Custom Text Classification using Amazon Comprehend Task List

Click on the tasks below to view instructions for the workshop. In order to finish the workshop, kindly complete tasks in order from the top to the bottom.

1: Pre-requisite
2. Prepare Data
3: Train the Model
4: Configure SageMaker Notebook
5: Create Client
6. Clean up

1. Pre-requisite

You need to have an AWS account with administrative access to complete the workshop. If you don't have an AWS account, kindly use the link (https://aws.amazon.com/free) to create free trial account for AWS. If you are participating an AWS event, please get your account from the event organizer.

2. Prepare Data

You will use a sample data to train the model to classify the text. The workshop uses a sample data from the Kaggle website. Download the data from the link (https://aws-dojo.com/ws40/news_test.csv). The sample data has news titles and their classification as Real or Fake. It is a csv file with the following fields:

Class - It classifies text as Real or Fake

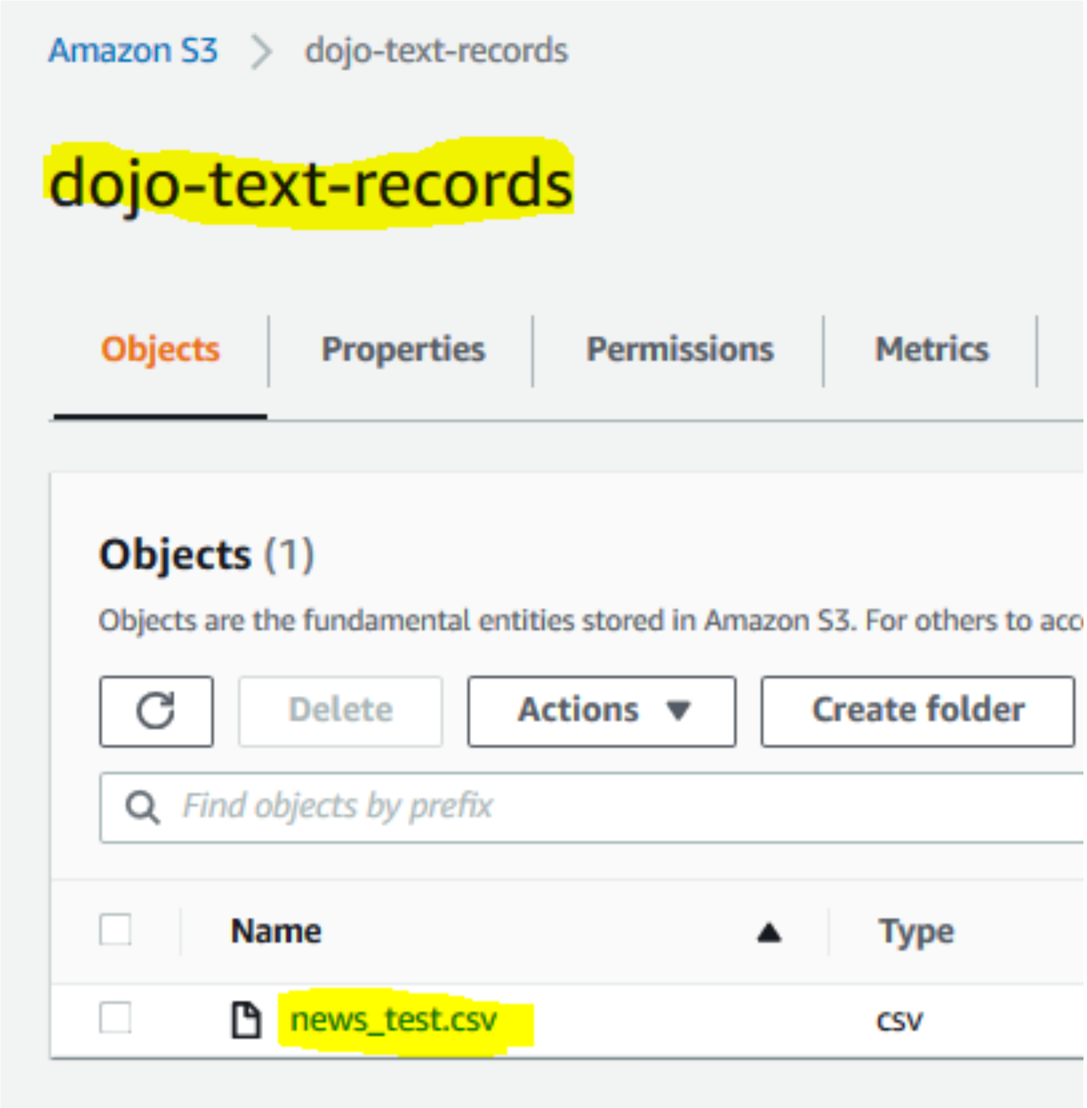
News Title - The news title text which is classified as Real or Fake

Please download the file to get familiar with the data and its format.

You upload the sample data file to a S3 bucket.

Login to AWS Console and choose Virginia as the region.

Create a bucket with name **dojo-text-records-[your name]** and upload news_test.csv file into the bucket. If this bucket name is not available, use a bucket name which is available.



Objects (1)

Objects are the fundamental entities stored in Amazon S3. For others to acc

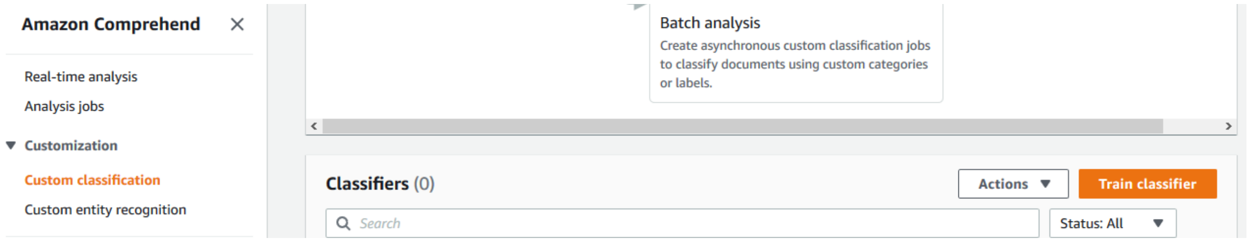
□	Name ▲	Type
□	news_test.csv	csv

The training data is ready. The next step is to build and train a model using this training data.

3: Train the Model

You use the sample data loaded in the S3 bucket to train a model for text classification. The model can predict whether a news title text is Real or Fake.

1. Go to the Amazon Comprehend console, click on the Custom classification menu in the left and then click on the Train classifier button.

