Assignment 3

How to execute:

To execute the code, you simply run:

python [filename]

To any of the file. This should automatically start the program. Parameter specification is not needed because this is done inside the program. All that it is needed is to start the program and the results will be outputted to the terminal.

The following files were attached to the zip folder:

hw3_logisticregression_with_stopwords.py

hw3_logisticregression_without_stopwords.py

hw3_naivebayes_with_stopwords.py

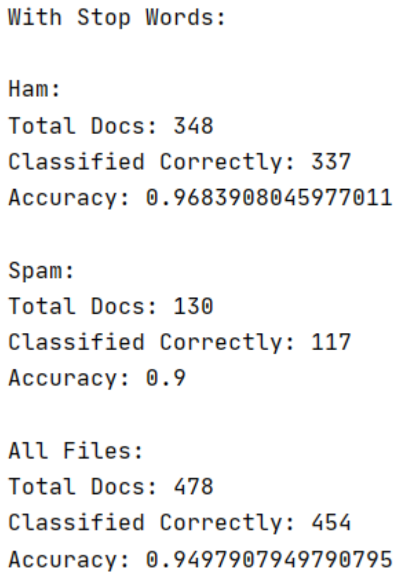
hw3_naivebayes_without_stopwords.py

Removing Stop Words

Removing stop words typically decreased my model's performance / accuracy. Please see below the accuracies with and without stop words. At first, I thought my accuracies were going to increase with the removal of stop words because typically stop words provide no meaning to the document. However, with the results given, it made me think that basically in emails stop words are important, or at least some of them. I think why my performance suffer is because certain stop words appear more in one classification than in others, at least for this training set. For example, in spam the stop word "you" might appear more often because the spam emails are trying to convince you to get their product, click on their link, etc. and with the removal of it, it decreases the performance. Another possibility can happen in ham where communication is often done with another person and "I" can be referred to a lot, while in spam not so much, because communication between people is not occurring. Thus, the removal of stop words like "you" and "I" can severely affect the probability of classifying one class compared to another. Finally, another possibility can be that this training set is not large enough to train without stop words. Due to its size, removing more words can result in decreased performance because there are not enough words to train on. Due to these reasons I believe my accuracy decreased with the removal of stop words.

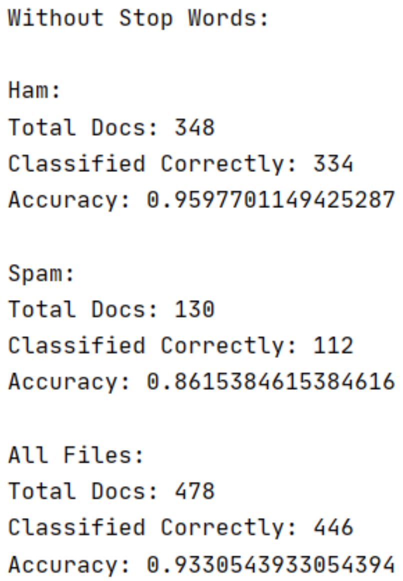
Accuracies

Naïve Bayes With Stop Words



```
With Stop Words:

Ham:
Total Docs: 348
Classified Correctly: 337
Accuracy: 0.9683908045977011

Spam:
Total Docs: 130
Classified Correctly: 117
Accuracy: 0.9

All Files:
Total Docs: 478
Classified Correctly: 454
Accuracy: 0.9497907949790795
```

