Name: Naveenraj Palanisamy
NetId: NXP154130
Machine Learning Assignment (CS 6375.001)-(Naive bayes and logistic Regression)
(Note: Runs in python 3.5)
(Note: File Path should be full path)

Zip file name: NXP154130_ML_Assignment2.zip
Folder Name: NXP154130_ML_Assignment2

Folder Structure:

This PC > OS (C:) > NXP154130_ML_Assignment2

Name	Date modified	Type	Size
__pycache__	3/16/2016 2:19 AM	File folder	
2	3/16/2016 2:25 AM	File folder	
3	3/16/2016 2:25 AM	File folder	
.project	3/8/2016 10:38 PM	PROJECT File	1 KB
.pydevproject	3/8/2016 10:38 PM	PYDEVPROJECT File	1 KB
data	3/9/2016 2:03 PM	File	2 KB
data	3/9/2016 9:01 PM	Text Document	0 KB
FileParsing	3/12/2016 11:49 PM	Python File	6 KB
FileParsing_stop_word	3/12/2016 11:50 PM	Python File	7 KB
FileParsing_test	3/12/2016 11:50 PM	Python File	6 KB
FileParsing_test_stop_words	3/12/2016 11:51 PM	Python File	7 KB
LogisticRegression	3/11/2016 10:24 A...	Python File	2 KB
Main_class	3/13/2016 11:13 A...	Python File	3 KB
NaiveBayes	3/10/2016 10:45 PM	Python File	2 KB
Stop_words	3/13/2016 1:10 AM	Python File	1 KB
STOP_WORDS	3/10/2016 4:50 PM	Text Document	10 KB

<u>Main_Class.py is the python file need to run.</u>

<u>Input and output files are with in folder name 2.</u>

This PC > OS (C:) > NXP154130_ML_Assignment2 > 2 >

Name	Date modified	Type	Size
hw2_test	3/16/2016 2:19 AM	File folder	
hw2_train	3/16/2016 2:19 AM	File folder	

hw2_train contains training files.

This PC > OS (C:) > NXP154130_ML_Assignment2 > 2 > hw2_train >

Name	Date modified	Type	Size
train	3/8/2016 10:41 PM	File folder	

This PC > OS (C:) > NXP154130_ML_Assignment2 > 2 > hw2_train > train >

Name	Date modified	Type	Size
ham	3/16/2016 2:19 AM	File folder	
spam	3/16/2016 2:19 AM	File folder	

hw2_test contains test files.

This PC > OS (C:) > NXP154130_ML_Assignment2 > 2 > hw2_test >

Name	Date modified	Type	Size
test	3/8/2016 10:41 PM	File folder	

This PC > OS (C:) > NXP154130_ML_Assignment2 > 2 > hw2_test > test >

Name	Date modified	Type	Size
ham	3/16/2016 2:19 AM	File folder	
spam	3/16/2016 2:19 AM	File folder	

Folder 3 is manually created documents to check for spam and ham.

This PC > OS (C:) > NXP154130_ML_Assignment2 > 3 >

Name	Date modified	Type	Size
hw2_test	3/9/2016 9:08 PM	File folder	
hw2_train	3/9/2016 9:08 PM	File folder	

Main_class.py -> python program that need to run.

Running the program steps: ->

1) Go to 'command prompt'

