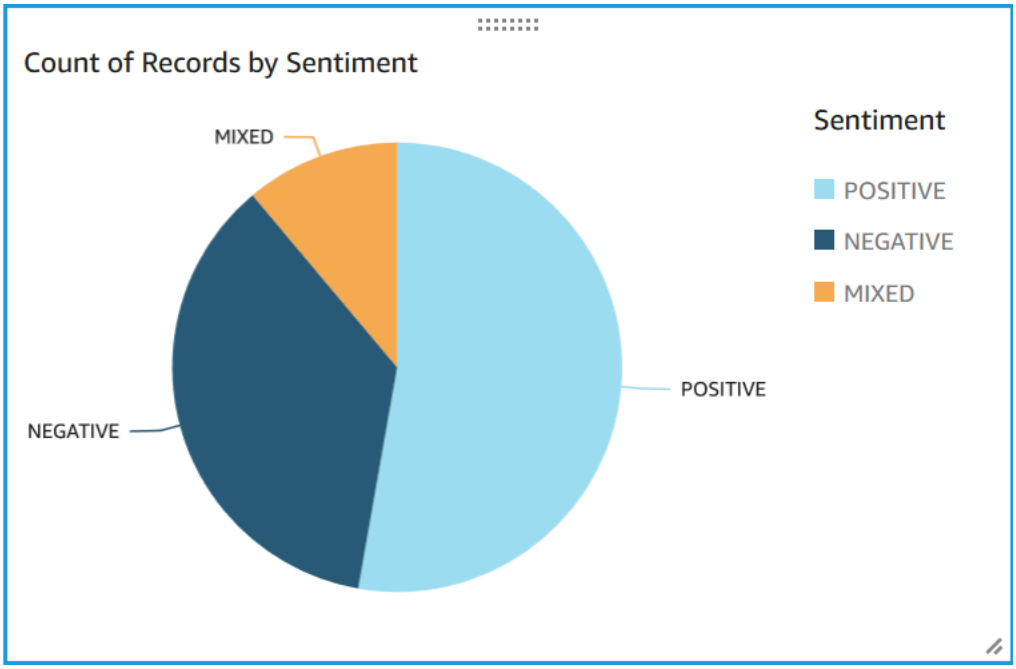
To visualize sentiment data

1. In the Amazon QuickSight console, choose **Analyses** and then choose **New analysis**.
2. From **Your Data Sets**, choose the sentiment dataset `sentiment_results_final` and then choose **Create analysis**.
3. In the visual editor, in **Fields list**, choose **sentiment**.

> **Note**
>
> The values in the **Fields list** depend on the column names you used to create the tables in Amazon Athena. If you changed the provided column names in the SQL queries, the **Fields list** names will be different than the names used in these visualization examples.

4. For **Visual types**, choose **Pie chart**.

A pie chart similar to the following with positive, neutral, mixed, and negative sections is displayed. To see the count and percentage of a section, hover over it.



