Getting Started with Amazon Comprehend: Custom Entity Recognition

**SPL-TF-100-MLACEN-1 - Version 1.0.6**

© 2024 Amazon Web Services, Inc. or its affiliates. All rights reserved. This work may not be reproduced or redistributed, in whole or in part, without prior written permission from Amazon Web Services, Inc. Commercial copying, lending, or selling is prohibited. All trademarks are the property of their owners.

Note: Do not include any personal, identifying, or confidential information into the lab environment. Information entered may be visible to others.

Corrections, feedback, or other questions? Contact us at [*AWS Training and Certification*](https://support.aws.amazon.com/#/contacts/aws-training).

**Lab overview**

Your company, AnyCompany Consulting, wants to integrate natural language processing (NLP) to extract key phrases, entities, and sentiment from documents. They have chosen Amazon Comprehend as their NLP service so they can integrate NLP into their existing applications. To start the process of creating models and endpoints for their documents, they want you to create a custom entity recognizer model to identify specific terms in a collection of documents.

In this lab, you create and test a custom entity recognition model using Amazon Comprehend.

OBJECTIVES

By the end of this lab, you will be able to:

* Create a custom entity recognition model using Amazon Comprehend.
* Conduct a real-time analysis with a custom entity recognition endpoint.

TECHNICAL KNOWLEDGE PREREQUISITES

To successfully complete this lab, you should be familiar with basic navigation of the AWS Management console and have knowledge of Amazon S3.

DURATION

This lab requires approximately *60* minutes to complete.