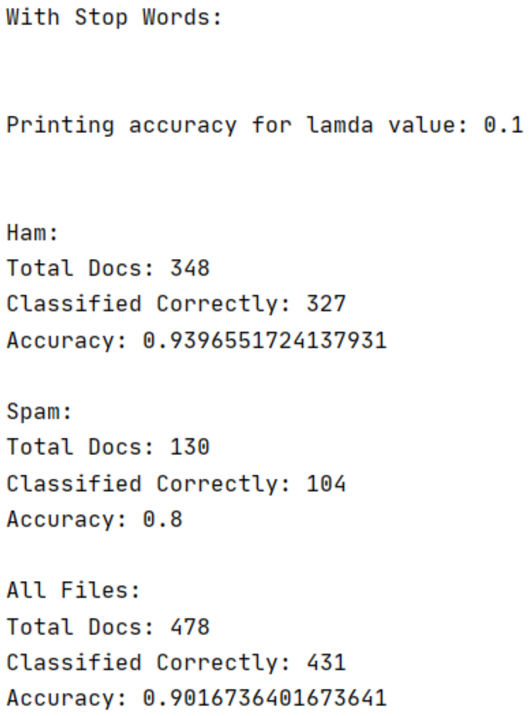
Naïve Bayes Without Stop Words



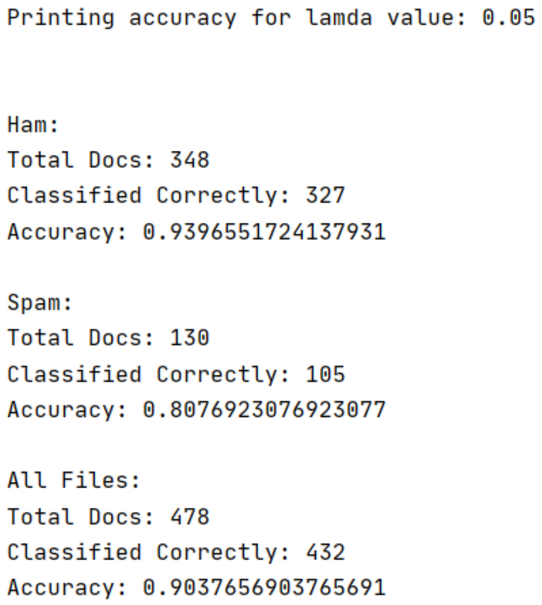
```
Without Stop Words:

Ham:
Total Docs: 348
Classified Correctly: 334
Accuracy: 0.9597701149425287

Spam:
Total Docs: 130
Classified Correctly: 112
Accuracy: 0.8615384615384616

All Files:
Total Docs: 478
Classified Correctly: 446
Accuracy: 0.9330543933054394
```

Logistic Regression With Stop Words

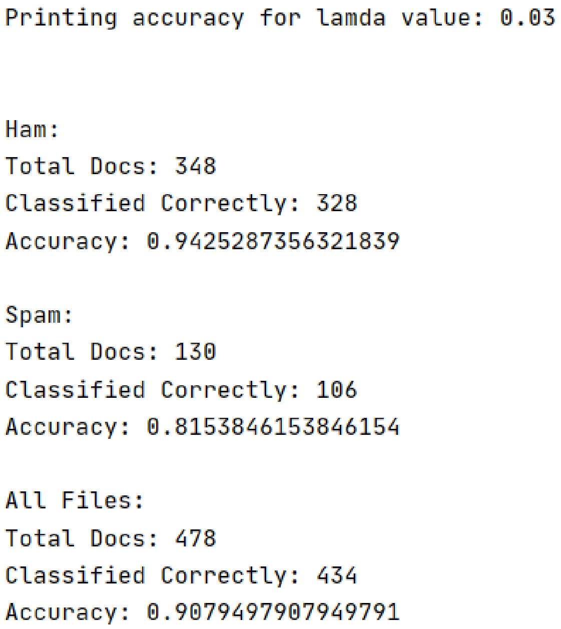
Note: Mu value is 0.05, iterations are 100, and the lambda values are 0.1, 0.05, 0.03, 0.01, 0.001. In total there will be 5 train and test with different lambda values. Mu and number of iterations stay the same.



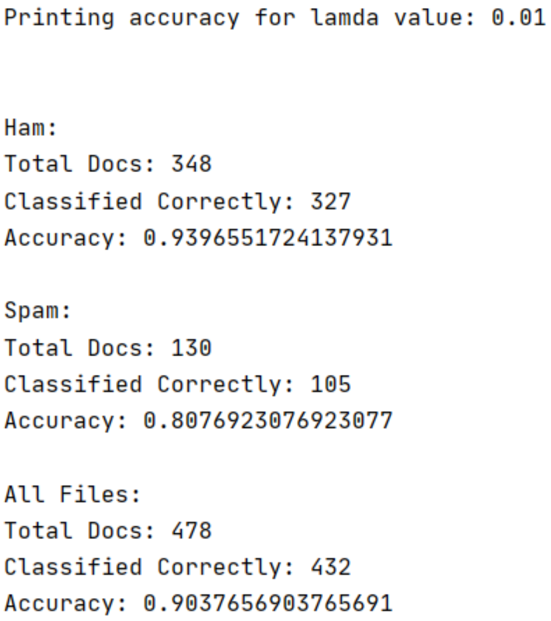
```
With Stop Words:


Printing accuracy for lamda value: 0.1


Ham:
Total Docs: 348
Classified Correctly: 327
Accuracy: 0.9396551724137931

Spam:
Total Docs: 130
Classified Correctly: 104
Accuracy: 0.8

All Files:
Total Docs: 478
Classified Correctly: 431
Accuracy: 0.9016736401673641
```



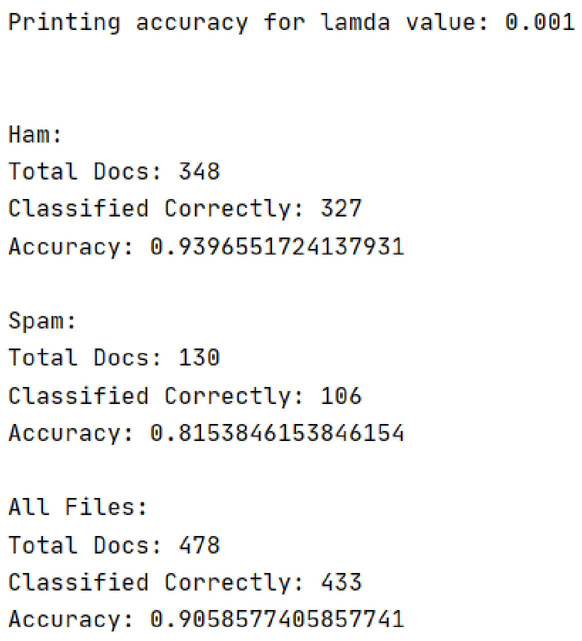
```
Printing accuracy for lamda value: 0.05


Ham:
Total Docs: 348
Classified Correctly: 327
Accuracy: 0.9396551724137931

Spam:
Total Docs: 130
Classified Correctly: 105
Accuracy: 0.8076923076923077

All Files:
Total Docs: 478
Classified Correctly: 432
Accuracy: 0.9037656903765691
```



```
Printing accuracy for lamda value: 0.03


Ham:
Total Docs: 348
Classified Correctly: 328
Accuracy: 0.9425287356321839

Spam:
Total Docs: 130
Classified Correctly: 106
Accuracy: 0.8153846153846154

All Files:
Total Docs: 478
Classified Correctly: 434
Accuracy: 0.9079497907949791
```



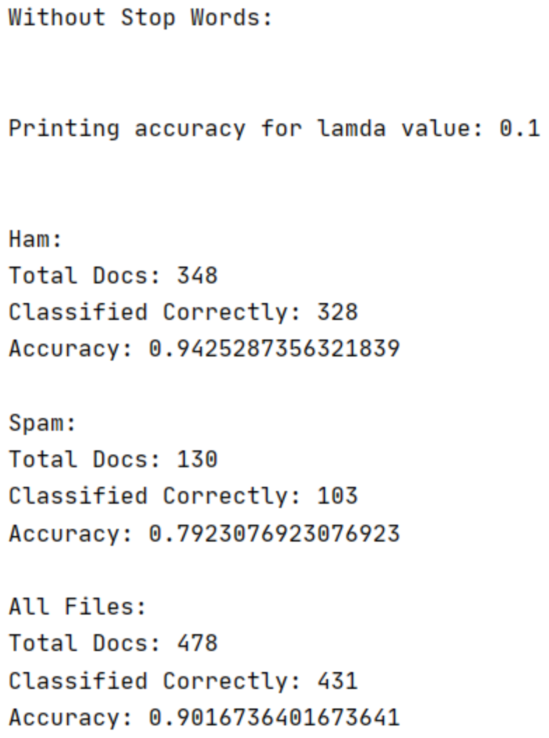
```
Printing accuracy for lamda value: 0.01


Ham:
Total Docs: 348
Classified Correctly: 327
Accuracy: 0.9396551724137931

Spam:
Total Docs: 130
Classified Correctly: 105
Accuracy: 0.8076923076923077

All Files:
Total Docs: 478
Classified Correctly: 432
Accuracy: 0.9037656903765691
```