Create an entities visualization

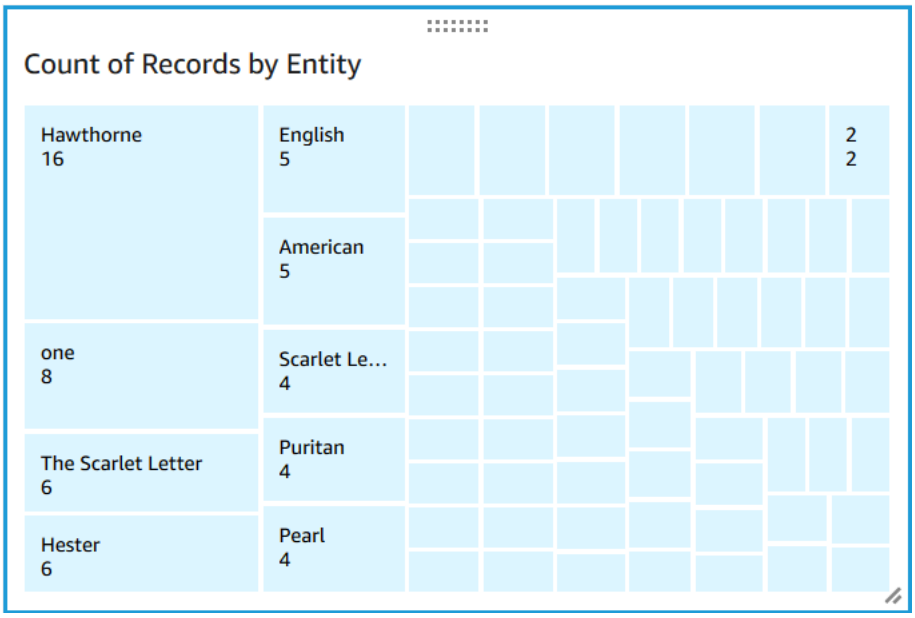
Now create a second visualization with the entities dataset. You create a tree map of the distinct entities in the data. Each block in the tree map represents an entity, and the size of the block correlates to the number of times that the entity appears in the dataset.

To visualize entities data

1. In the **Visualize** control pane, next to **Data set**, choose the **Add, edit, replace, and remove data sets** icon.
2. Choose **Add data set**.

3. For **Choose data set to add**, choose your entities dataset `entities_results_final` from the list of datasets and choose **Select**.
4. In the **Visualize** control pane, choose the **Data set** drop down menu and choose the entities dataset `entities_results_final`.
5. In **Fields list**, choose **entity**.
6. For **Visual types**, choose **Tree map**.

A tree map similar to the following is displayed next to your pie chart. To see the count of a specific entity, hover over a block.



Publish a dashboard

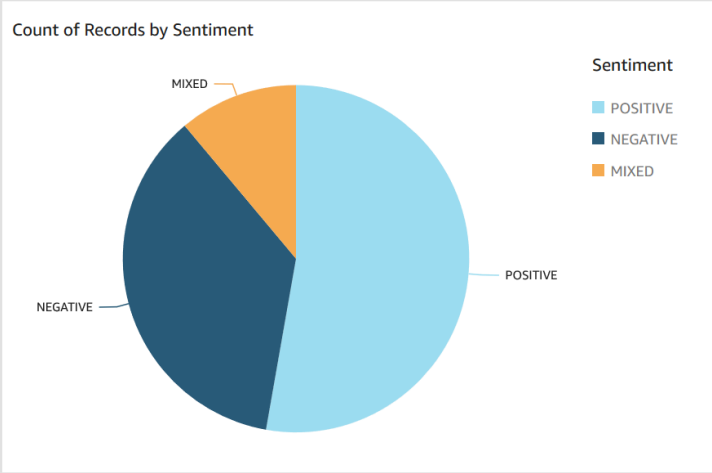
After creating the visualizations, you can publish them as a dashboard. You can perform various tasks with a dashboard, such as sharing it with users in your AWS account, saving it as a PDF, or emailing it as a report (limited to the Enterprise edition of Amazon QuickSight). In this step, you publish the visualizations as a dashboard in your account.

