**Please follow the below instructions:**

1. Unzip the Assignment Folder.
2. The execution is dived into two parts:
3. Naïve Bayes - **NaiveBayes.py**
4. Logistic Regression – **LogisticRegressionWithStopWords.py, LogisticRegressionWithoutStopWords.py**
5. Open command prompt, change your directory to point to the path where the file has been extracted and Run the **NaiveBayes.py** files with the arguments as given below:

**>>> python NaiveBayes.py <train-ham-path> <train-spam-path> <test-ham-path> <test-spam-path>**

C:\Emkae\UTD\Machine Learning\Assignment\Text Classification\Assignment2>python NaiveBayes.py hw2\_train\train\ham hw2\_train\train\spam hw2\_test\test\ham hw2\_test\test\spam

Naive Bayes Classification without stop words:

Number of ham files successfully classified: 322

Number of ham files successfully not classified: 26

Number of spam files successfully classified: 129

Number of spam files successfully not classified: 1

The total accuracy is: 94.35146443514645

Naive Bayes Classification with stop words:

Number of ham files successfully classified: 325

Number of ham files successfully not classified: 23

Number of spam files successfully classified: 129

Number of spam files successfully not classified: 1

The total accuracy is: 94.97907949790795

**Note:** The **stopwords.txt** file and the data is also present in the folder and there is no need to explicitly mention the path of the stopwords file. Please make sure the **NaiveBayes.py** file and the **stopwords.txt** file are in same folder.

1. Open command prompt, change your directory to point to the path where the file has been extracted and Run the **LogisticRegressionWithStopWords.py** files with the arguments as given below:

**>>> python LogisticRegressionWithStopWords.py** **<traindata -path> <testdata-path>**

C:\Emkae\UTD\Machine Learning\Assignment\Text Classification\Assignment2>python LogisticRegressionWithStopWords.py hw2\_train/train hw2\_test/test

Logistic Regression with stop words:

Number of ham files classified successsfully : 325

Number of ham files not classified successsfully: 23

Number of spam files classified successsfully: 106

Number of spam files not classified successsfully: 24

Accuracy: 92.25941422594143

1. Open command prompt, change your directory to point to the path where the file has been extracted and Run the **LogisticRegressionWithoutStopWords.py** files with the arguments as given below:

**>>> python LogisticRegressionWithoutStopWords.py** **<traindata -path> <testdata-path>**

C:\Emkae\UTD\Machine Learning\Assignment\Text Classification\Assignment2>python LogisticRegressionWithoutStopWords.py hw2\_train/train hw2\_test/test

Logistic Regression without stop words:

Number of ham files classified successsfully : 326

Ham files not classified successsfully: 22

Number of spam files classified successsfully: 119

Spam files not classified successsfully: 11

Accuracy: 95.18828451882845

1. Accuracy Improvements:
2. Naïve Bayes: There is not so significant difference in the accuracies of NB with and without stopwords. There can be two reasons for this:
3. Stop words were not collected properly and hence removal does not affect the accuracy.
4. Stop words are equal likely to be present in both ham and spam frequently and does not help in classifying one from other. Most of the stop words are present equal likely in both spam and ham, hence removing them do not have much impact.
5. Logistic Regression: