NLP Assignment 2

Name: Vamsi Tallam   UIN: 432001932

Naïve Bayes Classifier:

How to run the code:

1. Download and extract the code from SentimentAnalyzer_Fall22.zip
2. Navigate to the python folder and run the NaiveBayes.py
   (C:\Users\vamsi\OneDrive\Desktop\SentimentAnalyzer_Fall22\SentimentAnalyzer_Fall22\python)
3. Using Spyder, we can run the code using the following command:
   Part1: Simple Naïve Bayes Classifier without ignoring stop words and non-binary
   runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python/NaiveBayes.py', args='../data/imdb1',
   wdir='C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')

   Part2: Simple Naïve Bayes Classifier after ignoring stopwords
   runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python/NaiveBayes.py', args='-f ../data/imdb1',
   wdir='C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')

   Part3: binarized version of the multinomial Naïve Bayes Classifier
   runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python/NaiveBayes.py', args='-b ../data/imdb1',
   wdir='C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')

Results and Analysis:
- The average accuracy of the Naïve Bayes classifier is about 81.65% using 10-fold cross-validation. The lowest observed accuracy is about 74.5% and the highest is about 84.5%.
- The average accuracy of the Naïve Bayes classifier after removing stop words is about 80.1% using 10-fold cross-validation. The lowest observed accuracy is about 74.5% and the highest is about 82.5%.
  - Decrease in accuracy can be due to the removal of some essential stop words which are removed based on your stop word list.
- The average accuracy of the binarized version of Naïve Bayes classifier is about 82.95% using 10-fold cross-validation. The lowest observed accuracy is about 78.5% and the highest is about 86%.
  - The binarized version of Naïve Bayes classifier performs better as the occurrence of the word is more meaningful than the frequency of occurrence. This idea is reinforced from the request shown below.

```
In [58]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python/
NaiveBayes.py', args='../data/imdb1', wdir='C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/
SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.755000
[INFO]  Fold 1 Accuracy: 0.845000
[INFO]  Fold 2 Accuracy: 0.845000
[INFO]  Fold 3 Accuracy: 0.845000
[INFO]  Fold 4 Accuracy: 0.815000
[INFO]  Fold 5 Accuracy: 0.825000
[INFO]  Fold 6 Accuracy: 0.830000
[INFO]  Fold 7 Accuracy: 0.820000
[INFO]  Fold 8 Accuracy: 0.745000
[INFO]  Fold 9 Accuracy: 0.840000
[INFO]  Accuracy: 0.816500

In [59]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python/
NaiveBayes.py', args='-f ../data/imdb1', wdir='C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/
SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.745000
[INFO]  Fold 1 Accuracy: 0.815000
[INFO]  Fold 2 Accuracy: 0.825000
[INFO]  Fold 3 Accuracy: 0.810000
[INFO]  Fold 4 Accuracy: 0.790000
[INFO]  Fold 5 Accuracy: 0.825000
[INFO]  Fold 6 Accuracy: 0.820000
[INFO]  Fold 7 Accuracy: 0.810000
[INFO]  Fold 8 Accuracy: 0.750000
[INFO]  Fold 9 Accuracy: 0.820000
[INFO]  Accuracy: 0.801000

In [60]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python/
NaiveBayes.py', args='-b ../data/imdb1', wdir='C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/
SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.805000
[INFO]  Fold 1 Accuracy: 0.835000
[INFO]  Fold 2 Accuracy: 0.840000
[INFO]  Fold 3 Accuracy: 0.815000
[INFO]  Fold 4 Accuracy: 0.830000
[INFO]  Fold 5 Accuracy: 0.820000
[INFO]  Fold 6 Accuracy: 0.855000
[INFO]  Fold 7 Accuracy: 0.850000
[INFO]  Fold 8 Accuracy: 0.785000
[INFO]  Fold 9 Accuracy: 0.860000
[INFO]  Accuracy: 0.829500
```

Known Issues and Limitations:

- Assuming that the features are independent i.e, the occurrence of a word is not dependent on the previous word. This is not the case in our real world scenarios.

Perceptron Classifier:

How to run the code:

1. Download and extract the code from SentimentAnalyzer_Fall22.zip

2. Navigate to the python folder and run the Perceptron.py
   (C:\Users\vamsi\OneDrive\Desktop\SentimentAnalyzer_Fall22\SentimentAnalyzer_Fall22\python)

3. Using Spyder, we can run the code using the following command:
   We have to change the args for changing the iterations
   runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python/Perceptron.py', args='../data/imdb1 100',
   wdir='C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')

Results and Analysis:

- Implemented parameter averaging as described in section 2.1.1 of Hal Daume's thesis (http://users.umiacs.umd.edu/~hal/docs/daume06thesis.pdf).