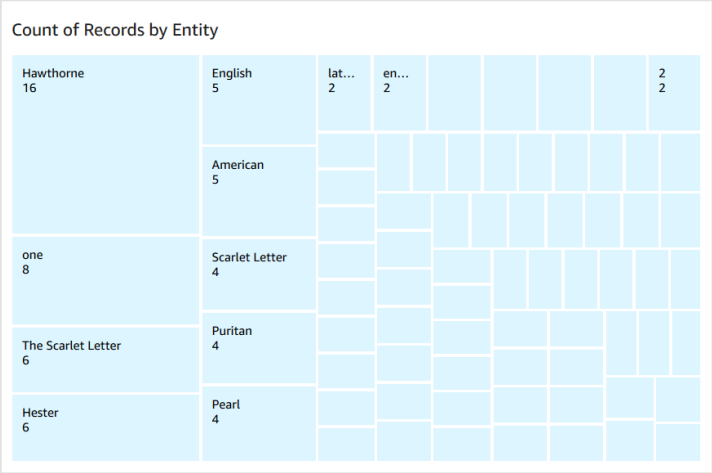
To publish your dashboard

1. In the navigation bar, choose **Share**.
2. Choose **Publish dashboard**.
3. Choose **Publish new dashboard as** and enter the name `comprehend-analysis-reviews` for the dashboard.

4. Choose **Publish dashboard**.
5. Close the **Share dashboard with users** pane by choosing the close button in the upper-right corner.
6. In the Amazon QuickSight console, in the navigation pane, choose **Dashboards**. A thumbnail of your new dashboard `comprehend-analysis-reviews` should appear under **Dashboards**. Choose the dashboard to view it.

You now have a dashboard with sentiment and entities visualizations that looks similar to the following example.





> **Tip**
>
> If you want to edit the visualizations in your dashboard, return to **Analyses** and edit the visualization that you want to update. Then, publish the dashboard again either as a new dashboard or as a replacement of the existing dashboard.

Clean up

After completing this tutorial, you might want to clean up any AWS resources you no longer want to use. Active AWS resources can continue to incur charges in your account.

The following actions can help prevent incurring ongoing charges:

- Cancel your Amazon QuickSight subscription. Amazon QuickSight is a monthly subscription service. To cancel your subscription, see Canceling your subscription in the *Amazon QuickSight User Guide*.

- Delete your Amazon S3 bucket. Amazon S3 charges you for storage. To clean up your Amazon S3 resources, delete your bucket. For information about deleting a bucket, see How do I delete an S3 Bucket? in the *Amazon Simple Storage Service User Guide*. Make sure that you save all of your important files before deleting your bucket.
- Clear your AWS Glue Data Catalog. The AWS Glue Data Catalog charges you monthly for storage. You can delete your databases to prevent incurring ongoing charges. For information about managing your AWS Glue Data Catalog databases, see Working with databases on the AWS Glue console in the *AWS Glue Developer Guide*. Make sure that you export your data before clearing any databases or tables.

Using Amazon S3 object Lambda access points for personally identifiable information (PII)

Use Amazon S3 Object Lambda Access Points for personally identifiable information (PII) to configure how documents are retrieved from your Amazon S3 bucket. You can control access to documents that contain PII and redact PII from documents. For more information on how Amazon Comprehend can detect PII in your documents, see Detecting PII entities. Amazon S3 Object Lambda Access Points use AWS Lambda functions to automatically transform the output of a standard Amazon S3 GET request. For more information see, Transforming objects with S3 object Lambda in the *Amazon Simple Storage Service User Guide*.

When you create an Amazon S3 Object Lambda Access Point for PII, documents are processed using Amazon Comprehend Lambda functions to control access of documents that contain PII and redact PII from documents.

When you create an Amazon S3 Object Lambda Access Point for PII, documents are processed using the following Amazon Comprehend Lambda functions:

- `ComprehendPiiAccessControlS3ObjectLambda` - Controls access to documents with PII stored in your S3 bucket. For more information about this Lambda function, sign in to the AWS Management Console to view the ComprehendPiiAccessControlS3ObjectLambda function in the AWS Serverless Application Repository.
- `ComprehendPiiRedactionS3ObjectLambda` - Redacts PII from documents in your Amazon S3 bucket. For more information about this Lambda function, sign in to the AWS Management Console to view the ComprehendPiiRedactionS3ObjectLambda function in the AWS Serverless Application Repository.

For information about how to deploy serverless applications from the AWS Serverless Application Repository, see Deploying applications in the *AWS Serverless Application Repository Developer Guide*.