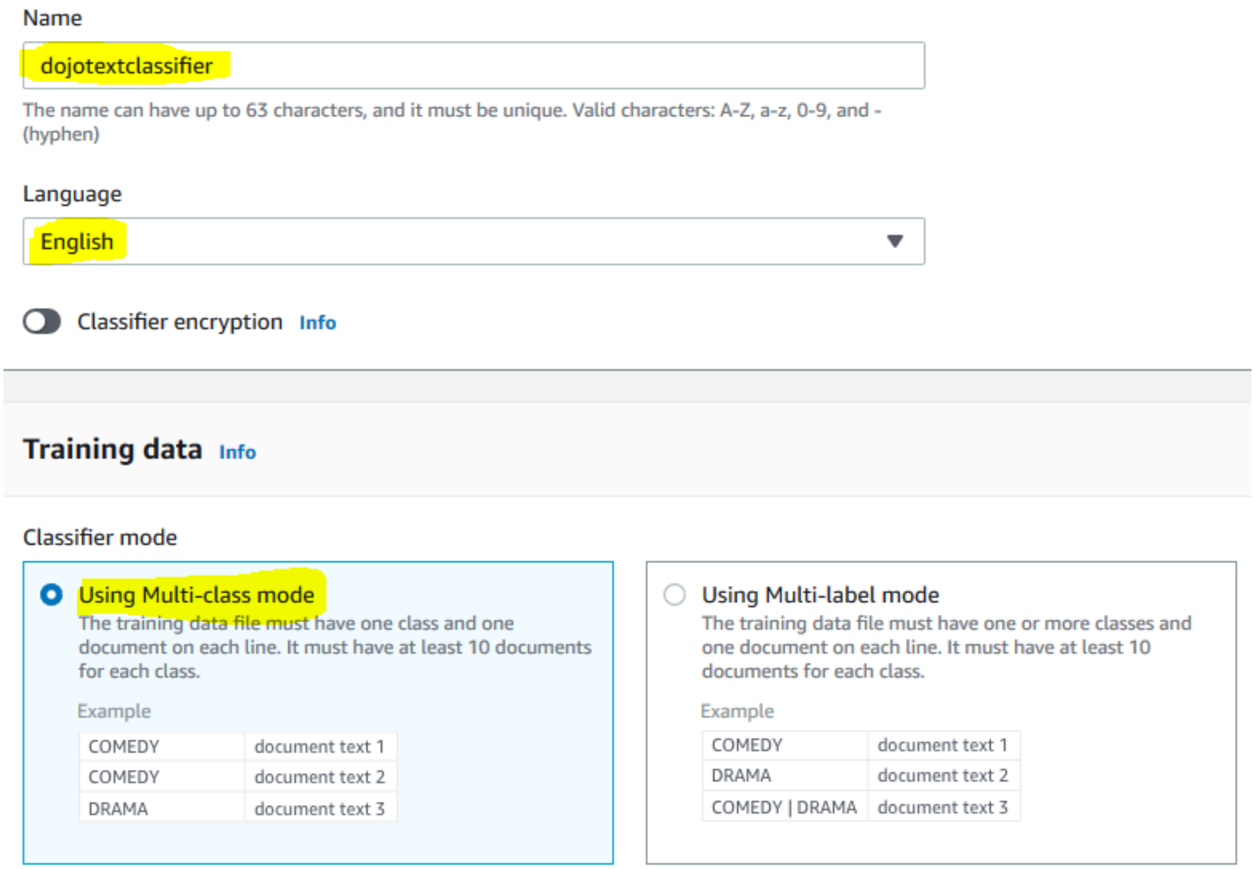


Batch analysis

Create asynchronous custom classification jobs to classify documents using custom categories or labels.

Classifiers (0)

2. On the next screen, type in **dojotextclassifier** for the name. Select English for the language. Select Using Multi-class mode option for the Classifier mode.



Name

dojotextclassifier

The name can have up to 63 characters, and it must be unique. Valid characters: A-Z, a-z, 0-9, and - (hyphen)

Language

English

Classifier encryption Info

Training data Info

Classifier mode

● Using Multi-class mode

The training data file must have one class and one document on each line. It must have at least 10 documents for each class.

Example

COMEDY	document text 1
COMEDY	document text 2
DRAMA	document text 3

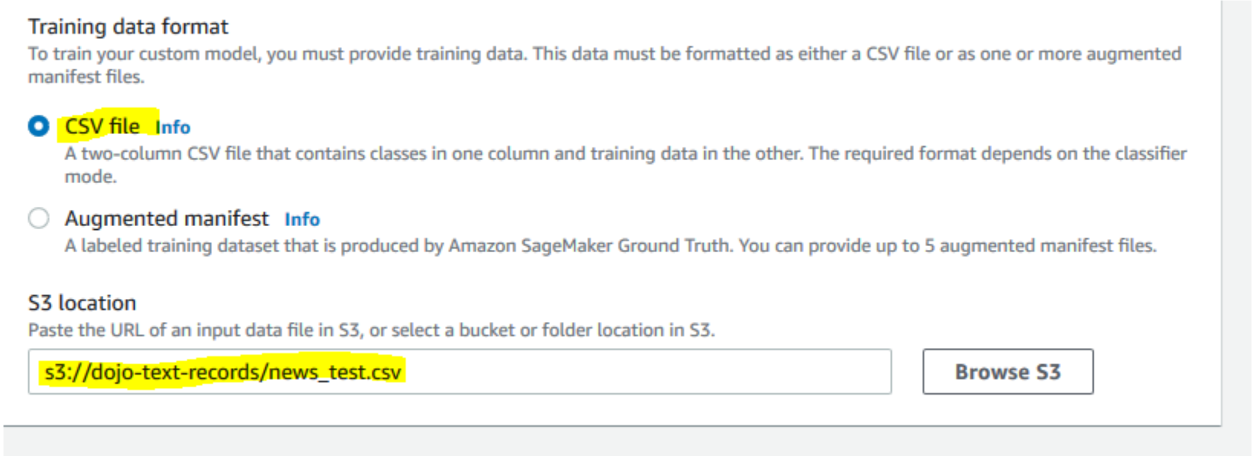
○ Using Multi-label mode

The training data file must have one or more classes and one document on each line. It must have at least 10 documents for each class.

Example

COMEDY	document text 1
DRAMA	document text 2
COMEDY | DRAMA	document text 3

3. On the same screen, select CSV file for the Training data format. Select s3://dojo-text-records-[your name]/news_test.csv for the S3 location. If you created bucket with a different name, then use that bucket.



Training data format

To train your custom model, you must provide training data. This data must be formatted as either a CSV file or as one or more augmented manifest files.

● CSV file Info

A two-column CSV file that contains classes in one column and training data in the other. The required format depends on the classifier mode.