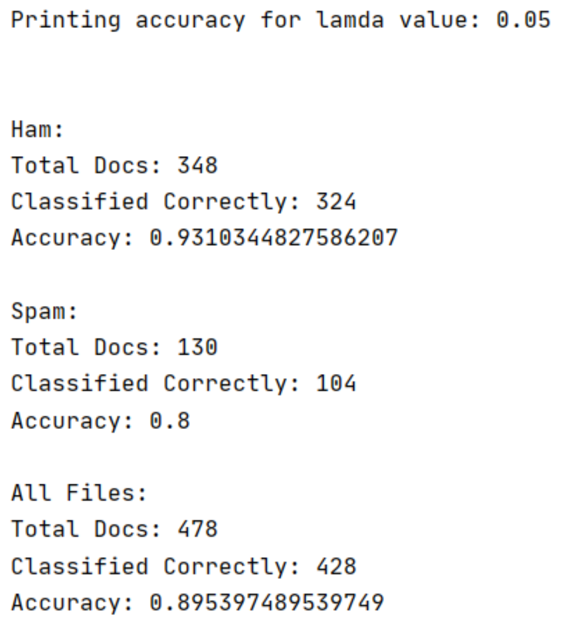
```
Printing accuracy for lamda value: 0.001


Ham:
Total Docs: 348
Classified Correctly: 327
Accuracy: 0.9396551724137931

Spam:
Total Docs: 130
Classified Correctly: 106
Accuracy: 0.8153846153846154

All Files:
Total Docs: 478
Classified Correctly: 433
Accuracy: 0.9058577405857741
```

Logistic Regression Without Stop Words

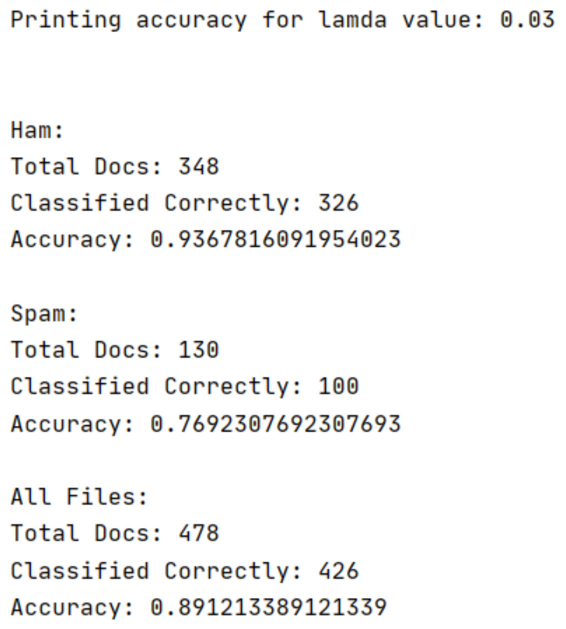
Note: Mu value is 0.05, iterations are 100, and the lambda values are 0.1, 0.05, 0.03, 0.01, 0.001. In total there will be 5 train and test with different lambda values. Mu and number of iterations stay the same.