Topics

Controlling access to documents with personally identifiable information (PII)

You can use an Amazon S3 Object Lambda Access Point to control access to documents with personally identifiable information (PII).

To ensure that only authorized users have access to documents that contain PII stored in your Amazon S3 bucket, you use the `ComprehendPiiAccessControlS3ObjectLambda` function. This Lambda function uses the ContainsPiiEntities operation when processing a standard Amazon S3 GET request on document objects.

For example, if you have documents in your S3 bucket that include PII such as credit card numbers or bank account information, you can configure the `ComprehendPiiAccessControlS3ObjectLambda` function to detect these PII entity types and restrict access to unauthorized users. For more information about supported PII entity types, see PII universal entity types.

For more information about this Lambda function, sign in to the AWS Management Console to view the ComprehendPiiAccessControlS3ObjectLambda function in the AWS Serverless Application Repository.

Creating an Amazon S3 object Lambda access point to control access to documents

The following example creates an Amazon S3 Object Lambda Access Point to control access to documents that contain social security numbers.

Creating an Amazon S3 object Lambda access point using the AWS Command Line Interface

Create an Amazon S3 Object Lambda Access Point configuration and save the configuration in a file called **config.json**.

```
{
    "SupportingAccessPoint": "s3-default-access-point-name-arn",
    "TransformationConfigurations": [
        {
            "Actions": [
                "s3:GetObject"
            ],
            "ContentTransformation": {
                "AwsLambda": {
                    "FunctionArn": "comprehend-pii-access-control-s3-object-lambda-
arn",
                    "FunctionPayload": "{\"pii_entities_types\": \"SSN\"}"
                }
            }
        }
    ]
}
```

The following example creates an Amazon S3 Object Lambda Access Point based on the configuration defined in the `config.json` file.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws s3control create-banner-access-point \
    --region region \
    --account-id account-id \
    --name s3-object-lambda-access-point \
    --configuration file://config.json
```

Invoking an Amazon S3 object Lambda access point to control access to documents

The following example invokes an Amazon S3 Object Lambda Access Point to control access to documents.

Invoking an Amazon S3 object Lambda access point using the AWS Command Line Interface

The following example invokes an Amazon S3 Object Lambda Access Point using the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws s3api get-object \
    --region region \
    --bucket s3-object-lambda-access-point-name-arn \
    --key object-prefix-key output-file-name
```

Redacting personally identifiable information (PII) from documents

You can use an Amazon S3 Object Lambda Access Point to redact personally identifiable information (PII) from documents.

To redact PII entity types from documents stored in an S3 bucket, you use the `ComprehendPiiRedactionS3ObjectLambda` function. This Lambda function uses the ContainsPiiEntities and DetectPiiEntities operations when processing a standard Amazon S3 GET request on document objects.

For example, if documents in your S3 bucket include PII such as credit card numbers or bank account information, you can configure the `ComprehendPiiRedactionS3ObjectLambda` function to detect PII and then return a copy of these documents in which PII entity types are redacted. For more information about supported PII entity types, see PII universal entity types.

For more information about this Lambda function, sign in to the AWS Management Console to view the ComprehendPiiRedactionS3ObjectLambda function in the AWS Serverless Application Repository.

Creating an Amazon S3 object Lambda access point to redact PII from documents

The following example creates an Amazon S3 Object Lambda Access Point to redeact credit card numbers from documents.