○ Augmented manifest Info

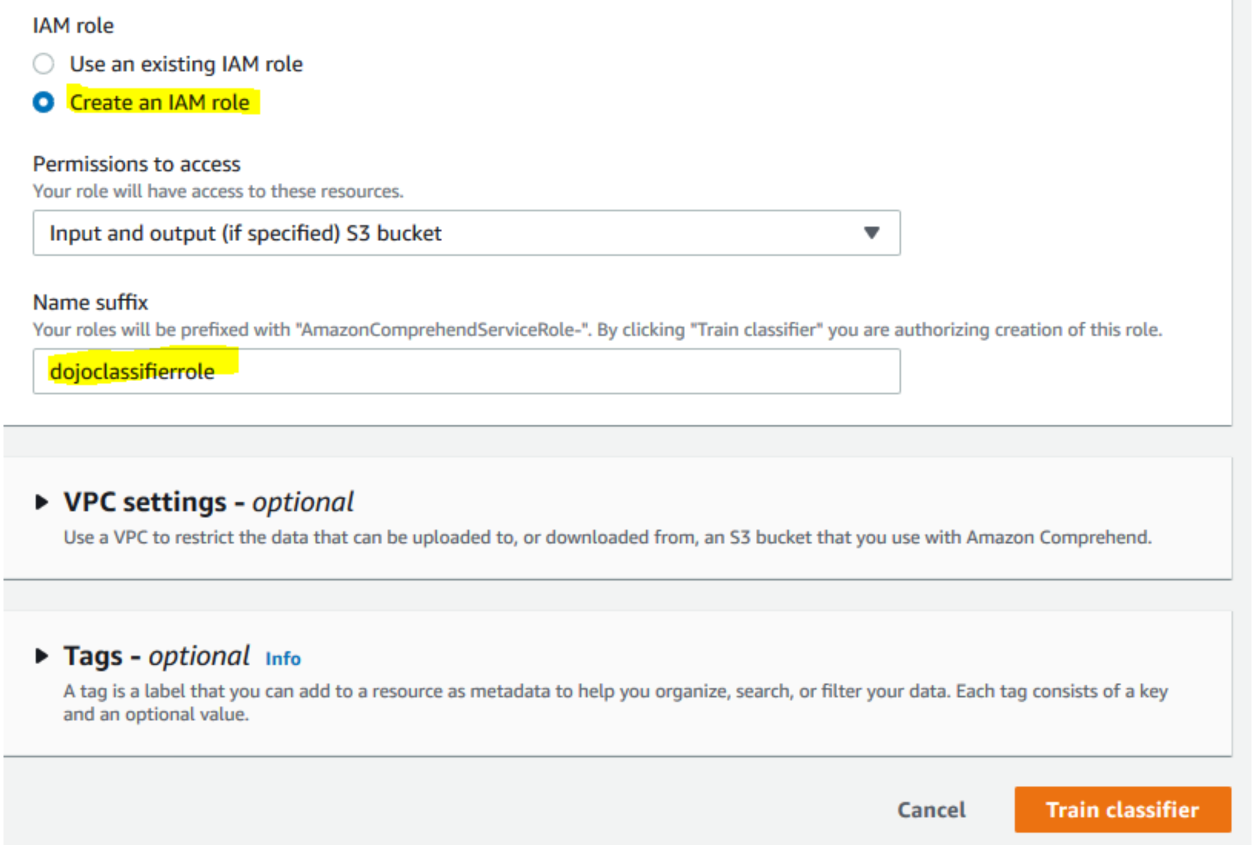
A labeled training dataset that is produced by Amazon SageMaker Ground Truth. You can provide up to 5 augmented manifest files.

S3 location

Paste the URL of an input data file in S3, or select a bucket or folder location in S3.

s3://dojo-text-records/news_test.csv

4. On the same screen, select Create an IAM role option. Type in **dojoclassifierrole** for the name suffix. Finally, click on the Train classifier button.



IAM role

○ Use an existing IAM role

● Create an IAM role

Permissions to access

Your role will have access to these resources.

Input and output (if specified) S3 bucket

Name suffix

Your roles will be prefixed with "AmazonComprehendServiceRole-". By clicking "Train classifier" you are authorizing creation of this role.

dojoclassifierrole

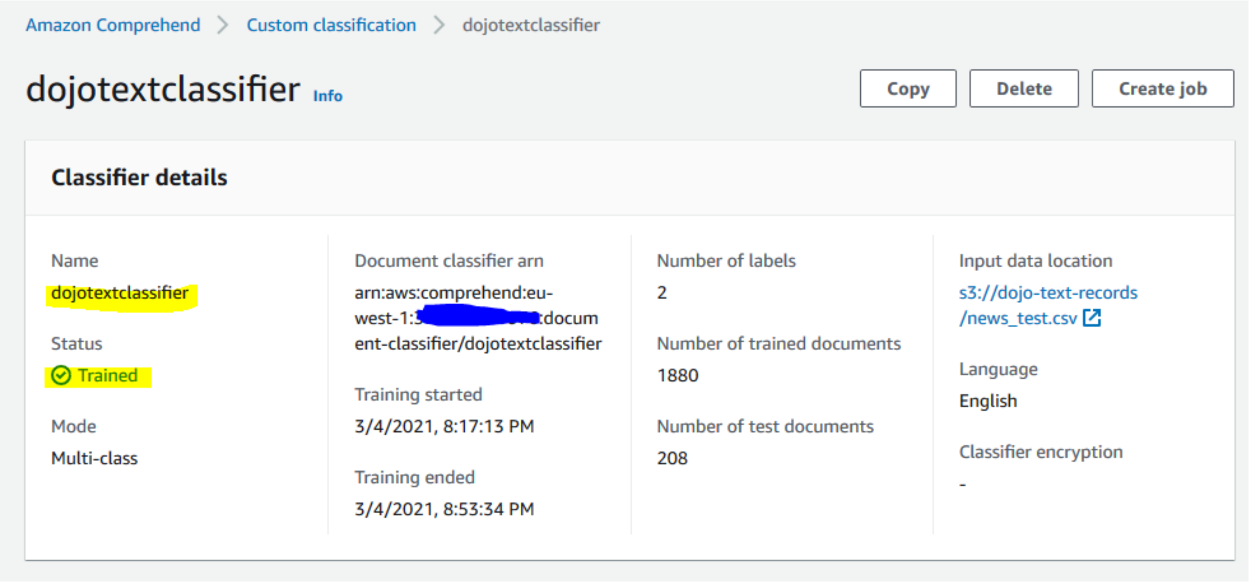
▶ **VPC settings** - *optional*

Use a VPC to restrict the data that can be uploaded to, or downloaded from, an S3 bucket that you use with Amazon Comprehend.

▶ **Tags** - *optional* Info

A tag is a label that you can add to a resource as metadata to help you organize, search, or filter your data. Each tag consists of a key and an optional value.

5. It will start training the model. Wait till the status changes to Trained. It might take up to an hour to train the model.


Amazon Comprehend > Custom classification > dojotextclassifier
Info
Copy
Delete
Create job

dojotextclassifier

Classifier details

Name
dojotextclassifier

Status
Trained

Mode
Multi-class

Document classifier arn
arn:aws:comprehend:eu-west-1:[redacted]document-classifier/dojotextclassifier