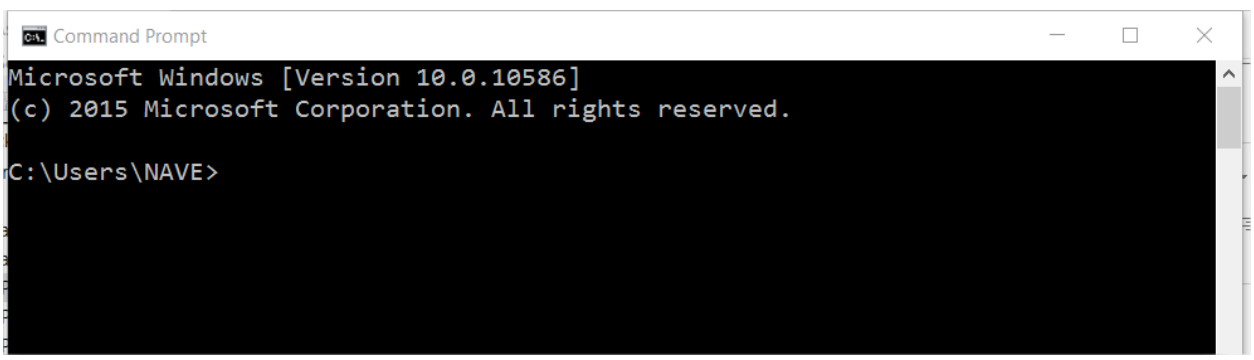
2) Go to folder where python is installed.

3) Given command as (python.exe 'full path to-> Main_class.py' 'Training_ham_folder_location' 'Training_spam_folder_location' 'Test_ham_folder_location' 'Test_spam_folder_location' Learning_Rate Lambda 'Stop_words_file_location')

4) Stop_words_file_location is the location of the file 'STOP_WORDS.txt' it is main folder itself.

Sample Run:

1)Start->cmd



2) Copy the path where python 3.5 is installed.

For me path is:

C:\Users\NAVE\AppData\Local\Programs\Python\Python35-32\python.exe

3) Copy the path where folder is downloaded.

For me it is:

C:\NXP154130_ML_Assignment2\Main_class.py

4) Copy the path of training ham folder.

For me it is:

C:\NXP154130_ML_Assignment2\2\hw2_train\train\ham

5) Copy the path of training spam folder.

For me it is:
C:\NXP154130_ML_Assignment2\2\hw2_train\train\spam

6) Copy the path of test ham folder.

For me it is:
C:\NXP154130_ML_Assignment2\2\hw2_test\test\ham

7) Copy the path of test spam folder.

For me it is:
C:\NXP154130_ML_Assignment2\2\hw2_test\test\spam

8) Copy the path of stop words.

For me it is:
C:\NXP154130_ML_Assignment2\STOP_WORDS.txt

9)Decide values for Learning Rate and Lambda.

I have Lambda dif values from (-1 to -18) and Learning Rate as 0.1

Now run the program:

```
C:\Users\NAVE\AppData\Local\Programs\Python\Python35-32\python.exe
C:\NXP154130_ML_Assignment2\Main_class.py
C:\NXP154130_ML_Assignment2\2\hw2_train\train\ham
C:\NXP154130_ML_Assignment2\2\hw2_train\train\spam
C:\NXP154130_ML_Assignment2\2\hw2_test\test\ham
C:\NXP154130_ML_Assignment2\2\hw2_test\test\spam                    0.1                    -10
C:\NXP154130_ML_Assignment2\STOP_WORDS.txt
```

Note: Program might take more than 4 minutes for a run. Please be patience.

```
C:\Users\NAVE>C:\Users\NAVE\AppData\Local\Programs\Python\Python35-32\python.exe
 C:\NXP154130_ML_Assignment2\Main_class.py C:\NXP154130_ML_Assignment2\2\hw2_tra
in\train\ham C:\NXP154130_ML_Assignment2\2\hw2_train\train\spam C:\NXP154130_ML_
Assignment2\2\hw2_test\test\ham C:\NXP154130_ML_Assignment2\2\hw2_test\test\spam
 0.1 -10 C:\NXP154130_ML_Assignment2\STOP_WORDS.txt
running....... Running Naive Bayes with out stop words
running....... Running Logistic Regression with out stop words
running....... Running Naive Bayes with stop words
running....... Running Logistic Regression with stop words
naive bayes accuracy 74.68619246861925
Logistic Regression accuracy 70.50209205020921
naive bayes with stop words accuracy 78.8702928870293
Logistic Regression with stop words accuracy 82.21757322175732

C:\Users\NAVE>
```