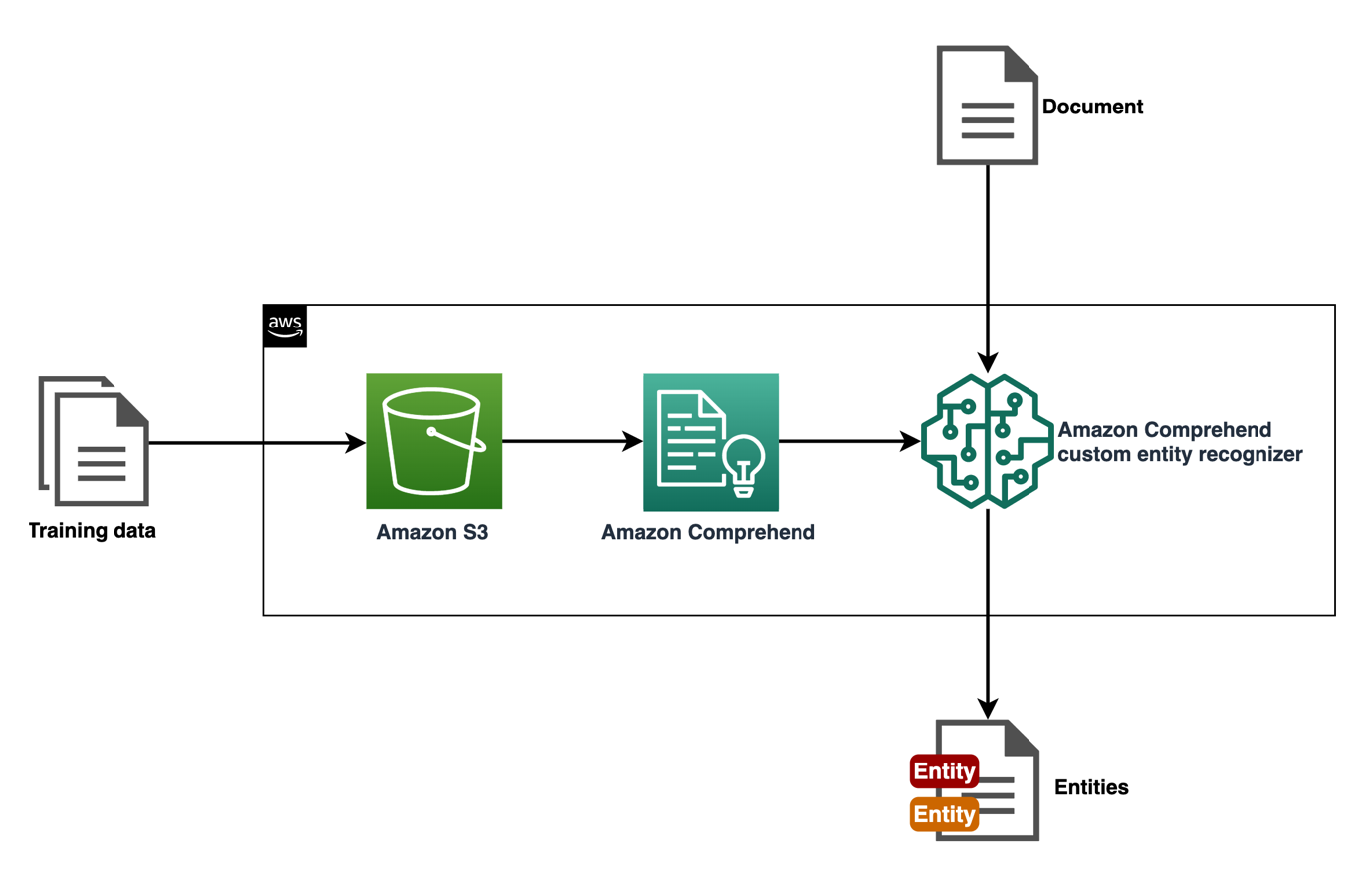
ICON KEY

Various icons are used throughout this lab to call attention to different types of instructions and notes. The following list explains the purpose for each icon:

* **Note:** A hint, tip, or important guidance.
* **Learn more:** Where to find more information.
* **Caution:** Information of special interest or importance (not so important to cause problems with the equipment or data if you miss it, but it could result in the need to repeat certain steps).
* **Consider:** A moment to pause to consider how you might apply a concept in your own environment or to initiate a conversation about the topic at hand.
* **File contents:** A code block that displays the contents of a script or file you need to run that has been pre-created for you.

ENVIRONMENT OVERVIEW

The following diagram shows the basic architecture of the lab environment:



*In the preceding diagram, training data is uploaded to an Amazon S3 bucket. Amazon Comprehend uses the training data in the bucket to train a custom entity recognition model. When documents are analyzed in real-time using an endpoint, the trained model finds any entities in the document.*

**Start lab**

1. To launch the lab, at the top of the page, choose **Start lab**.

 You must wait for the provisioned AWS services to be ready before you can continue.

1. To open the lab, choose **Open Console**.

You are automatically signed in to the AWS Management Console in a new web browser tab.

**Do not change the Region unless instructed.**

COMMON SIGN-IN ERRORS

**Error: You must first sign out**



If you see the message, **You must first log out before logging into a different AWS account:**

* Choose the **click here** link.
* Close your **Amazon Web Services Sign In** web browser tab and return to your initial lab page.
* Choose **Open Console** again.

**Error: Choosing Start Lab has no effect**

In some cases, certain pop-up or script blocker web browser extensions might prevent the **Start Lab** button from working as intended. If you experience an issue starting the lab:

* Add the lab domain name to your pop-up or script blocker’s allow list or turn it off.
* Refresh the page and try again.

**Task 1: Review the documents in the S3 bucket**

AnyCompany Consulting has given you a set of annotated data about AWS services to train your first custom entity recognition model using Amazon Comprehend. There is a documents.txt file that contains excerpts about AWS services and an annotations.csv file that contains annotations for the documents.txt file to support the model training. Review the documents before you create your model.

1. At the top of the AWS Management Console, in the search bar, search for and choose

S3

.

1. Choose the link for the bucket name that starts with **databucket**.

You can create a custom entity recognition model in Amazon Comprehend using either annotations or entity lists. When you use annotations, you provide a dataset containing annotated entities for model training. When you use entity lists, you provide a list of entities and their type label, and a set of unannotated documents containing those entities for model training. In this lab, you use a dataset containing annotated entities.

There is a text file in the bucket named **documents.txt** that contains text about AWS services. There is a CSV file in the bucket named **annotations.csv** that contains annotations for the text file. There are 1000 records in the dataset that are treated as individual documents during model training.

**File contents:** The **documents.txt** file contains data similar to this:

"Pearson Boosts Security and Productivity Using Amazon Elasticsearch Service"

"2020"

"Global educational media company Pearson needed a more efficient way to analyze and gain insights from its log data. With a number of teams in various locations using Elasticsearch\u2014the popular open-source tool for search and log analytics\u2014Pearson found that keeping track of log data and managing updates led to high operating costs. Faced with this, as well as increasingly complex security log management and analysis, the company found a solution on Amazon Web Services (AWS). Pearson quickly saw improvements by migrating from its self-managed open-source Elasticsearch architecture to Amazon Elasticsearch Service, a fully managed service that makes it easy to deploy, secure, and run Elasticsearch cost effectively at scale. Rather than spending considerable time and resources on managing the Elasticsearch clusters on its own, Pearson used the managed Amazon Elasticsearch Service as part of its initiative to modernize its products. "

**File contents:** The **annotations.csv** file contains data similar to this:

File,Line,Begin Offset,End Offset,Type

documents.txt,0,47,75,AWS\_SERVICE

documents.txt,2,167,180,AWS\_SERVICE

documents.txt,2,453,479,AWS\_SERVICE

documents.txt,2,590,610,AWS\_SERVICE

documents.txt,2,860,888,AWS\_SERVICE

documents.txt,5,17,45,AWS\_SERVICE

documents.txt,7,0,26,JOB\_TITLE

documents.txt,7,31,56,JOB\_TITLE

**Consider:** How many types are included in the annotations file?

There are two types included in the annotations file: **AWS\_SERVICE** and **JOB\_TITLE**.