Training started
3/4/2021, 8:17:13 PM

Training ended
3/4/2021, 8:53:34 PM

Number of labels
2

Number of trained documents
1880

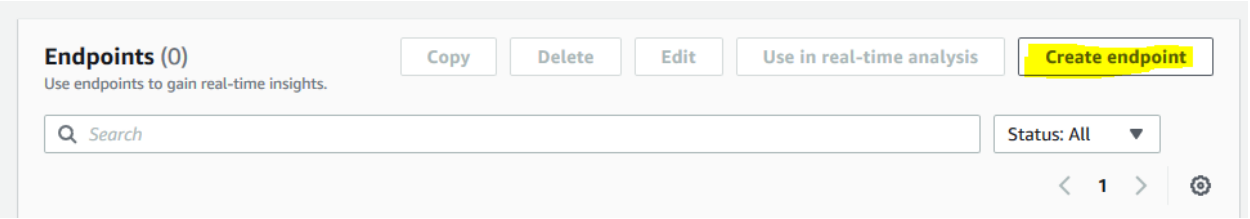
Number of test documents
208

Input data location
s3://dojo-text-records/news_test.csv

Language
English

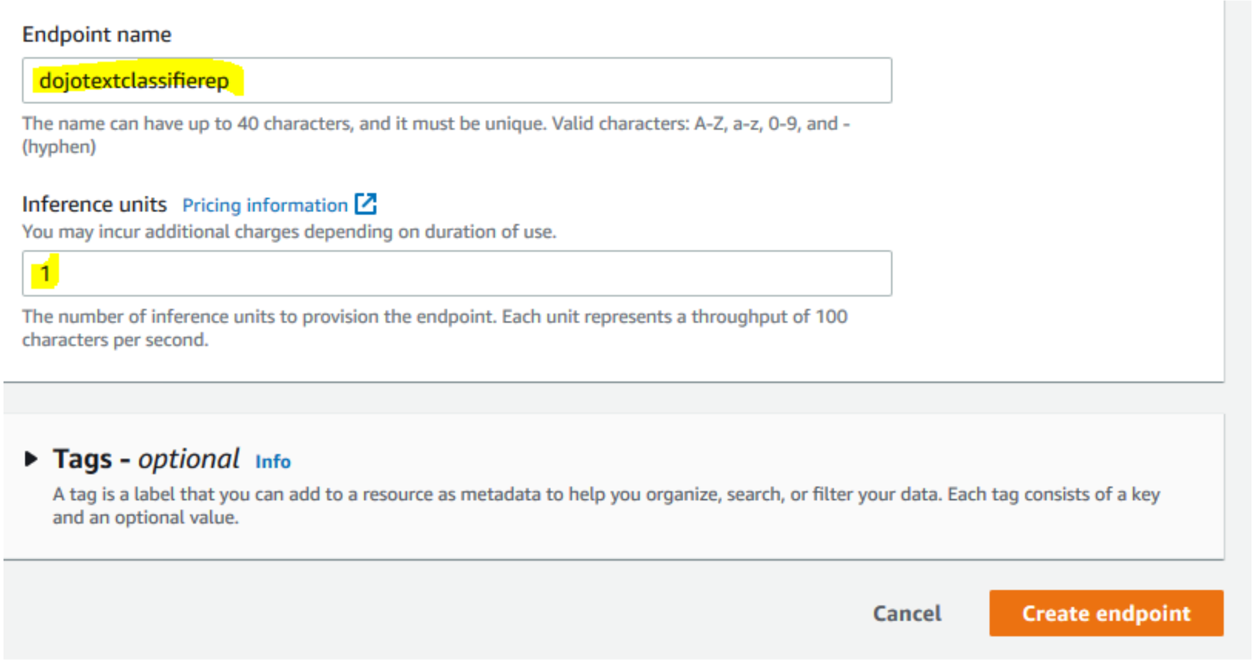
Classifier encryption
-

6. Once the model is trained, on the classifier details screen, click on the Create endpoint button.



Endpoints (0)
Use endpoints to gain real-time insights.

7. On the next screen, type in **dojotextclassifierep** as the endpoint name. Type in 1 for Inference units. Finally, click on the Create endpoint button.



Endpoint name

dojotextclassifierep

The name can have up to 40 characters, and it must be unique. Valid characters: A-Z, a-z, 0-9, and - (hyphen)

Inference units Pricing information

You may incur additional charges depending on duration of use.

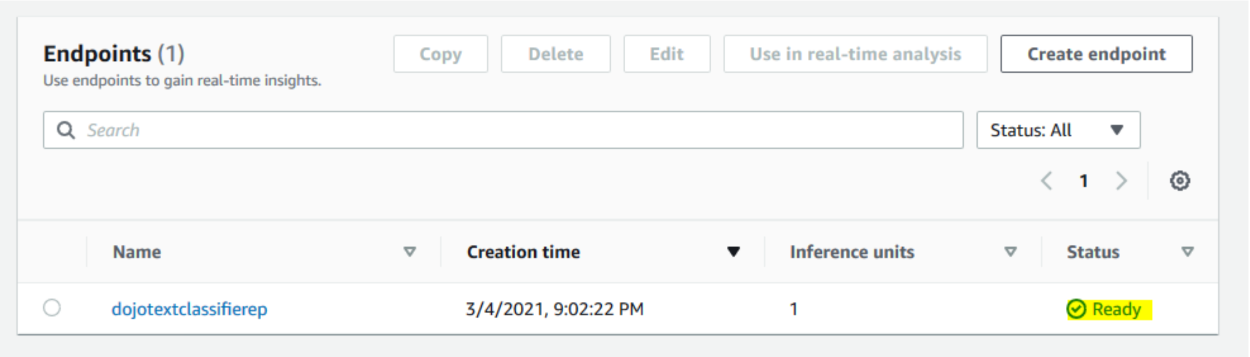
1

The number of inference units to provision the endpoint. Each unit represents a throughput of 100 characters per second.

▶ **Tags** - *optional* Info

A tag is a label that you can add to a resource as metadata to help you organize, search, or filter your data. Each tag consists of a key and an optional value.

8. It will throw pop up and ask for the confirmation, click on the Confirm button in the pop up. The endpoint creation will start. Wait till the status changes to Ready.

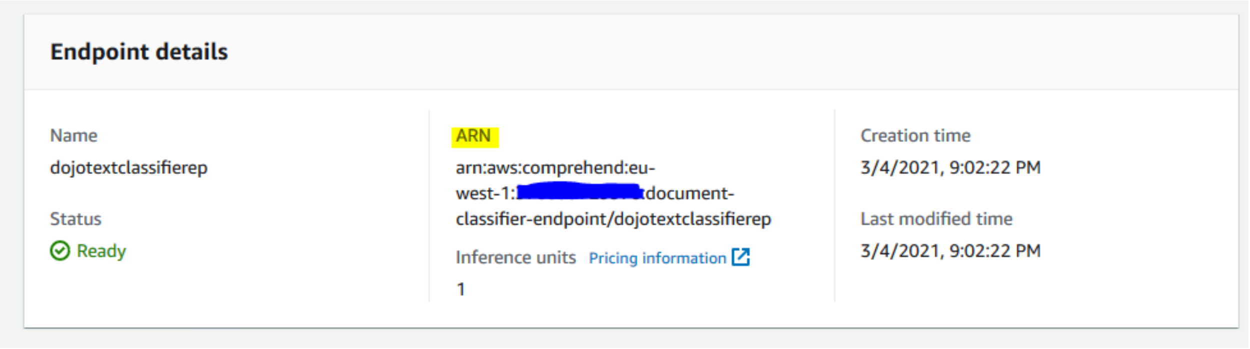


Endpoints (1)

Use endpoints to gain real-time insights.

	Name	Creation time	Inference units	Status
○	dojotextclassifierep	3/4/2021, 9:02:22 PM	1	✓ Ready

9. Once the endpoint is ready, make note of the endpoint ARN as you need it later in the workshop.



Endpoint details

Name
dojotextclassifierep

Status
✓ Ready

ARN
arn:aws:comprehend:eu-west-1:[redacted]document-classifier-endpoint/dojotextclassifierep