Creating an Amazon S3 object Lambda access point using the AWS Command Line Interface

Create an Amazon S3 Object Lambda Access Point configuration and save the configuration in a file called `config.json`.

```
{
    "SupportingAccessPoint": "s3-default-access-point-name-arn",
    "TransformationConfigurations": [
        {
            "Actions": [
```

```
                "s3:GetObject"
            ],
            "ContentTransformation": {
                "AwsLambda": {
                    "FunctionArn": "comprehend-pii-redaction-s3-object-lambda-arn",
                    "FunctionPayload": "{\"pii_entities_types\": \"CREDIT_DEBIT_NUMBER
\"}"
                }
            }
        }
    ]
}
```

The following example demonstrates creating an Amazon S3 Object Lambda Access Point based on the configuration defined in the `config.json`

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws s3control create-access-point-for-object-lambda \
    --region region \
    --account-id account-id \
    --name s3-object-lambda-access-point \
    --configuration file://config.json
```

Invoking an Amazon S3 object Lambda access point to redact PII from documents

The following examples invoke an Amazon S3 Object Lambda Access Point to redact PII from documents.

Invoking an Amazon S3 object Lambda access point using the AWS Command Line Interface

The following example invokes an Amazon S3 Object Lambda Access Point using the AWS CLI.

The example is formatted for Unix, Linux, and macOS. For Windows, replace the backslash (\) Unix continuation character at the end of each line with a caret (^).