**Note:** If you want to download and view a file, select the filename and choose **Download**.

 Congratulations! You have successfully reviewed the documents in the S3 bucket.

**Task 2: Create a custom entity recognition model**

AnyCompany wants a model that can recognize AWS service names and job titles. In this task, you create a custom entity recognition model that recognizes AWS\_SERVICE and JOB\_TITLE types using Amazon Comprehend.

1. At the top of the AWS Management Console, in the search bar, search for and choose

Amazon Comprehend

.

1. Choose **Launch Amazon Comprehend**.
2. In the navigation pane at the left of the page, in the **Customization** section, choose **Custom entity recognition**.
3. Choose **Create new model**.
4. On the **Create new model** page, in the **Model settings** section:

* For **Model name**, enter

aws-entity-recognizer

.

* For **Version name**, enter

1

.

* For **Custom entity type**, enter

AWS\_SERVICE

.

* Choose **Add type**.
* For **Custom entity type**, enter

JOB\_TITLE

.

* Choose **Add type**.

You have added two custom labels for the recognizer to identify in your dataset. These types match the types in your **annotations.csv** file.

1. In the **Data specifications** section, under **Training dataset**:

* For **Training type**, select **Using annotations and training docs**.
* For **Annotation location on S3**, copy and paste the **AnnotationDataLocation** value that is listed to the left of these instructions.
* For **Training data location on S3**, copy and paste the **TrainingDataLocation** value that is listed to the left of these instructions.

1. In the **IAM role** section:

* For **IAM role**, choose **Use an existing IAM role**.
* For **Role name**, choose **ComprehendEntityRecognitionRole**.

1. At the bottom of the page, choose **Create**.

In the **Recognizer models** section, an **aws-entity-recognizer** model appears with a **Version status** of **Submitted**.

**Note:** This model takes 15-20 minutes to train. While you wait for the model to train, explore the Amazon Comprehend documentation.

**Learn more:** Refer to [Custom entity recognition](https://docs.aws.amazon.com/comprehend/latest/dg/custom-entity-recognition.html) for more information about how custom entity recognition works in Amazon Comprehend and the options for creating a custom model.

**Learn more:** Refer to [Preparing entity recognizer training data](https://docs.aws.amazon.com/comprehend/latest/dg/prep-training-data-cer.html) for more information about preparing training data and when to use annotations or entity lists.

**Learn more:** Refer to [Training custom entity recognizer models](https://docs.aws.amazon.com/comprehend/latest/dg/training-recognizers.html) for more information about training custom recognition models using Amazon Comprehend.

**Note:** While you wait for the model to train, you can also explore the real-time analysis features in Amazon Comprehend. Amazon Comprehend can use built-in or custom models to analyze text in real-time. You can recognize entities, extract key phrases, detect dominant languages, detect Personally Identifiable Information (PII), determine sentiment, detect targeted sentiment, or analyze syntax. You can also use custom models to detect entities or classify documents. To explore real-time analysis in Amazon Comprehend, in the navigation pane at the left of the page, choose **Real-time analysis**. Then, explore the real-time analysis features.

 Congratulations! You have successfully created a custom entity recognition model.

**Task 3: Use a custom entity recognition model**

Your model is trained and is ready for real-time analysis. In this task, you view the model performance, create an endpoint for custom entity recognition, and use the model in a real-time analysis.

TASK 3.1: VIEW THE MODEL PERFORMANCE

First, view the **aws-entity-recognizer** model to see how the model training performed.

1. In the navigation pane at the left of the page, in the **Customization** section, choose **Custom entity recognition**.

**Refresh:** To see the current **Version status**, refresh your browser tab.

**Caution:** If the **aws-entity-recognizer** model has not finished training, wait until **Version status** changes to **Trained**.

1. Choose **aws-entity-recognizer**.

**Note:** You can retrain your custom entity recognition model with more documents. If you make a new version of the model, the version appears in this section.

1. Choose **1** to open the model.

The **Version details** section contains information about the model, version, how long the training took, the number of trained documents, the number of test documents, and other important model training details.

1. Choose the **Performance** tab.

**Consider:** What are the precision and recall scores for the model?

The model scores are 100 for both custom entity types. Scores can range from 0 to 100 for these metrics. Scores of 100 for all metrics and entity types might indicate overfitting or underfitting. You can improve the model by adding more records to the training data, cleaning the data, or balancing the number of train mentions for the two custom entity types. In this lab, a model with scores of 100 is sufficient to identify AWS services and job titles.