Inference units Pricing information
1

Creation time
3/4/2021, 9:02:22 PM

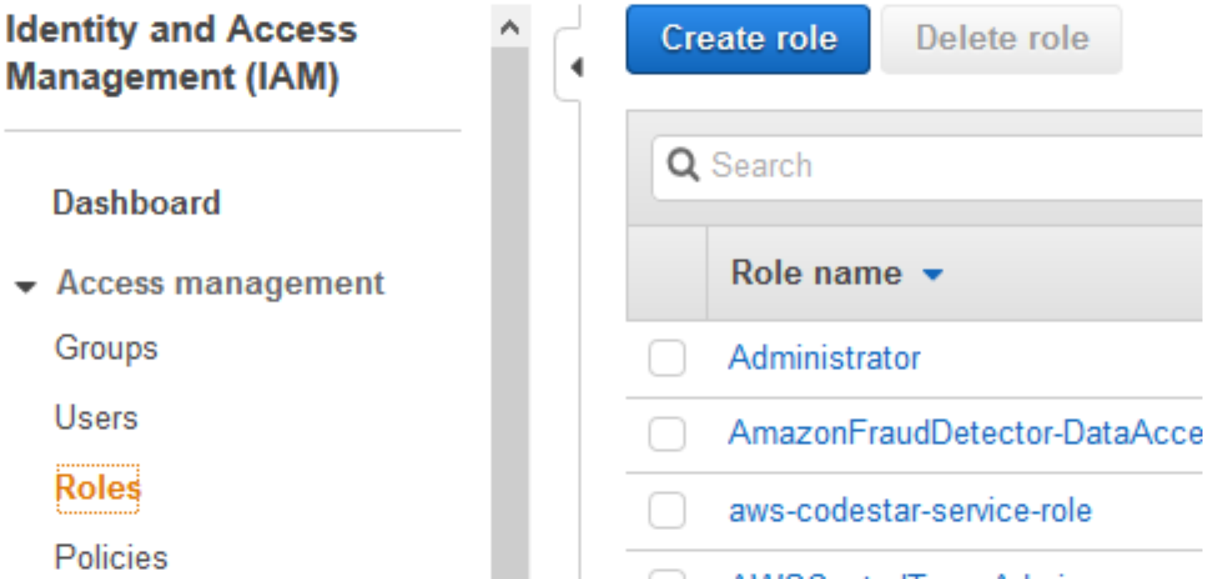
Last modified time
3/4/2021, 9:02:22 PM

10. The model and endpoint are ready. Create SageMaker Notebook Instance in the next step which you use to call the endpoint.

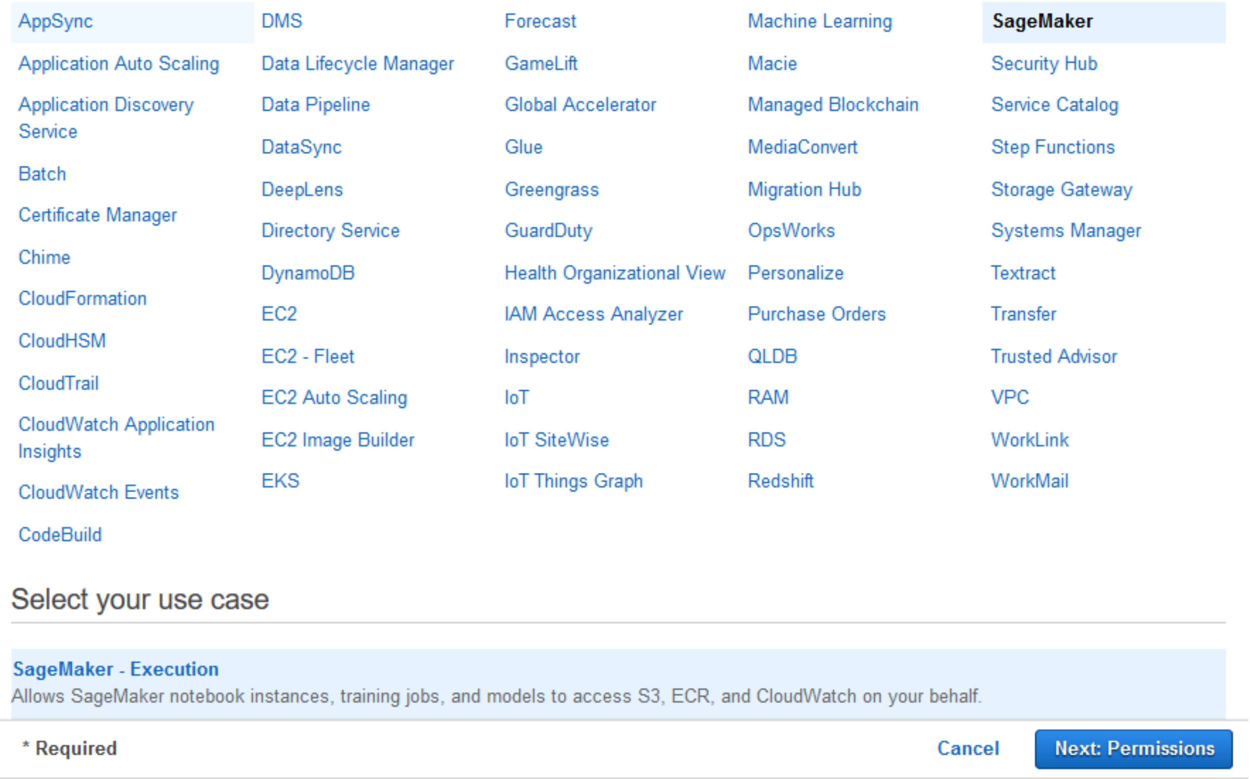
4: Configure SageMaker Notebook

The workshop will use Amazon SageMaker Notebook as the client to call classifier endpoint to check if the news title text is Real or Fake. Before you launch the SageMaker Notebook, you create an IAM role which SageMaker Notebook will use to authorize call to the Amazon Comprehend.

1. Goto the IAM Management console, click on the Roles menu in the left and then click on the Create role button.



2. On the next screen, select SageMaker as the service and click on the Next: Permissions button.

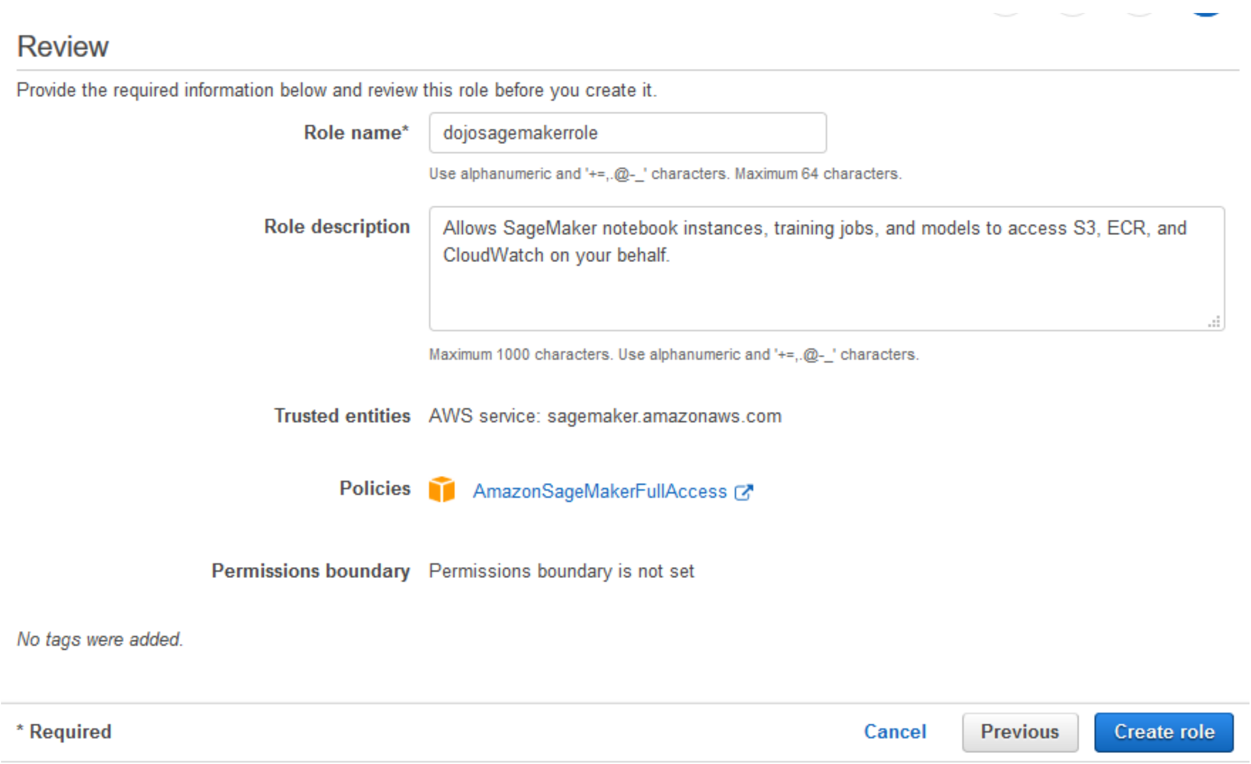


Select your use case

SageMaker - Execution
Allows SageMaker notebook instances, training jobs, and models to access S3, ECR, and CloudWatch on your behalf.

* Required

3. On the next screen, click on the Next: Tags button.
4. On the next screen, click on the Next: Review button.
5. On the next screen, type in **dojosagemakerrole** as the role name and click on the Create role button.



Review

Provide the required information below and review this role before you create it.

Role name* dojosagemakerrole

Use alphanumeric and '+=,.@-_' characters. Maximum 64 characters.

Role description Allows SageMaker notebook instances, training jobs, and models to access S3, ECR, and CloudWatch on your behalf.

Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.

Trusted entities AWS service: sagemaker.amazonaws.com

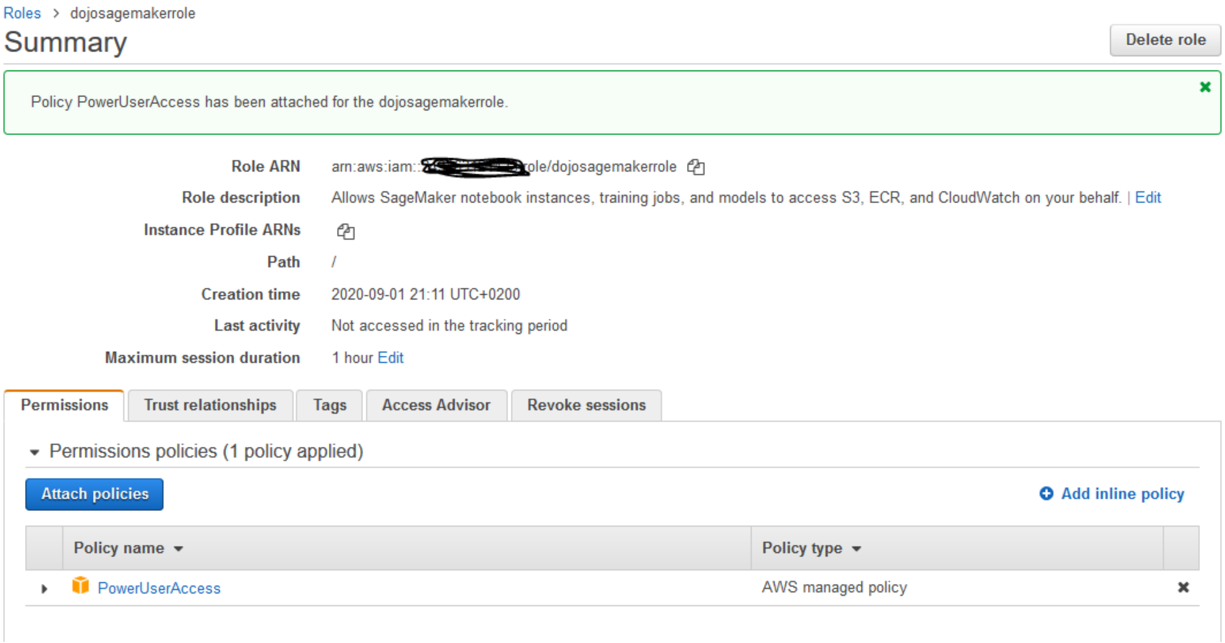
Policies AmazonSageMakerFullAccess

Permissions boundary Permissions boundary is not set

No tags were added.

* Required

6. The role is created in no time. Open the dojosagemakerrole role details, remove AmazonSageMakerFullAccess policy and attach PowerUserAccess policy to the role.



Roles > dojosagemakerrole

Summary

Policy PowerUserAccess has been attached for the dojosagemakerrole.

Role ARN arn:aws:iam::[redacted]ole/dojosagemakerrole

Role description Allows SageMaker notebook instances, training jobs, and models to access S3, ECR, and CloudWatch on your behalf. | Edit

Instance Profile ARNs

Path /

Creation time 2020-09-01 21:11 UTC+0200

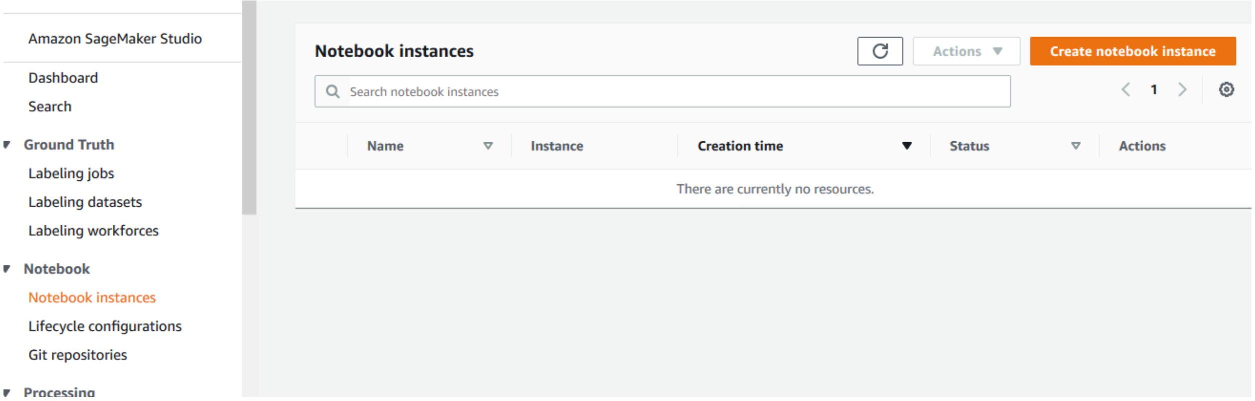
Last activity Not accessed in the tracking period

Maximum session duration 1 hour Edit

Permissions policies (1 policy applied)

