# Contents



# List of Figures

# PlagCaps: Prediction of Plagiarised Text on a Corpus Dataset using Deep Learning Algorithms

Prathmesh Shukla

x19231644

## 1 Introduction

The prerequisites for experimenting are detailed in this setup handbook. It also serves as the step-by-step guide for implementing the necessary code changes to conduct the experiment and obtain the desired results.

## 2 System Specification

The hardware and software settings on which the experiment was conducted are detailed below.

### 2.1 Hardware Specification

The experiment was conducted on the below h/w:

- CPU and Core - Intel i5-8250U CPU, 8th Generation @1.80 GHz
- RAM - 8.00 GB
- Operating System - 64 bit, Windows 10 OS

### 2.2 Software Specification

- Python - Python is a programming language that allows you to work faster and more effectively with your systems. [1]
- Anaconda Navigator - Anaconda Individual Edition is a free, easy-to-install package manager, environment manager, and Python distribution with a collection of 1,500+ open source packages with free community support. Anaconda is platform-agnostic [2].
- Google Colab - Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with Zero configuration required, Free access to GPUs, Easy sharing.

## 3 Installation of Anaconda Navigator Environment

- The software for Anaconda Navigator can be installed by following the steps mentioned in the link here: `https://docs.anaconda.com/anaconda/install/windows/`
- Once the installation is completed, verify if the installation is done using the link here: `https://docs.anaconda.com/anaconda/install/verify-install/`
- Once the installation is verified, create an environment for python so that it can be used in Jupyter Lab using the link here: `https://docs.anaconda.com/anaconda/navigator/tutorials/manage-environments/#creating-a-new-environment`
- Click on the Jupyter Lab icon present on the ANE to start Jupyter Lab.

[1] `https://www.python.org/about/`
[2] `https://docs.anaconda.com/`

# 4 Python Libraries and Packages used

Once the Anaconda Navigator is installed make sure the below libraries are available in the ANE as they are used as part of the solution -

- matplotlib
- scipy
- sklearn
- itertools
- pandas
- mpl toolkits
- !pip3 uninstall keras-nightly
- !pip3 uninstall -y tensorflow
- !pip3 install keras==2.1.6
- !pip3 install tensorflow==1.15.0
- !pip3 install h5py==2.10.0
- !pip install pipenv
- !pip install numpy==1.16.1

# 5 Python version

The python version used for this experiment was 3.8.5.

# 6 Download required files

The files required for running the experiment are available in the link here: `https://www.kaggle.com/c/quora-question-pairs/data`. For running the experiment make sure the below files are available in the local environment.

Glove, word embedding file is available at the below link: `https://www.kaggle.com/rtatman/glove-global-vectors-for-word-representation`

Once the files are downloaded and unzipped and available for use in the local environment, edit the code to point to these datasets which is elaborated in section 7

# 7 Steps to Execute the code

There are two approaches in which the experiment can be replicated namely -

1. Using Jupyter NB
2. Using Google Colab

Below is the elaboration for the same.

## 7.1 Approach 1 - Jupyter Lab Steps for execution

Before running the code the only requirement is to change the path where the file is being read. Below is the screenshot for places to change the source file path to the one where the user has stored the source files:

```
df_train = pd.read_csv("D:\\MScDataAnalytics\\Sem_3\\Research\\train.csv")
print(df_train.head(1))
```

```
   id  qid1  qid2                                                question1  \
0   0     1     2  What is the step by step guide to invest in sh...

                                            question2  is_duplicate
0  What is the step by step guide to invest in sh...             0
```

Figure 1: Jupyter NB - change path for Train Quora Question Pair file

```
df_test = pd.read_csv("D:\\MScDataAnalytics\\Sem_3\\Research\\test.csv")
print(df_test.head(1))
```

```
   test_id                                            question1  \
0        0  How does the Surface Pro himself 4 compare wit...

                                           question2
0  Why did Microsoft choose core m3 and not core ...
```

Figure 2: Jupyter NB - change path for Test Quora Question Pair file

## 7.2 Approach 2 - Google Colab Steps for execution

The only thing you need to do before executing the code is modify the location where the file is being read. The picture below shows where you may modify the source file path to the location where the user has saved the source files:

```
[3] df_train = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/train.csv")
    print(df_train.head(1))
```

```
   id  qid1  ...                                          question2 is_duplicate
0   0     1  ...  What is the step by step guide to invest in sh...            0

[1 rows x 6 columns]
```

Figure 3: Jupyter NB - change path for Train Quora Question Pair file

```
[4]  df_test = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/test.csv")
     print(df_test.head(1))

        test_id  ...                                          question2
     0        0  ...  Why did Microsoft choose core m3 and not core ...

     [1 rows x 3 columns]
```

Figure 4: Jupyter NB - change path for Test Quora Question Pair file

Once the code is executed the same can be run with a single step by clicking on the :

- For Jupyter NB: Click on Kernel button present on the ribbon of Jupyter NB and then choosing Restart and Run All from the drop down to execute all the cells in which the code is present.
- For Google Colab: Click on Runtime button present on the ribbon of Google Colab and then choosing Run All from the drop down to execute all the steps.

# 8 Evaluation

After executing all the cells in the code file, an accuracy and loss curve will be generated along with a confusion matrix. After the execution of the epoches, accuracy can be seen in the last epoch.Prediction can be made using y_pred and a given question pair is plagiarised or not can be checked.

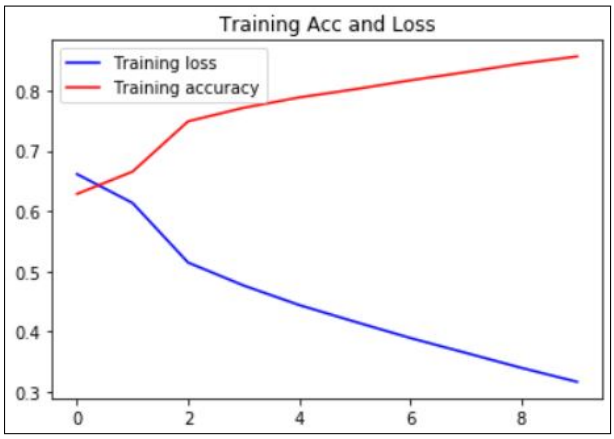


Figure 5: LSTM: Accuracy Vs Loss curve

In the figure 5, curve is representing the acc vs loss curve and one can say as the epoches were increased the learning rate and accuracy also increased and at the same time loss during the training got reduced to almost 0.1.