Classification Naive Bayes and Logistic Regression

# Naive Bayes for Text Classification:

## Accuracy:

### Accuracy of Naive Bayes without removing Stop Words:

Spam % Accuracy: 97.6923076923077

Ham % Accuracy: 93.39080459770115

### Accuracy of Naive Bayes after removal of Stop Words:

Spam % Accuracy: 96.92307692307692

Ham % Accuracy: 92.24137931034483

# Logistic Regression for Text Classification:

Accuracy observed for various values of leaning rate penalty regularization term – Lambda.

And number of iterations (hard-limit) to come to convergence. Please note that it takes a significant amount of time to complete the hard-limit number of iterations.

#### For, Learning rate: 0.01

Lambda=0.01 Limit=50

without removing Stop Words:

Testing Spam test files.....

Accuracy on spam 65.38461538461539

Testing ham test files.....

Accuracy on Ham 80.17241379310344

after removal of Stop Words:

Testing Spam test files.....

Accuracy on spam 69.23076923076923

Testing ham test files.....

Accuracy on Ham 89.9425287356322

#### Lambda=0.01 Limit=100

without removing Stop Words:

Testing Spam test files.....

Accuracy on spam 56.15384615384615

Testing ham test files.....

Accuracy on Ham 73.85057471264368

after removal of Stop Words:

Testing Spam test files.....

Accuracy on spam 64.61538461538461

Testing ham test files.....

Accuracy on Ham 84.77011494252874

#### Lambda=0.01 Limit=500

without removing Stop Words:

Testing Spam test files.....

Accuracy on spam 53.07692307692308

Testing ham test files.....

Accuracy on Ham 81.03448275862068

after removal of Stop Words:

Testing Spam test files.....

Accuracy on spam 60.76923076923077

Testing ham test files.....

Accuracy on Ham 83.62068965517241

#### Lambda=0.1 Limit=500

without removing Stop Words:

Testing Spam test files.....

Accuracy on spam 57.692307692307686

Testing ham test files.....

Accuracy on Ham 74.42528735632183

after removal of Stop Words:

Testing Spam test files.....

Accuracy on spam 65.38461538461539

Testing ham test files.....

Accuracy on Ham 88.2183908045977

##### Lambda=0.5 Limit=500

without removing Stop Words:

Testing Spam test files.....

Accuracy on spam 60.76923076923077

Testing ham test files.....

Accuracy on Ham 79.02298850574712

after removal of Stop Words:

Testing Spam test files.....

Accuracy on spam 65.38461538461539

Testing ham test files.....

Accuracy on Ham 84.19540229885058

##### For, Learning rate=2

##### Lambda=2 Limit=500

without removing Stop Words:

Testing Spam test files.....

Accuracy on spam 76.92307692307693

Testing ham test files.....

Accuracy on Ham 84.19540229885058

after removal of Stop Words:

Testing Spam test files.....

Accuracy on spam 80.76923076923077

Testing ham test files.....

Accuracy on Ham 91.0919540229885

##### Learning rate=3

##### Lambda=3 Limit=500

without removing Stop Words:

Testing Spam test files.....

Accuracy on spam 63.84615384615384

Testing ham test files.....

Accuracy on Ham 85.0574712643678

after removal of Stop Words:

Testing Spam test files.....

Accuracy on spam 65.07692307692307

Testing ham test files.....

Accuracy on Ham 91.37931034482759

Does the accuracy improve?

Yes there is a 3 to 5% improvement in the accuracy after removal of stop word

Explain why the accuracy improves or why it does not?

When we remove stop words from the documents our vocabulary contains less words than the dictionary constructed without considering stop words. There is slight reduction in both cases – Ham and Spam accuracy. This might be due to distribution of Stop words in either case.

Also, there is no strict relation between the stop words and the documents, since if a document contains more stop words, and then we just get rid of the stops words before processing. This restricts the size of the word set in consideration.

Logistic regression:

The values of the parameters – Learning rate – dictates how fast we move towards the convergence point.

Also, the parameter regularization term – Lambda controls the penalty that we add to the weight calculation. Since it takes more computation power to find the exact convergence point, we restrict on the number of iterations our model takes to get closer to convergence point.

From the result set, I observe that when I increase the iterations in general, there is an increase in accuracy. However, if the learning rate is increased, the value of accuracy varies, it either increases for some cases or decreases and this is due to the fact that if we move at a faster rate there we converge to a wrong point. Also, the variations in Lambda in combination with the value of learning rate and number of iteration.