**Learn more:** The metrics provide an insight into how your custom entity recognizer performs during an entity recognition job. Refer to [Custom entity recognizer metrics](https://docs.aws.amazon.com/comprehend/latest/dg/cer-metrics.html) for more information about Amazon Comprehend custom entity recognition model scores.

You have viewed the model performance and are ready to create an endpoint.

TASK 3.2: CREATE AN ENDPOINT

Next, create an endpoint to use in a real-time analysis.

1. Choose the **Endpoints** tab.
2. Choose **Create endpoint**.
3. For **Endpoint name**, enter

aws-entity-recognizer-endpoint

.

1. For **Custom model type**, choose **Custom entity recognition**.
2. For **Entity recognition models and flywheels**, choose **aws-entity-recognizer**.
3. For **Version**, choose **1**.
4. For **Number of inference units (IUs)**, enter

1

.

1. Choose the **Acknowledge** message to approve the endpoint charges.

**Learn more:** You do not incur separate charges by using the endpoint in this lab. The endpoint gets deleted automatically when the lab ends. If you create an endpoint in your production environment, there are additional charges. Refer to [Amazon Comprehend Pricing](https://aws.amazon.com/comprehend/pricing/) for more information.

1. Choose **Create endpoint**.
2. Wait until the **Status** changes from **Creating** to **Active**.

**Refresh:** To see the current **Status**, refresh your browser tab.

**Note:** This endpoint takes 10-15 minutes to create.

**Learn more:** Refer to [Managing Amazon Comprehend endpoints](https://docs.aws.amazon.com/comprehend/latest/dg/manage-endpoints.html) for more information about using, monitoring, updating, and deleting Amazon Comprehend endpoints.

You have created an endpoint and are ready to use it in a real-time analysis.

TASK 3.3: USE THE MODEL IN A REAL-TIME ANALYSIS

Finally, use your *aws-entity-recognizer* custom entity recognition model in a real-time analysis, testing AWS service documentation and customer testimonials to see which AWS service and job titles the model can identify.

1. In the navigation pane at the left of the page, choose **Real-time analysis**.
2. In the **Input data** section:

* For **Analysis type**, choose **Custom**.
* For **Custom model type**, choose **Custom entity recognition**.
* For **Endpoint**, choose **aws-entity-recognizer-endpoint**.
* For **Input text**, clear the text and enter:

Amazon HealthLake is a HIPAA-eligible service that healthcare providers can use to store, transform, query, and analyze large-scale health data.

Health data is frequently incomplete and inconsistent. It's also often unstructured, with information contained in clinical notes, lab reports, insurance claims, medical images, recorded conversations, and time-series data (for example, heart ECG or brain EEG traces).

1. Choose **Analyze**.

The results appear in the **Insights** section.

1. For **Input text**, clear the text and enter:

Amazon Textract is based on the same proven, highly scalable, deep-learning technology that was developed by Amazon's computer vision scientists to analyze billions of images and videos daily. You don't need any machine learning expertise to use it, as Amazon Textract includes simple, easy-to-use API operations that can analyze image files and PDF files. Amazon Textract is always learning from new data, and Amazon is continually adding new features to the service.

Vice President of Security Engineering for AnyCompany Engineering helped lead an initiative to catalogue all of their unstructured PDF documents in a DynamoDB database using Textract. Collaborating with the Senior Director of Cloud Engineering, the project transitioned 80% of their on-prem workloads to the cloud.

1. Choose **Analyze**.

The results appear in the **Insights** section.

CHALLENGE A

Take a moment to write your own AWS service text or copy text from the AWS documentation. Analyze the text and see what entities the model recognizes.

Expand the **Hint** or **Solution** sections if you want help solving the challenge.

**Hint**

**Solution**

 Congratulations! You have successfully used a custom entity recognition model in a real-time analysis.

**Conclusion**

 Congratulations! You now have successfully:

* Created a custom entity recognition model using Amazon Comprehend.
* Conducted a real-time analysis with a custom entity recognition endpoint.

**End lab**

Follow these steps to close the console and end your lab.

1. Return to the **AWS Management Console**.
2. At the upper-right corner of the page, choose **AWSLabsUser**, and then choose **Sign out**.
3. Choose **End lab** and then confirm that you want to end your lab.

For more information about AWS Training and Certification, see [*https://aws.amazon.com/training/*](https://aws.amazon.com/training/).

*Your feedback is welcome and appreciated.*  
If you would like to share any feedback, suggestions, or corrections, please provide the details in our [*AWS Training and Certification Contact Form*](https://support.aws.amazon.com/#/contacts/aws-training).