Plagiarism Detection Model

Now that you've created training and test data, you are ready to define and train a model. Your goal in this notebook, will be to train a binary classification model that learns to label an answer file as either plagiarized or not, based on the features you provide the model.

This task will be broken down into a few discrete steps:

- Upload your data to S3.
- Define a binary classification model and a training script.
- Train your model and deploy it.
- Evaluate your deployed classifier and answer some questions about your approach.

To complete this notebook, you'll have to complete all given exercises and answer all the questions in this notebook.

> All your tasks will be clearly labeled **EXERCISE** and questions as **QUESTION**.

It will be up to you to explore different classification models and decide on a model that gives you the best performance for this dataset.

Load Data to S3

In the last notebook, you should have created two files: a `training.csv` and `test.csv` file with the features and class labels for the given corpus of plagiarized/non-plagiarized text data.

> The below cells load in some AWS SageMaker libraries and creates a default bucket. After creating this bucket, you can upload your locally stored data to S3.

Save your train and test `.csv` feature files, locally. To do this you can run the second notebook "2_Plagiarism_Feature_Engineering" in SageMaker or you can manually upload your files to this notebook using the upload icon in Jupyter Lab. Then you can upload local files to S3 by using `sagemaker_session.upload_data` and pointing directly to where the training data is saved.

In [1]:

```python
import pandas as pd
import boto3
import sagemaker
```

In [2]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
# session and role
sagemaker_session = sagemaker.Session()
role = sagemaker.get_execution_role()

# create an S3 bucket
bucket = sagemaker_session.default_bucket()
```

EXERCISE: Upload your training data to S3

Specify the `data_dir` where you've saved your `train.csv` file. Decide on a descriptive `prefix` that defines where your data will be uploaded in the default S3 bucket. Finally, create a pointer to your training data by calling `sagemaker_session.upload_data` and passing in the required parameters. It may help to look at the Session documentation (https://sagemaker.readthedocs.io/en/stable/session.html#sagemaker.session.Session.upload_data) or previous SageMaker code examples.

You are expected to upload your entire directory. Later, the training script will only access the `train.csv` file.

In [3]:

```
# should be the name of directory you created to save your features data
data_dir = 'plagiarism_data'

# set prefix, a descriptive name for a directory
prefix = 'plagiarism-detection'

# upload all data to S3
input_data = sagemaker_session.upload_data(path=data_dir, bucket=bucket, key_prefix=prefix)
print(input_data)
```

```
s3://sagemaker-eu-central-1-174885060101/plagiarism-detection
```

Test cell

Test that your data has been successfully uploaded. The below cell prints out the items in your S3 bucket and will throw an error if it is empty. You should see the contents of your `data_dir` and perhaps some checkpoints. If you see any other files listed, then you may have some old model files that you can delete via the S3 console (though, additional files shouldn't affect the performance of model developed in this notebook).

In [4]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
# confirm that data is in S3 bucket
empty_check = []
for obj in boto3.resource('s3').Bucket(bucket).objects.all():
    empty_check.append(obj.key)
    print(obj.key)

assert len(empty_check) !=0, 'S3 bucket is empty.'
print('Test passed!')
```

```
plagiarism-detection/test.csv
plagiarism-detection/train.csv
Test passed!
```

Modeling

Now that you've uploaded your training data, it's time to define and train a model!

The type of model you create is up to you. For a binary classification task, you can choose to go one of three routes:

- Use a built-in classification algorithm, like LinearLearner.
- Define a custom Scikit-learn classifier, a comparison of models can be found [here (https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html)](https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html).
- Define a custom PyTorch neural network classifier.

It will be up to you to test out a variety of models and choose the best one. Your project will be graded on the accuracy of your final model.

EXERCISE: Complete a training script

To implement a custom classifier, you'll need to complete a `train.py` script. You've been given the folders `source_sklearn` and `source_pytorch` which hold starting code for a custom Scikit-learn model and a PyTorch model, respectively. Each directory has a `train.py` training script. To complete this project **you only need to complete one of these scripts**; the script that is responsible for training your final model.