- The average accuracy of the perceptron classifier is about 71.65% using 10-fold cross-validation and with #iterations as 1. While using #iterations as 10 and 50 is 82.4% and 84.2% respectively. Similarly the results for 100 iterations and 1000 iterations can be seen below.

- Increasing the number of iterations will increase the accuracy but after 50 iterations the increase is not significant. This may be due to the fact that we achieve convergence around 50 and increasing any further does not the performance.

- Using np.shuffle greatly improved the accuracy by introducing randomness and greater chance to look at the points.

```
In [61]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/
python/Perceptron.py', args='../data/imdb1 1', wdir='C:/Users/vamsi/OneDrive/Desktop/
SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.730000
[INFO]  Fold 1 Accuracy: 0.720000
[INFO]  Fold 2 Accuracy: 0.710000
[INFO]  Fold 3 Accuracy: 0.700000
[INFO]  Fold 4 Accuracy: 0.730000
[INFO]  Fold 5 Accuracy: 0.735000
[INFO]  Fold 6 Accuracy: 0.735000
[INFO]  Fold 7 Accuracy: 0.655000
[INFO]  Fold 8 Accuracy: 0.765000
[INFO]  Fold 9 Accuracy: 0.685000
[INFO]  Accuracy: 0.716500

In [62]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/
python/Perceptron.py', args='../data/imdb1 10', wdir='C:/Users/vamsi/OneDrive/Desktop/
SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.820000
[INFO]  Fold 1 Accuracy: 0.850000
[INFO]  Fold 2 Accuracy: 0.825000
[INFO]  Fold 3 Accuracy: 0.835000
[INFO]  Fold 4 Accuracy: 0.780000
[INFO]  Fold 5 Accuracy: 0.820000
[INFO]  Fold 6 Accuracy: 0.830000
[INFO]  Fold 7 Accuracy: 0.825000
[INFO]  Fold 8 Accuracy: 0.815000
[INFO]  Fold 9 Accuracy: 0.840000
[INFO]  Accuracy: 0.824000

In [63]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/
python/Perceptron.py', args='../data/imdb1 50', wdir='C:/Users/vamsi/OneDrive/Desktop/
SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.820000
[INFO]  Fold 1 Accuracy: 0.855000
[INFO]  Fold 2 Accuracy: 0.840000
[INFO]  Fold 3 Accuracy: 0.845000
[INFO]  Fold 4 Accuracy: 0.835000
[INFO]  Fold 5 Accuracy: 0.860000
[INFO]  Fold 6 Accuracy: 0.855000
[INFO]  Fold 7 Accuracy: 0.825000
[INFO]  Fold 8 Accuracy: 0.815000
[INFO]  Fold 9 Accuracy: 0.870000
[INFO]  Accuracy: 0.842000
```

```
In [64]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/
python/Perceptron.py', args='../data/imdb1 100', wdir='C:/Users/vamsi/OneDrive/Desktop/
SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.805000
[INFO]  Fold 1 Accuracy: 0.820000
[INFO]  Fold 2 Accuracy: 0.800000
[INFO]  Fold 3 Accuracy: 0.845000
[INFO]  Fold 4 Accuracy: 0.830000
[INFO]  Fold 5 Accuracy: 0.845000
[INFO]  Fold 6 Accuracy: 0.875000
[INFO]  Fold 7 Accuracy: 0.825000
[INFO]  Fold 8 Accuracy: 0.840000
[INFO]  Fold 9 Accuracy: 0.865000
[INFO]  Accuracy: 0.835000

In [65]: runfile('C:/Users/vamsi/OneDrive/Desktop/SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/
python/Perceptron.py', args='../data/imdb1 1000', wdir='C:/Users/vamsi/OneDrive/Desktop/
SentimentAnalyzer_Fall22/SentimentAnalyzer_Fall22/python')
[INFO]  Fold 0 Accuracy: 0.815000
[INFO]  Fold 1 Accuracy: 0.835000
[INFO]  Fold 2 Accuracy: 0.835000
[INFO]  Fold 3 Accuracy: 0.865000
[INFO]  Fold 4 Accuracy: 0.850000
[INFO]  Fold 5 Accuracy: 0.855000
[INFO]  Fold 6 Accuracy: 0.860000
[INFO]  Fold 7 Accuracy: 0.830000
[INFO]  Fold 8 Accuracy: 0.835000
[INFO]  Fold 9 Accuracy: 0.855000
[INFO]  Accuracy: 0.843500
```

Known Issues and Limitations:

- Running time greatly increases with an increase in the number of iterations. There is no significant increase in using higher iterations. Hence limiting the number of iterations to 50 is optimal for our data set.