



AWS Workshop

Amazon Comprehend

*Getting Started with Comprehend Custom – Custom
Entity Recognition*

Lab Overview

This lab shows you how to prepare a dataset and train a custom entity recognizer using Amazon Comprehend Custom Entity Recognizer. We will use a Twitter customer support dataset for this exercise and the goal is to create a custom entity recognizer to recognize smartphone devices.

The steps in this lab include:

- ☐ Create an S3 Bucket for storing training and test data as well as our entity list
- ☐ Setup your SageMaker notebook instance environment that will be used to explore data & build your Custom Entity Recognizer using Comprehend APIs
- ☐ Setup your IAM Role & Permissions
- ☐ Download Twitter dataset & execute lab from your SageMaker Notebook Instance

For this lab, the bulk of the instructions will be contained and executed within the Jupyter notebook itself so will not be repeated within this document.

Create a new S3 Bucket

For the model we will be training in this lab, we will need to create a new S3 bucket to store our:

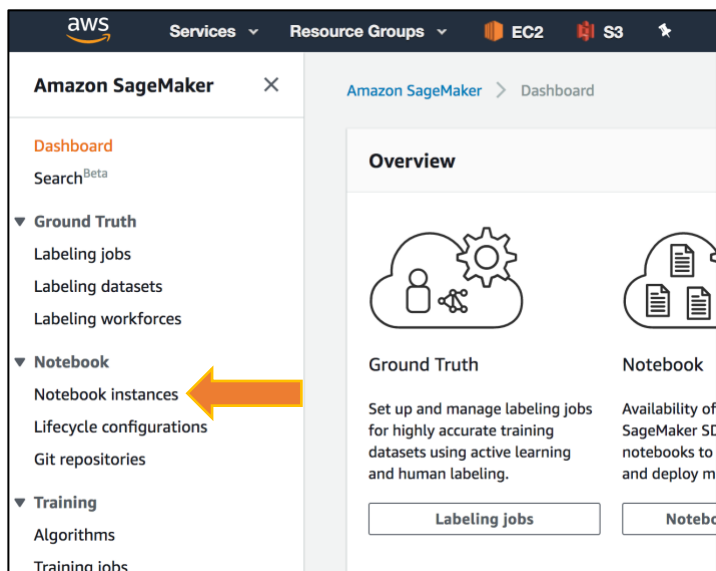
- (1) training & test dataset
- (2) entity list

1. In the upper-right corner of the AWS Management Console, confirm you are in the desired AWS region (e.g., N.Virginia).
2. Click on **Services**, then **S3** on the top menu.
3. Click **Create Bucket**, name your bucket (ex. comprehend-lab-**<your-initials>**).
4. Ensure your bucket region is listed as the region you are working in (Ex. US East). Leave all other options as the default, then scroll down and click **Create Bucket**

Create a new SageMaker Notebook Instance

In this portion of the lab, you will need to create a new notebook instance in SageMaker. This notebook instance will be used for the majority of the lab to explore our Twitter data and train our model using Comprehend APIs.

1. In the upper-right corner of the AWS Management Console, confirm you are in the desired AWS region (e.g., N.Virginia).
2. Click on **Services**, then **Amazon SageMaker** on the top menu. From the left menu, select **Notebook instances**.



5. Click the **Create notebook instance** button on the upper right
6. Under Create notebook instance:
7. **Notebook instance name:** **'comprehend-custom-*<your-initials>*'** replacing *<your-initials>* with your own initials (ex. AIML-Workshop-sde).
8. **Notebook instance type:** ml.t2.medium
9. Under **Permissions and encryption:**
 - * IAM Role -> select **'Create a new role'** from the dropdown
 - * On the pop-up, select **'Any S3 bucket'** , then click **Create Role**

Create an IAM role

×

Passing an IAM role gives Amazon SageMaker permission to perform actions in other AWS services on your behalf. Creating a role here will grant permissions described by the [AmazonSageMakerFullAccess](#) IAM policy to the role you create.

The IAM role you create will provide access to:

☒ S3 buckets you specify - *optional*

☐ Specific S3 buckets

Example: bucket-name-1, bucket-name-2, bucket-name-3

Comma delimited. ARNs, "*" and "/" are not supported.