A typical training script:

- Loads training data from a specified directory
- Parses any training & model hyperparameters (ex. nodes in a neural network, training epochs, etc.)
- Instantiates a model of your design, with any specified hyperparams
- Trains that model
- Finally, saves the model so that it can be hosted/deployed, later

Defining and training a model

Much of the training script code is provided for you. Almost all of your work will be done in the `if __name__ == '__main__':` section. To complete a `train.py` file, you will:

1. Import any extra libraries you need
2. Define any additional model training hyperparameters using `parser.add_argument`
3. Define a model in the `if __name__ == '__main__':` section
4. Train the model in that same section

Below, you can use `!pygmentize` to display an existing `train.py` file. Read through the code; all of your tasks are marked with `TODO` comments.

Note: If you choose to create a custom PyTorch model, you will be responsible for defining the model in the `model.py` file, and a `predict.py` file is provided. If you choose to use Scikit-learn, you only need a `train.py` file; you may import a classifier from the `sklearn` library.

In [21]:

```
# directory can be changed to: source_sklearn or source_pytorch
!pygmentize source_sklearn/train.py
```

```
    # Read in csv training file
    training_dir = args.data_dir
    train_data = pd.read_csv(os.path.join(training_dir, "train.csv"), head
er=None, names=None)

    # Labels are in the first column
    train_y = train_data.iloc[:,0]
    train_x = train_data.iloc[:,1:]


    ## --- Your code here --- ##


    ## TODO: Define a model
    model = LinearSVC()


    ## TODO: Train the model
```

Provided code

If you read the code above, you can see that the starter code includes a few things:

- Model loading (`model_fn`) and saving code
- Getting SageMaker's default hyperparameters
- Loading the training data by name, `train.csv` and extracting the features and labels, `train_x` , and `train_y`

If you'd like to read more about model saving with joblib for sklearn (https://scikit-learn.org/stable/modules/model_persistence.html) or with torch.save (https://pytorch.org/tutorials/beginner/saving_loading_models.html), click on the provided links.

Create an Estimator

When a custom model is constructed in SageMaker, an entry point must be specified. This is the Python file which will be executed when the model is trained; the `train.py` function you specified above. To run a custom training script in SageMaker, construct an estimator, and fill in the appropriate constructor arguments:

- **entry_point**: The path to the Python script SageMaker runs for training and prediction.
- **source_dir**: The path to the training script directory `source_sklearn` OR `source_pytorch` .
- **entry_point**: The path to the Python script SageMaker runs for training and prediction.
- **source_dir**: The path to the training script directory `train_sklearn` OR `train_pytorch` .
- **entry_point**: The path to the Python script SageMaker runs for training.
- **source_dir**: The path to the training script directory `train_sklearn` OR `train_pytorch` .
- **role**: Role ARN, which was specified, above.
- **train_instance_count**: The number of training instances (should be left at 1).
- **train_instance_type**: The type of SageMaker instance for training. Note: Because Scikit-learn does not natively support GPU training, Sagemaker Scikit-learn does not currently support training on GPU instance types.
- **sagemaker_session**: The session used to train on Sagemaker.
- **hyperparameters** (optional): A dictionary `{'name':value, ..}` passed to the train function as hyperparameters.

Note: For a PyTorch model, there is another optional argument **framework_version**, which you can set to the latest version of PyTorch, `1.0` .

EXERCISE: Define a Scikit-learn or PyTorch estimator

To import your desired estimator, use one of the following lines:

```
from sagemaker.sklearn.estimator import SKLearn
```

```
from sagemaker.pytorch import PyTorch
```

In [22]:

```
# your import and estimator code, here
from sagemaker.sklearn.estimator import SKLearn

estimator = SKLearn(entry_point="train.py",
                    source_dir="source_sklearn",
                    role=role,
                    train_instance_count=1,
                    train_instance_type='ml.m4.xlarge')
```

EXERCISE: Train the estimator

Train your estimator on the training data stored in S3. This should create a training job that you can monitor in your SageMaker console.