Policy name	Policy type
PowerUserAccess	AWS managed policy

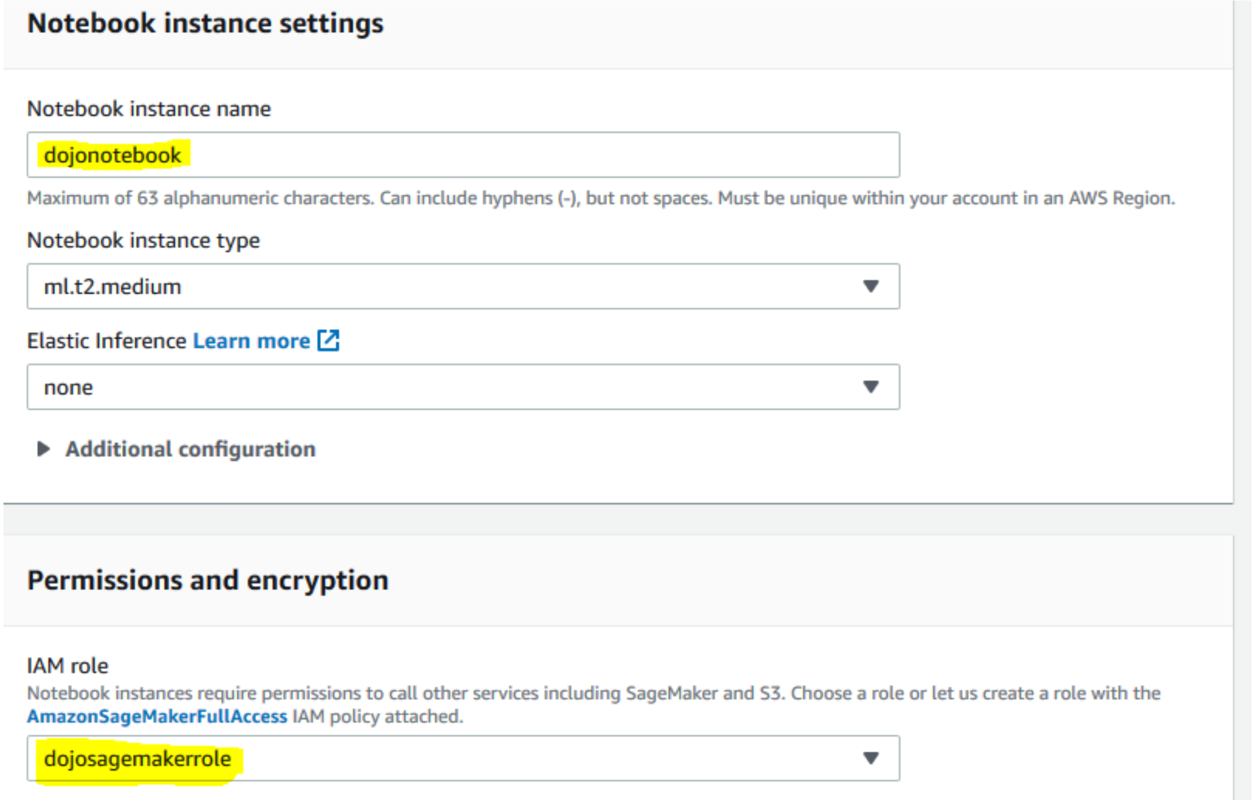
7. The role is ready. Let's launch the Notebook Instance next. Go to Amazon SageMaker console. Select Notebook instances in the left and then click on the Create notebook instance button.



Notebook instances

There are currently no resources.

8. On the next screen, type in dojonotebook as the notebook instance name, select **dojosagemakerrole** as the IAM role. Leave rest of the configuration as the default and click on the Create notebook instance button.



Notebook instance settings

Notebook instance name: dojonotebook

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

Notebook instance type: ml.t2.medium

Elastic Inference: none

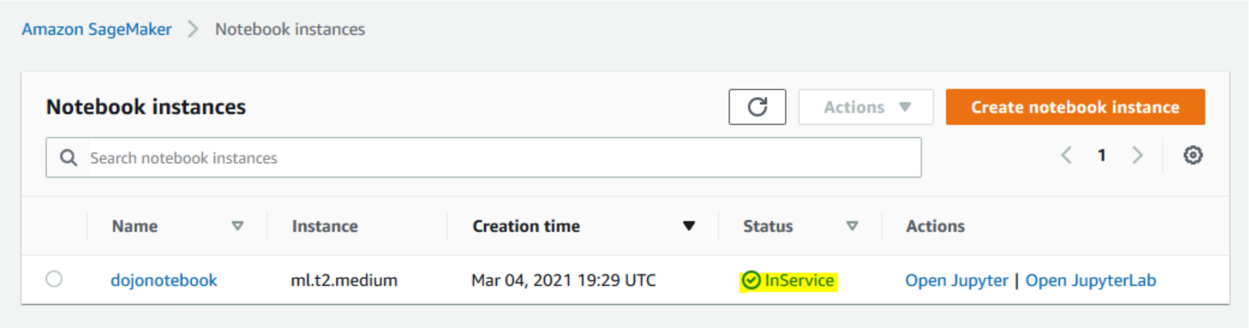
Permissions and encryption

IAM role

Notebook instances require permissions to call other services including SageMaker and S3. Choose a role or let us create a role with the AmazonSageMakerFullAccess IAM policy attached.

dojosagemakerrole

9. The notebook instance launch starts. Wait till the status changes to InService.


Amazon SageMaker > Notebook instances
Actions ▼
Create notebook instance
Search notebook instances
< 1 >

Notebook instances

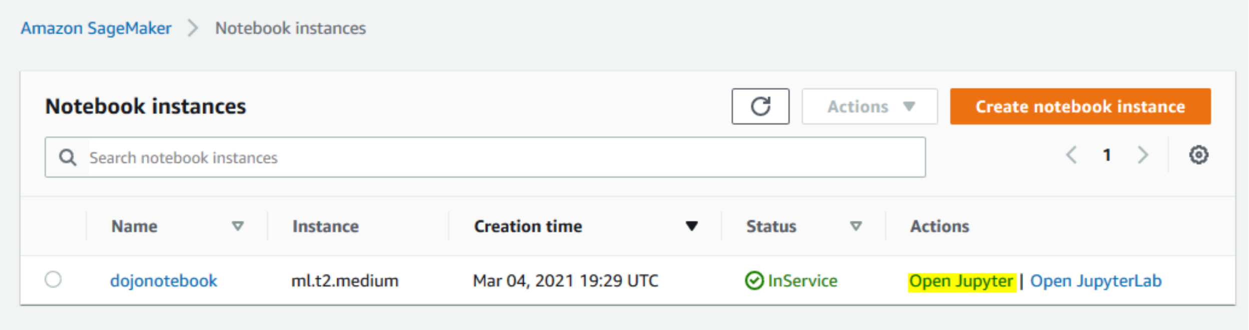
	Name	Instance	Creation time	Status	Actions
○	dojonotebook	ml.t2.medium	Mar 04, 2021 19:29 UTC	InService	Open Jupyter | Open JupyterLab

10. The notebook is ready. Let’s write client code for the text classification in the next step.

5: Create Client

The notebook instance is ready. You now write code which calls classifier endpoint to detect whether a text is Real or Fake.

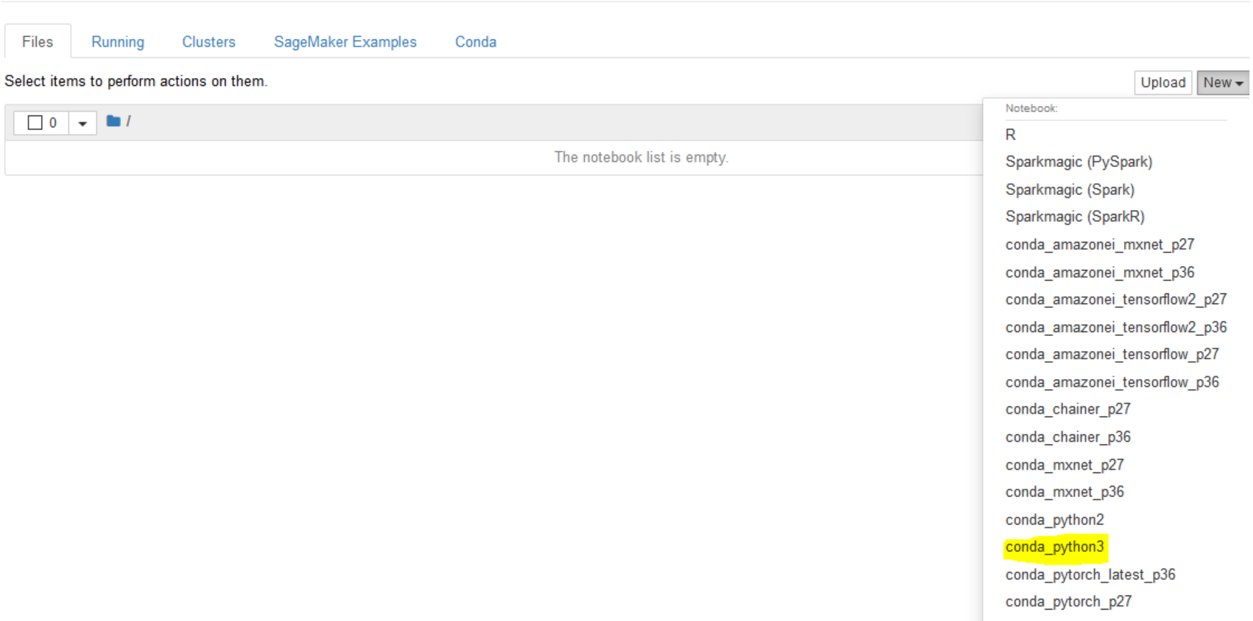
1. In the Amazon SageMaker console, select dojonotebook instance and click on the Open Jupyter link.