☒ Any S3 bucket

Allow users that have access to your notebook instance access to any bucket and its contents in your account.

☐ None

☒ Any S3 bucket with "sagemaker" in the name

☒ Any S3 object with "sagemaker" in the name

☒ Any S3 object with the tag "sagemaker" and value "true"
 [See Object tagging](#)

☒ S3 bucket with a Bucket Policy allowing access to SageMaker
 [See S3 bucket policies](#)

Cancel

Create role

10. Under **Git repositories**

* Select 'Clone a public Git repository to this notebook instance only'

* For **Git repository URL**, enter:

<https://github.com/aws-samples/amazon-comprehend-custom-entity>

▼ Git repositories - *optional*

▼ Default repository

Repository

Jupyter will start in this repository. Repositories are added to your home directory.

Clone a public Git repository to this notebook instance only

⌵

⌵

Git repository URL

Clone a repository to use for this notebook instance only.

https://github.com/aws-samples/amazon-comprehend-custom-entity

⌵

[Add additional repository](#)

11. **Leave all other options as default**

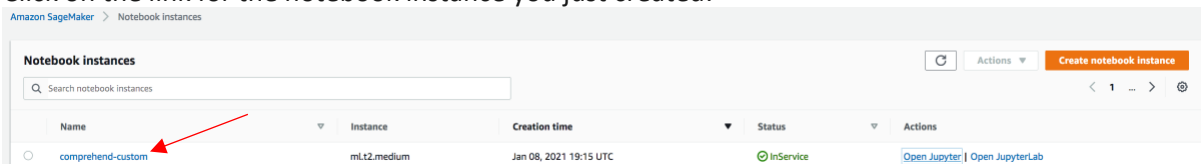
12. Click '**Create Notebook Instance**' at the bottom of the page

13. You will be returned to the Notebook Instance dashboard where you should see your notebook instance in a 'Pending' state until it is successfully created. When the notebook is ready for use you will see it change into a 'InService' status. (Timing Estimate: < 5 Minutes)

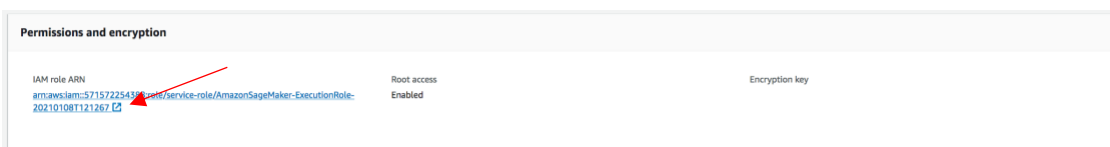
Add IAM Permissions to your Notebook Instance role

We now need to add permissions to the IAM role attached to your notebook instance so that we can interact with the Comprehend service via our notebook instance.

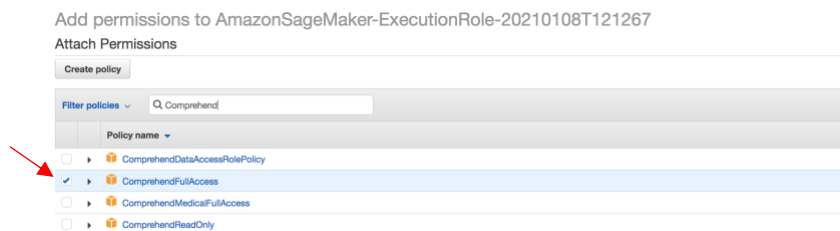
1. Click on the link for the notebook instance you just created:



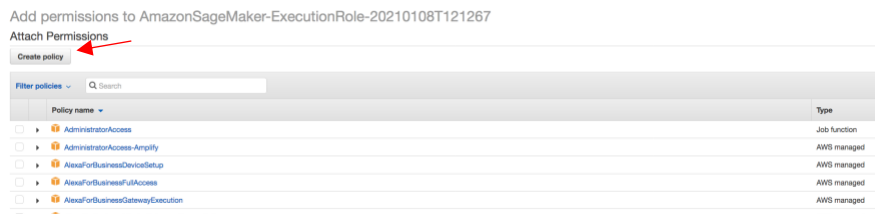
2. Scroll to **Permissions and encryption** and click on the IAM role attached to your notebook instance:



3. Click **Attach policies**, then search for **Comprehend** and select **ComprehendFullAccess**



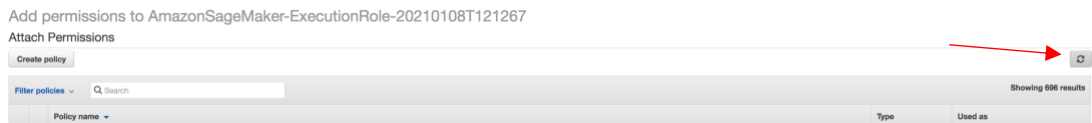
4. Click **Attach policy** in the bottom right corner
5. Click **Attach policies** again by clicking the button under the Permissions tab. This time click the **Create policy** button as shown below:



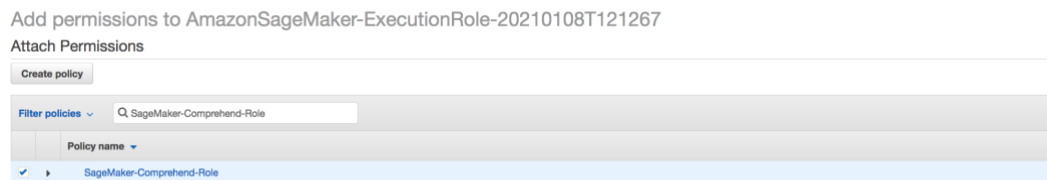
6. Click the JSON tab and copy/paste the contents of the JSON below:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Action": [
        "comprehend:*",
        "iam:ListRoles",
        "iam:GetRole",
        "iam:PassRole",
        "s3:ListAllMyBuckets",
        "s3:ListBucket",
        "s3:GetBucketLocation"
      ],
      "Effect": "Allow",
      "Resource": "*"
    }
  ]
}
```

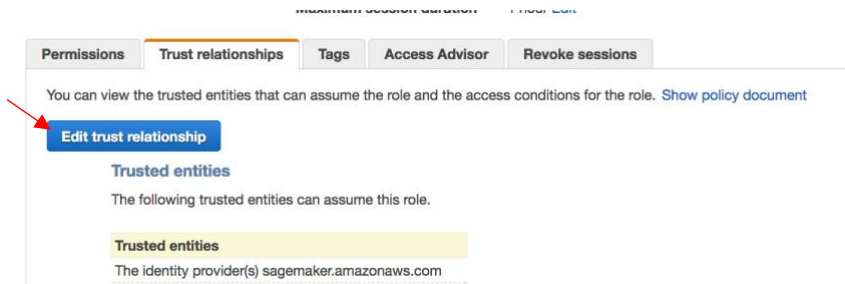
7. Click **Review policy** in the bottom right corner
8. Under **Review policy** enter a name for the role ***SageMaker-Comprehend-Role***, then click **Create policy** in the bottom right corner
9. We now need to attach to this new custom policy to our IAM role so from our previous screen/tab click refresh:



10. Now search for and select the role we just created:



11. Click **Attach policy**
12. Finally, click on the **Trust Relationships** tab of your IAM Role, click **Edit trust relationship**:



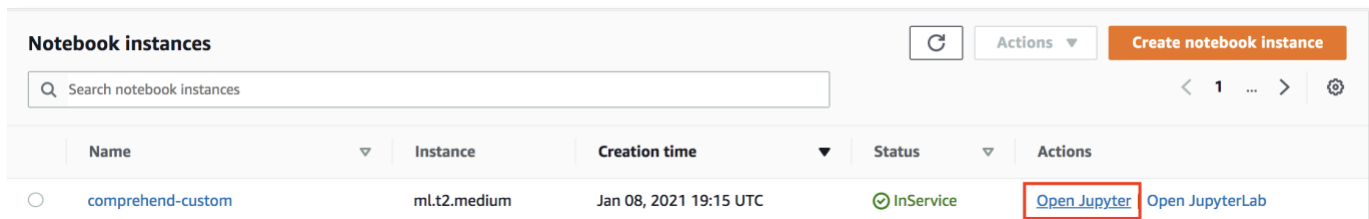
13. Copy & paste the json below replacing existing content:

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": [
          "comprehend.amazonaws.com",
          "s3.amazonaws.com",
          "sagemaker.amazonaws.com"
        ]
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

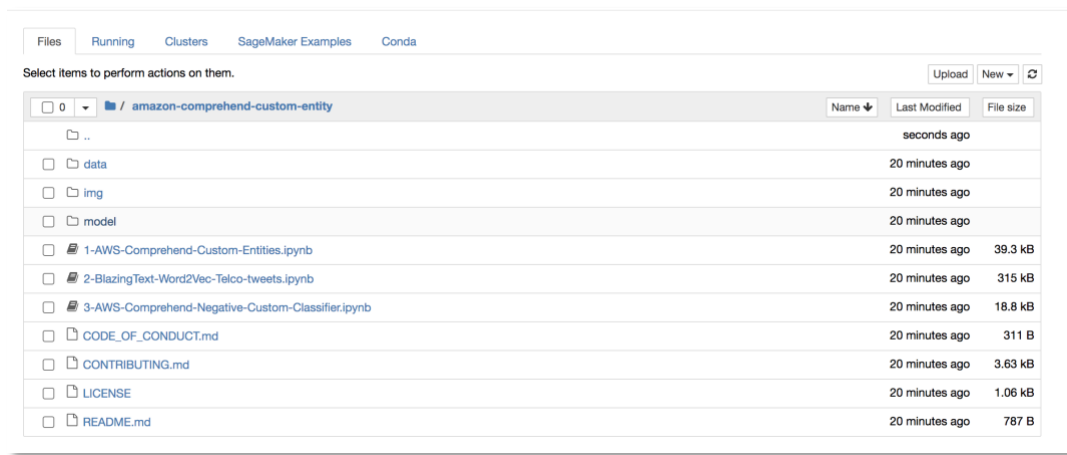
14. Click **Update Trust Policy**

Download Twitter Dataset & Execute Lab

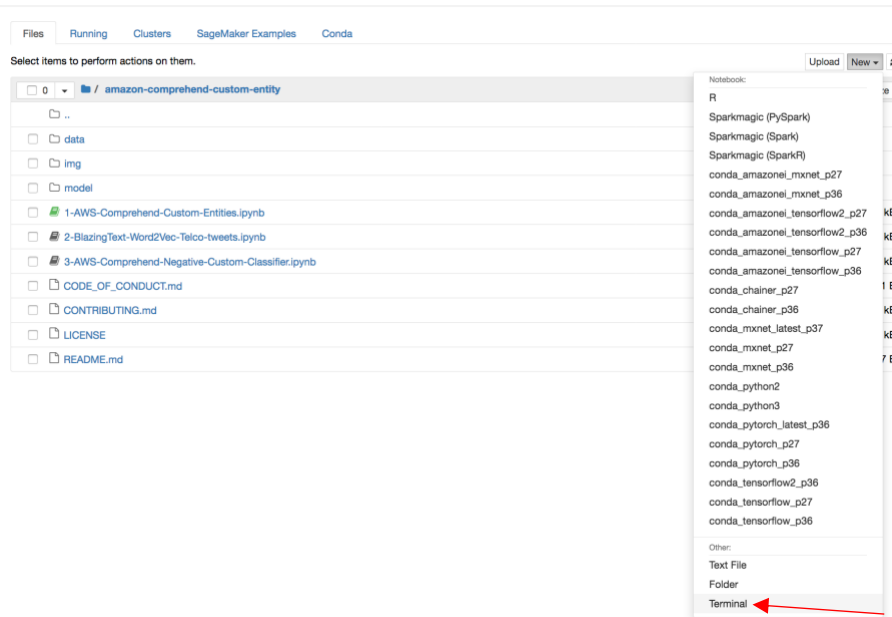
1. Click on **Services**, then **Amazon SageMaker** on the top menu. From the left menu, select **Notebook instances**.
2. Open the SageMaker notebook created in the previous step by selecting '**Open Jupyter**'



3. When you open your notebook, you should see the cloned git repository files under the 'Files' tab as shown below:



4. In this lab, we use a twitter dataset so before we dive into our notebook we need to download the dataset to our notebook instance. To do this:
 - a. Open a terminal on our notebook by, selecting **New** → **Terminal** from the drop down in the upper right corner:



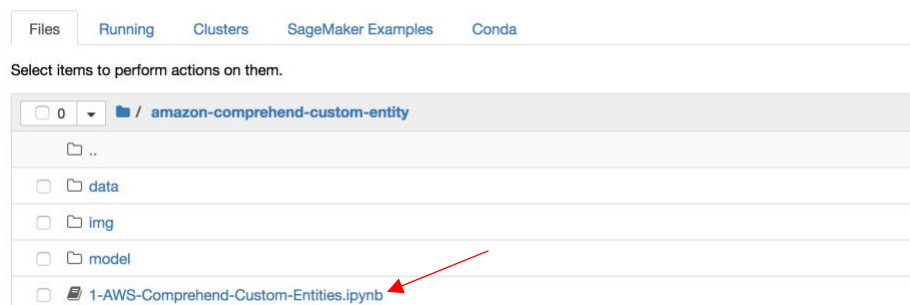
b. From the terminal copy & paste the following commands to download & unzip our data:

```
aws s3 cp s3://phi-demo-london/twcs/twcs.zip /home/ec2-user/SageMaker/amazon-comprehend-custom-
entity/data/twcs.zip
```

```
cd /home/ec2-user/SageMaker/amazon-comprehend-custom-entity/data
```

```
unzip twcs.zip
```

5. We're now ready to execute the lab from our notebook. Go back to your file view and select the first notebook **1-AWS-Comprehend-Custom-Entities**. This will bring up a copy of the Jupyter notebook in your environment. The rest of the lab will be executed from your notebook instance.



LAB NOTES:

- In the first steps of the lab, there are notes on setting up your IAM Role & Getting the dataset. You can ignore those notebook steps as we already did these steps via the detailed instructions above.
- Make sure you update your S3 bucket name as indicated in the notebook to use the S3 bucket we

created above.

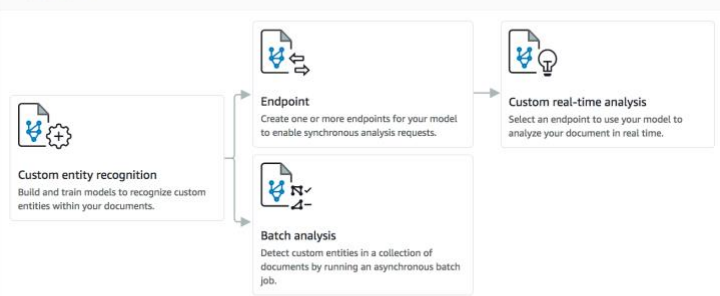
- This lab is timed to complete the first notebook only. There are two other notebooks in this example that you are welcome to continue with or explore as time permits.
- Although we are executing Comprehend API's from a SageMaker notebook instance to facilitate data exploration, the same steps to interact with Comprehend can be performed directly from the Console as well. You're encouraged to also look at the Comprehend service directly in the console to see what those API calls are generating behind the scenes. Example below:

Amazon Comprehend > Custom entity recognition

Custom entity recognition [Info](#)

Automatically train the recognizer to label words or sets of adjacent words with custom entity types. Automatic training requires having two types of information: sample documents and the entity list or annotations. Once the recognizer is trained, you can use it to detect custom entities in your documents. You can quickly analyze a small body of text in real time, or you can analyze a large set of documents with an asynchronous batch job.

▼ Overview



Entity recognizers (1)

Search

Status: All < 1 > ⌂

Name	Training data for...	Custom entity type	Training started	Training ended	Endpoints
custom-device-r...	CSV file	DEVICE	1/8/2021, 2:31:01 PM	1/8/2021, 3:01:15 PM	-