In [23]:

```
%%time

# Train your estimator on S3 training data
estimator.fit({'train': input_data})
```

```
SM_MODEL_DIR=/opt/ml/model
SM_MODULE_DIR=s3://sagemaker-eu-central-1-174885060101/sagemaker-scikit-le
arn-2020-05-21-18-23-57-115/source/sourcedir.tar.gz
SM_TRAINING_ENV={"additional_framework_parameters":{},"channel_input_dir
s":{"train":"/opt/ml/input/data/train"},"current_host":"algo-1","framework
_module":"sagemaker_sklearn_container.training:main","hosts":["algo-1"],"h
yperparameters":{},"input_config_dir":"/opt/ml/input/config","input_data_c
onfig":{"train":{"RecordWrapperType":"None","S3DistributionType":"FullyRep
licated","TrainingInputMode":"File"}},"input_dir":"/opt/ml/input","is_mast
er":true,"job_name":"sagemaker-scikit-learn-2020-05-21-18-23-57-115","log_
level":20,"master_hostname":"algo-1","model_dir":"/opt/ml/model","module_d
ir":"s3://sagemaker-eu-central-1-174885060101/sagemaker-scikit-learn-2020-
05-21-18-23-57-115/source/sourcedir.tar.gz","module_name":"train","network
_interface_name":"eth0","num_cpus":4,"num_gpus":0,"output_data_dir":"/opt/
ml/output/data","output_dir":"/opt/ml/output","output_intermediate_dir":"/
opt/ml/output/intermediate","resource_config":{"current_host":"algo-1","ho
sts":["algo-1"],"network_interface_name":"eth0"},"user_entry_point":"trai
n.py"}
SM_USER_ARGS=[]
SM_OUTPUT_INTERMEDIATE_DIR=/opt/ml/output/intermediate
```

EXERCISE: Deploy the trained model

After training, deploy your model to create a `predictor`. If you're using a PyTorch model, you'll need to create a trained `PyTorchModel` that accepts the trained `<model>.model_data` as an input parameter and points to the provided `source_pytorch/predict.py` file as an entry point.

To deploy a trained model, you'll use `<model>.deploy`, which takes in two arguments:

- **initial_instance_count**: The number of deployed instances (1).
- **instance_type**: The type of SageMaker instance for deployment.

Note: If you run into an instance error, it may be because you chose the wrong training or deployment instance_type. It may help to refer to your previous exercise code to see which types of instances we used.

In [24]:

```
%%time

# uncomment, if needed
# from sagemaker.pytorch import PyTorchModel


# deploy your model to create a predictor
predictor = estimator.deploy(initial_instance_count=1, instance_type='ml.t2.medium')
```

```
-----------------!CPU times: user 292 ms, sys: 6.07 ms, total: 298 ms
Wall time: 8min 32s
```

Evaluating Your Model

Once your model is deployed, you can see how it performs when applied to our test data.

The provided cell below, reads in the test data, assuming it is stored locally in `data_dir` and named `test.csv` . The labels and features are extracted from the `.csv` file.

In [16]:

```
"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
import os

# read in test data, assuming it is stored locally
test_data = pd.read_csv(os.path.join(data_dir, "test.csv"), header=None, names=None)

# labels are in the first column
test_y = test_data.iloc[:,0]
test_x = test_data.iloc[:,1:]
```

EXERCISE: Determine the accuracy of your model

Use your deployed `predictor` to generate predicted, class labels for the test data. Compare those to the *true* labels, `test_y` , and calculate the accuracy as a value between 0 and 1.0 that indicates the fraction of test data that your model classified correctly. You may use sklearn.metrics (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics) for this calculation.

To pass this project, your model should get at least 90% test accuracy.

In [25]:

```
# First: generate predicted, class labels
test_y_preds = predictor.predict(test_x)


"""
DON'T MODIFY ANYTHING IN THIS CELL THAT IS BELOW THIS LINE
"""
# test that your model generates the correct number of labels
assert len(test_y_preds)==len(test_y), 'Unexpected number of predictions.'
print('Test passed!')
```

```
Test passed!
```

In [26]:

```
# Second: calculate the test accuracy
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report

accuracy = accuracy_score(test_y, test_y_preds)
print(accuracy)
target_names = ['Original', 'Plagiarism']
print(classification_report(test_y.values, test_y_preds, target_names=target_names))

## print out the array of predicted and true labels, if you want
print('\nPredicted class labels: ')
print(test_y_preds)
print('\nTrue class labels: ')
print(test_y.values)
```

```
0.96
              precision    recall  f1-score   support

    Original       0.91      1.00      0.95        10
  Plagiarism       1.00      0.93      0.97        15

   micro avg       0.96      0.96      0.96        25
   macro avg       0.95      0.97      0.96        25
weighted avg       0.96      0.96      0.96        25


Predicted class labels: 
[0 0 0 1 1 1 0 1 1 1 0 1 1 0 0 0 1 1 1 1 0 1 1 0 0]

True class labels: 
[0 0 0 1 1 1 0 1 1 1 0 1 1 1 0 0 1 1 1 1 0 1 1 0 0]
```

Question 1: How many false positives and false negatives did your model produce, if any? And why do you think this is?