```
aws s3api get-object \
    --region region \
    --bucket s3-object-lambda-access-point-name-arn \
    --key object-prefix-key output-file-name
```

Solution: Analyzing text with Amazon Comprehend and OpenSearch

AWS provides a reference implementation of text analysis using Amazon Comprehend and the OpenSearch service. Amazon Comprehend provides text analysis and OpenSearch provides document indexing, searching, and visualization.

For more information, see Analyzing text with OpenSearch and Amazon Comprehend.

API reference

The API reference is now a separate document. For more information, see Amazon Comprehend API Reference.

Document history for Amazon Comprehend

The following table describes the documentation for this release of Amazon Comprehend.

Change	Description	Date
Custom classifier training with native documents	Amazon Comprehend now supports custom classifier training with native documents. For more information, see Training classification models in Amazon Comprehend.	April 19, 2023
Flywheels for managing custom models	Amazon Comprehend now supports flywheels to help you manage the training and tracking of model versions for custom models. For more information, see Flywheels in Amazon Comprehend.	February 28, 2023
Updated IAM security topics	Updated the IAM security topics to include federated identities. For more informati on, see Identity and Access Management for Amazon Comprehend.	December 22, 2022
One-step processing for inference with custom models	Amazon Comprehend now automatically performs the text extraction for image, PDF, or Word input documents prior to running custom classification or custom entity recognition. For more information, see	December 1, 2022

	Document processing in Amazon Comprehend.	
Synchronous APIs for targeted sentiment	Amazon Comprehend now supports synchronous APIs and console real-time analysis for targeted sentiment . Targeted sentiment determines the sentiment associated with specific entities in a document. For more information, see Targeted sentiment in Amazon Comprehend.	September 21, 2022
Lower minimum annotations for training recognizers	Amazon Comprehend has reduced the minimum requirements for training a recognizer with plaintext CSV annotation files. You can now build a custom entity recognition model with as few as three annotated documents and at least 25 annotations per entity type. For more information, see Preparing the training data.	August 3, 2022
Increased input document size for real-time APIs	Amazon Comprehend now supports up to a 100KB input document for most real-time APIs. For more information, see Guidelines and quotas.	July 18, 2022

Additional PII entity types	Additional PII entity types now detected by Amazon Comprehend. For more information, see Detecting PII entities in Amazon Comprehend.	May 20, 2022
Table of Contents restructure	Restructured the Amazon Comprehend Developer Guide table of contents for easier navigation. For more information, see What is Amazon Comprehend.	April 7, 2022
Targeted sentiment	Amazon Comprehend now supports targeted sentiment analysis, which determine s the sentiment associate d with specific entities in a document. For more informati on, see Targeted sentiment in Amazon Comprehend.	March 9, 2022
New feature	Amazon Comprehend now allows you to analyze images for custom entitiy recogniti on. For more information, see Detecting custom entities in Amazon Comprehend.	February 28, 2022
New feature	You can now copy trained custom models between AWS accounts. For more informati on, see Copying custom models between accounts in Amazon Comprehend.	February 2, 2022

New feature	You can now use AWS Trusted Advisor to view recommendations that can help you optimize the cost and security of your Amazon Comprehend endpoints. For more information, see Using Trusted Advisor with Amazon Comprehend.	September 29, 2021
New feature	Amazon Comprehend has launched a suite of features for Comprehend Custom which enable continuous model improvements by giving you the ability to create new model versions, continuously test on specific test sets, and perform live migration to new model endpoints.	September 21, 2021
New feature	Amazon Comprehend now allows you to analyze PDF and Word documents for custom entitiy recognition. With PDF and Word formats, you can extract information from documents containing headers, lists and tables.	September 14, 2021

New feature	Amazon Comprehend has launched a new endpoints overview feature which provides you a global view of your endpoints. From the endpoints overview page, you can view all of your endpoints in one place to understand your endpoint usage versus your actual resource usage.	August 24, 2021
New feature	Amazon Comprehend Medical now allows you to establish a private connection with your Virtual Private Cloud (VPC) by creating an interface VPC endpoint. For more informati on, see VPC endpoints (PrivateLink).	June 13, 2021
Language expansion	Amazon Comprehend has added four additional languages for the dominant language feature: Hausa (ha), Lao (lo), Maltese (mt), and Oromo (om). For more information, see Supported languages in Amazon Comprehend.	May 10, 2021
New feature	With Amazon Comprehen d, you can now encrypt custom models using a customer managed key (CMK). For more information, see KMS encryption in Amazon Comprehend.	March 31, 2021

New feature	You can now use Amazon S3 Object Lambda Access Points to configure how documents that contain personally identifiable information (PII) are retrieved from your Amazon S3 bucket. You can control access of documents that contain PII and redact PII from documents. For more information, see Using Amazon S3 object Lambda access points for personally identifiable information (PII).	March 18, 2021
New feature	You can now label a document with personall y identifiable information (PII). Amazon Comprehend can analyze your document for the presence of PII and return the labels of identifie d PII entity types such as name, address, bank account number, or phone number. For more information, see Label document with PII.	March 11, 2021

New feature	With Amazon Comprehend, you can now detect events in a set of documents. When you create an asynchronous events detection job, Amazon Comprehend can detect supported types of financial events. For more information, see Detect events.	November 24, 2020