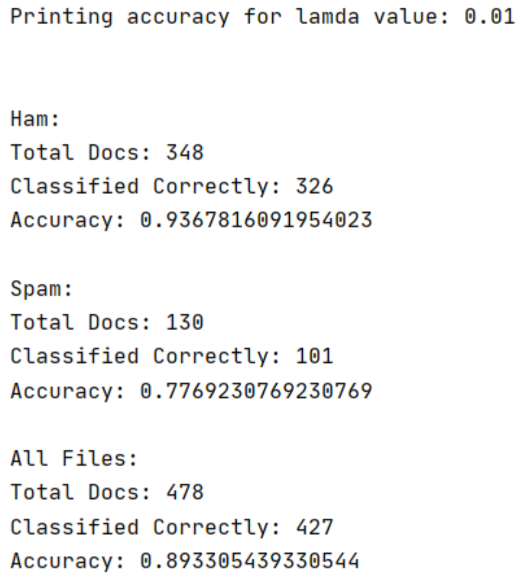
```
Without Stop Words:


Printing accuracy for lamda value: 0.1


Ham:
Total Docs: 348
Classified Correctly: 328
Accuracy: 0.9425287356321839

Spam:
Total Docs: 130
Classified Correctly: 103
Accuracy: 0.7923076923076923

All Files:
Total Docs: 478
Classified Correctly: 431
Accuracy: 0.9016736401673641
```



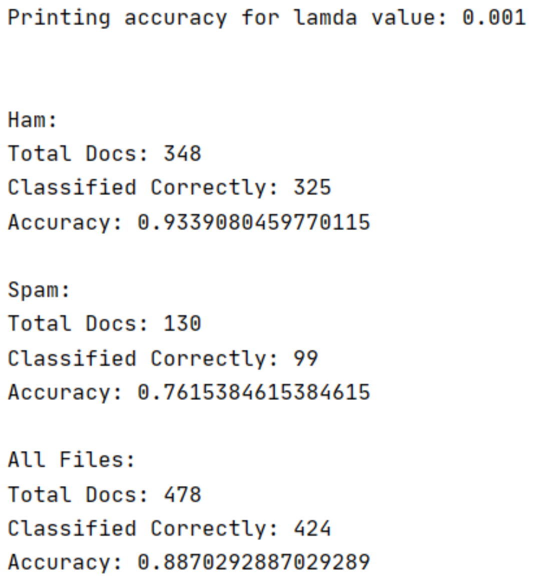
```
Printing accuracy for lamda value: 0.05


Ham:
Total Docs: 348
Classified Correctly: 324
Accuracy: 0.9310344827586207

Spam:
Total Docs: 130
Classified Correctly: 104
Accuracy: 0.8

All Files:
Total Docs: 478
Classified Correctly: 428
Accuracy: 0.895397489539749
```



```
Printing accuracy for lamda value: 0.03


Ham:
Total Docs: 348
Classified Correctly: 326
Accuracy: 0.9367816091954023

Spam:
Total Docs: 130
Classified Correctly: 100
Accuracy: 0.7692307692307693

All Files:
Total Docs: 478
Classified Correctly: 426
Accuracy: 0.891213389121339
```



```
Printing accuracy for lamda value: 0.01


Ham:
Total Docs: 348
Classified Correctly: 326
Accuracy: 0.9367816091954023

Spam:
Total Docs: 130
Classified Correctly: 101
Accuracy: 0.7769230769230769

All Files:
Total Docs: 478
Classified Correctly: 427
Accuracy: 0.893305439330544
```



```
Printing accuracy for lamda value: 0.001


Ham:
Total Docs: 348
Classified Correctly: 325
Accuracy: 0.9339080459770115

Spam:
Total Docs: 130
Classified Correctly: 99
Accuracy: 0.7615384615384615

All Files:
Total Docs: 478
Classified Correctly: 424
Accuracy: 0.8870292887029289
```

Lambda Values

As you can see different lambda values give different results. The best lambda value for logistic regression with stop words is 0.03 which gives us an overall 90.7% accuracy rate. The best lambda value for logistic regression without stop words is 0.1 which gives us an overall 90.1% accuracy rate. As you can see, the accuracies went down with the removal of stop words. Please see “Removing Stop Words” section to see why the accuracies went down.

If you have any questions regarding this document or my programs. Please let me know thank you.