** Answer**:

- K-Nearest Neighbor: 1 False positives, accuracy = 0.92
- Random Forest: 1 False positive and 1 false negative, accuracy = 0.88
- Linear Support Vector Classification: 1 False negative, accuracy = 0.96

All 3 models come up with nice results, since the data is not so big and only 3 features are taken into account

Question 2: How did you decide on the type of model to use?

** Answer**: After all, Linear Support Vector Classification scored the best on this binary classification problem. Linear SVC is also good at dealing with binary classification problem when the number of samples scales bigger.

EXERCISE: Clean up Resources

After you're done evaluating your model, **delete your model endpoint**. You can do this with a call to `.delete_endpoint()` . You need to show, in this notebook, that the endpoint was deleted. Any other resources, you may delete from the AWS console, and you will find more instructions on cleaning up all your resources, below.

In [27]:

```
predictor.delete_endpoint()
```

Deleting S3 bucket

When you are *completely* done with training and testing models, you can also delete your entire S3 bucket. If you do this before you are done training your model, you'll have to recreate your S3 bucket and upload your training data again.

In [28]:

```
# deleting bucket, uncomment lines below

bucket_to_delete = boto3.resource('s3').Bucket(bucket)
bucket_to_delete.objects.all().delete()
```

Out[28]:

```
[{'ResponseMetadata': {'RequestId': '3D4E876BF451F79F',
   'HostId': 'W714hQavAvYyzKHJ71oACFTmHmY1PS2mjPBrGnWmNCH+qGbpH6Z/CVfJQmB1DD
oAJ5NmktYnxgc=',
   'HTTPStatusCode': 200,
   'HTTPHeaders': {'x-amz-id-2': 'W714hQavAvYyzKHJ71oACFTmHmY1PS2mjPBrGnWmNC
H+qGbpH6Z/CVfJQmB1DDoAJ5NmktYnxgc=',
    'x-amz-request-id': '3D4E876BF451F79F',
    'date': 'Thu, 21 May 2020 18:41:45 GMT',
    'connection': 'close',
    'content-type': 'application/xml',
    'transfer-encoding': 'chunked',
    'server': 'AmazonS3'},
   'RetryAttempts': 0},
  'Deleted': [{'Key': 'sagemaker-scikit-learn-2020-05-21-18-00-42-368/debug-
output/training_job_end.ts'},
   {'Key': 'sagemaker-scikit-learn-2020-05-21-18-23-57-115/output/model.tar.
gz'},
   {'Key': 'plagiarism-detection/train.csv'},
   {'Key': 'sagemaker-scikit-learn-2020-05-21-17-51-59-425/source/sourcedir.
tar.gz'},
   {'Key': 'sagemaker-scikit-learn-2020-05-21-18-00-42-368/source/sourcedir.
tar.gz'},
   {'Key': 'plagiarism-detection/test.csv'},
   {'Key': 'sagemaker-scikit-learn-2020-05-21-18-23-57-115/debug-output/trai
ning_job_end.ts'},
   {'Key': 'sagemaker-scikit-learn-2020-05-21-17-55-58-279/source/sourcedir.
tar.gz'},
   {'Key': 'sagemaker-scikit-learn-2020-05-21-18-00-42-368/output/model.tar.
gz'},
   {'Key': 'sagemaker-scikit-learn-2020-05-21-18-23-57-115/source/sourcedir.
tar.gz'}]}]
```

Deleting all your models and instances

When you are *completely* done with this project and do **not** ever want to revisit this notebook, you can choose to delete all of your SageMaker notebook instances and models by following these instructions (https://docs.aws.amazon.com/sagemaker/latest/dg/ex1-cleanup.html). Before you delete this notebook instance, I recommend at least downloading a copy and saving it, locally.

Further Directions

There are many ways to improve or add on to this project to expand your learning or make this more of a unique project for you. A few ideas are listed below:

- Train a classifier to predict the *category* (1-3) of plagiarism and not just plagiarized (1) or not (0).
- Utilize a different and larger dataset to see if this model can be extended to other types of plagiarism.
- Use language or character-level analysis to find different (and more) similarity features.
- Write a complete pipeline function that accepts a source text and submitted text file, and classifies the submitted text as plagiarized or not.
- Use API Gateway and a lambda function to deploy your model to a web application.

These are all just options for extending your work. If you've completed all the exercises in this notebook, you've completed a real-world application, and can proceed to submit your project. Great job!