Accuracy details for different Lambda value:

Lambda : -10

Naïve Bayes without stop words: 74.68

Naïve Bayes with stop words: 78.87

Logistic Regression without stop words: 70.50

Logistic Regression with stop words: 82.21

Lambda: -15

Naïve Bayes without stop words: 74.68

Naïve Bayes with stop words: 78.87

Logistic Regression without stop words: 70.711

Logistic Regression with stop words: 82.42

Lambda: -5

Naïve Bayes without stop words: 74.68

Naïve Bayes with stop words: 78.87

Logistic Regression without stop words: 71.33

Logistic Regression with stop words: 80.96

Lambda: -1

Naïve Bayes without stop words: 74.68

Naïve Bayes with stop words: 78.87

Logistic Regression without stop words: 72.17

Logistic Regression with stop words: 78.45

Lambda: 0

Naïve Bayes without stop words: 74.68

Naïve Bayes with stop words: 78.87

Logistic Regression without stop words: 73.87

Logistic Regression with stop words: 71.19

Lambda: -18

Naïve Bayes without stop words: 74.68

Naïve Bayes with stop words: 78.87

Logistic Regression without stop words: 68.619

Logistic Regression with stop words: 81.58

Lambda: 1

Naïve Bayes without stop words: 74.68

Naïve Bayes with stop words: 78.87

Logistic Regression without stop words: 72.17

Logistic Regression with stop words: 75.94

Word Details:

Total number of words different words: 10316

Total number of different words without stop words: 9811

Manual Test:

To manually test the accuracy. Spam and Ham documents are created and tested with accuracy.

Just replace folder name 2 with 3 to test this file.

C:\Users\NAVE\AppData\Local\Programs\Python\Python35-32\python.exe
C:\NXP154130_ML_Assignment2\Main_class.py
C:\NXP154130_ML_Assignment2\3\hw2_train\train\ham
C:\NXP154130_ML_Assignment2\3\hw2_train\train\spam
C:\NXP154130_ML_Assignment2\3\hw2_test\test\ham
C:\NXP154130_ML_Assignment2\3\hw2_test\test\spam 0.1 -10
C:\NXP154130_ML_Assignment2\STOP_WORDS.txt

```
C:\Users\NAVE>C:\Users\NAVE\AppData\Local\Programs\Python\Python35-32\python.exe
 C:\NXP154130_ML_Assignment2\Main_class.py C:\NXP154130_ML_Assignment2\3\hw2_tra
in\train\ham C:\NXP154130_ML_Assignment2\3\hw2_train\train\spam C:\NXP154130_ML_
Assignment2\3\hw2_test\test\ham C:\NXP154130_ML_Assignment2\3\hw2_test\test\spam
 0.1 -10 C:\NXP154130_ML_Assignment2\STOP_WORDS.txt
running....... Running Naive Bayes with out stop words
running....... Running Logistic Regression with out stop words
running....... Running Naive Bayes with stop words
running....... Running Logistic Regression with stop words
naive bayes accuracy 100.0
Logistic Regression accuracy 100.0
naive bayes with stop words accuracy 100.0
Logistic Regression with stop words accuracy 100.0
```

For small perfect data we are getting 100% accuracy. This proves correctness of algorithm.

Notes:

- *Laplace smoothing used: 1*
- *All the multiplication is done based of log values.*
- *Algorithm is always run for 50 iterations. I printed weight values and checked that it is converged on 30th iteration itself.*
- *Without* ***stop words accuracy is increased for naïve Bayes****. Reason is that stop words will appear in spam and ham folders. Since ham folders have more files and spam here there is more chance it might classify all the document to ham because of the stop words. Thus removing some stop words increased accuracy for us. Naïve Bayes works based on count of occurrence of words in spam and ham so removing stop words removing accuracy.*