Amazon SageMaker > Notebook instances
Actions ▼
Create notebook instance
Search notebook instances
< 1 >

Notebook instances

	Name	Instance	Creation time	Status	Actions
○	dojonotebook	ml.t2.medium	Mar 04, 2021 19:29 UTC	InService	Open Jupyter | Open JupyterLab

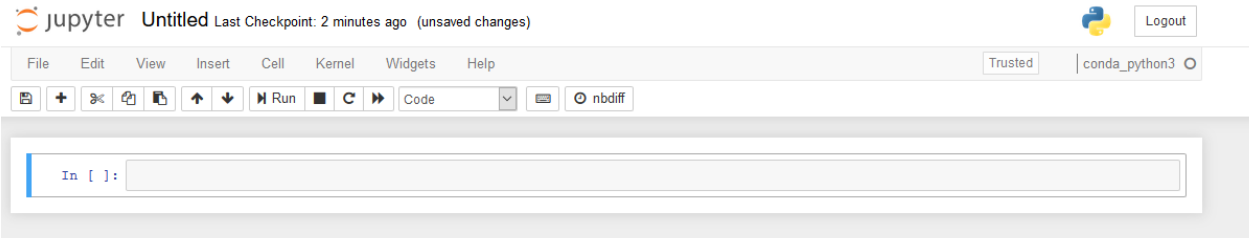
2. It will open Jupyter in a new browser tab or window. Select conda_python3 option under the New menu. Basically, you are starting a notebook with Python3. Such notebook also comes with Python Boto3 SDK deployed which will help in calling the endpoint.



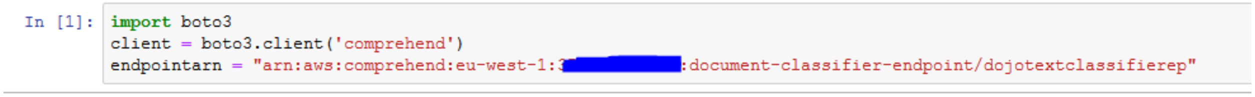
Select items to perform actions on them.

The notebook list is empty.

3. It will open a notebook in a new browser tab or window.



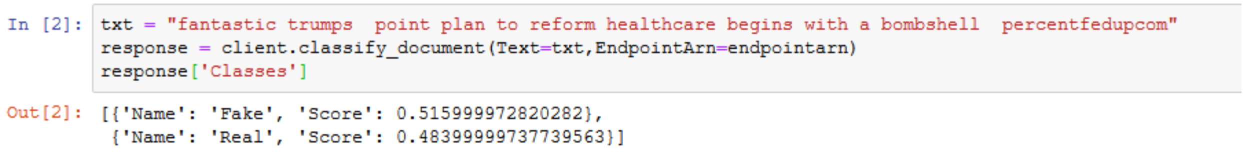
4. Copy-paste and run the following code in the notebook to import boto3 module and initiate Amazon Comprehend client. Replace {ENDPOINT_ARN} with classifier endpoint ARN you make note of in the previous step.



```
In [1]: import boto3
        client = boto3.client('comprehend')
        endpointarn = "arn:aws:comprehend:eu-west-1:[redacted]:document-classifier-endpoint/dojotextclassifierep"
```

```
import boto3
client = boto3.client('comprehend')
endpointarn = "{ENDPOINT_ARN}"
```

5. Copy-paste and run the following code in the notebook to classify a news title. It is calling classify_document method passing text and Endpoint ARN as parameters. The code then prints the classification. You can see the model predicts the news title text to be Fake with 52% confidence and real with 48% confidence. It is not a very confident prediction. Let's use another example.

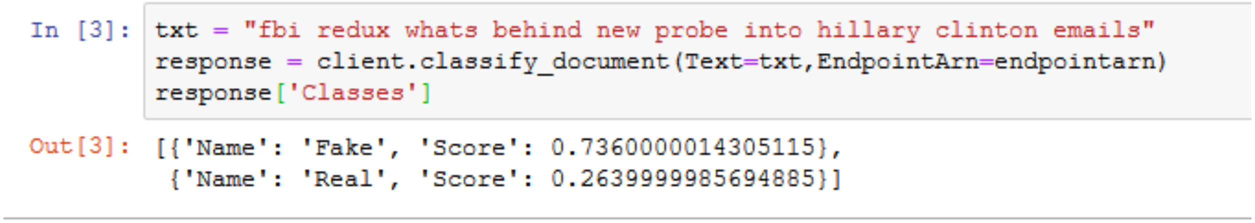


```
In [2]: txt = "fantastic trumps  point plan to reform healthcare begins with a bombshell  percentfedupcom"
        response = client.classify_document(Text=txt,EndpointArn=endpointarn)
        response['Classes']

Out[2]: [{'Name': 'Fake', 'Score': 0.515999972820282},
         {'Name': 'Real', 'Score': 0.48399999737739563}]
```

```
txt = "fantastic trumps  point plan to reform healthcare begins with a bombshell percentfedupcom"
response = client.classify_document(Text=txt,EndpointArn=endpointarn)
response['Classes']
```

6. Copy-paste and run the following code in the notebook to classify another news title. You can see the model predicts this title to be Fake with 73% confidence.



```
In [3]: txt = "fbi redux whats behind new probe into hillary clinton emails"
        response = client.classify_document(Text=txt,EndpointArn=endpointarn)
        response['Classes']

Out[3]: [{'Name': 'Fake', 'Score': 0.7360000014305115},
         {'Name': 'Real', 'Score': 0.2639999985694885}]
```

```
txt = "fbi redux whats behind new probe into hillary clinton emails"
response = client.classify_document(Text=txt,EndpointArn=endpointarn)
response['Classes']
```

7. The workshop finishes here. Goto the next task to clean-up the resources so that you don’t incur any cost post the workshop.

6. Clean up

1. Delete the dojonotebook notebook instance.
2. Delete the dojosagemakerrole IAM role.
3. Delete the dojo-text-records S3 bucket. If you created bucket with a different name then delete that one.
4. Delete dojotextclassifierep endpoint and dojotextclassifier classifier in the Amazon Comprehend.

Congratulations. You have successfully completed the workshop. Hope you enjoyed it.