New feature	Amazon Comprehend now allows you to use auto scaling for custom entity recognizer endpoints. With auto scaling, you can automatically set endpoint provisioning to fit your capacity needs. For more information, see Auto scaling with endpoints.	September 28, 2020
New feature	To train custom classifiers or entity recognizers, you can now provide augmented manifest files, which are labeled datasets that are produced by Amazon SageMaker Ground Truth. For more information about these files, and for examples, see Multi-class mode, Multi-label mode, and Annotations.	September 22, 2020

New tutorial	Amazon Comprehend now has a tutorial that walks you through a multi-ser vice workflow of analyzing customer reviews and visualizi ng the analysis results. For more information, see Tutorial: Analyzing insights from reviews.	September 17, 2020
New feature	With Amazon Comprehend, you can now detect entities in your text that contain personally identifiable information (PII), such as addresses, bank account numbers, or phone numbers. Amazon Comprehend can provide the location of each PII entity in your text, or it can provide a copy of your text in which the PII is redacted. For more information, see Detect personally identifiable information (PII).	September 17, 2020
New feature	Previously, you could only train a model on up to 12 custom entities. Now Amazon Comprehend allows you to train a model on up to 25 custom entities at a time. For more information, see Custom entity recognition.	August 12, 2020

Language expansion	Amazon Comprehend has added five additional languages for the custom entity recognition feature: German (de), Spanish (es), French (fr), Italian (it), and Portuguese (pt). For more information, see Supported languages in Amazon Comprehend.	August 12, 2020
New feature	Amazon Comprehend now allows you to establish a private connection with your Virtual Private Cloud (VPC) by creating an interface VPC endpoint. For more informati on, see VPC endpoints (AWS PrivateLink).	August 11, 2020
New feature	With Amazon Comprehend, you can now quickly detect custom entities in individua l text documents by running real-time analysis. For more information, see Detecting custom entities in real time with amazon comprehend.	July 9, 2020

New feature added	Amazon Comprehend now provides support for a second mode in asynchronous Custom Classification for documents that provides greater flexibility when applying custom classes to documents. While multi-class mode associates only a single class with each document, the new multi-label mode can associate more than one. For example, a movie can be classified as both science fiction and action at the same time. For more information, see Multi-class and multi-label modes in custom classification.	December 19, 2019
New feature added	Amazon Comprehend now provides support for real-time Custom Classification for documents with unstructured text. Customers can use real-time custom classification to understand, label and route information based on their own business rules, synchronously. For more information, see Real-time analysis with custom classification.	November 25, 2019

New languages added	Amazon Comprehend has added six additional languages for several of its features: Arabic (ar), Hindi (hi), Japanese (ja), Korean (ko), simplified Chinese (zh), and traditional Chinese (zh-TW). These new languages are supported only for Determine Sentiment, Detect Key Phrases, and non-custom Detect Entities operations. For more information, see Supported languages.	November 6, 2019
New feature	Previously, you could only train a model on a single custom entity. As a result, you could only search for that one entity with an entity recognition operation. Amazon Comprehend has changed this and you can now train a model on up to 12 custom entities at a time. For more information, see Custom entity recognition	July 9, 2019

New feature	Amazon Comprehend now provides a multi-class confusion matrix for added ability to analyze metrics when training a Custom Classifier. This is currently supported using the APIs only. For more informati on, see Tagging resources in Amazon Comprehend	April 5, 2019
New feature	Amazon Comprehend provides tags for Custom Classifiers and Custom Entity Recognizers, which can be used as metadata that enables you to organize, filter, and control access to your resources with a finer level of control than ever. For more information, see Tagging resources in Amazon Comprehend	April 3, 2019

New feature	Amazon S3 already enables you to encrypt your input documents, and Amazon Comprehend extends this even farther. By using your own KMS key, you can not only encrypt the output results of your job, but also the data on the storage volume attached to the compute instance that processes the analysis job. The result is end-to-end security. For more informati on, see KMS encryption in Amazon Comprehend	March 28, 2019
New feature	Custom entity recognition extends the capability of Amazon Comprehend by enabling you to identify new entity types not supported as one of the preset generic entity types. This means you can analyze documents and extract entities like product codes or business-specific entities that fit your particular needs. For more information, see Custom entity recognition	November 16, 2018

New feature	You can use Amazon Comprehend to build your own models for custom classification, assigning a document to a class or a category. For more informati on, see Document classific ation.	November 15, 2018
Region expansion	Amazon Comprehend is now available in Europe (Frankfurt) (eu-central-1).	October 10, 2018
Language expansion	In addition to English and Spanish Amazon Comprehend can now also examine documents in French, German, Italian, and Portuguese. For more information, see Supported languages in Amazon Comprehend.	October 10, 2018
Region expansion	Amazon Comprehend is now available in Asia Pacific (Sydney) (ap-southeast-2).	August 15, 2018
New feature	Amazon Comprehend now parses documents to discover the syntax of a document and the part of speech for each word. For more information, see Syntax.	July 17, 2018

New feature	Amazon Comprehend now supports asynchronous batch processing for language, key phrase, entity, and sentiment detection. For more information, see Asynchronous batch processing.	June 27, 2018
New guide	This is the first release of the *Amazon Comprehend Developer Guide*.	November 29, 2017

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the *AWS Glossary Reference*.