Assignment 2

CS 6375: MACHINE LEARNING

Spring 2016

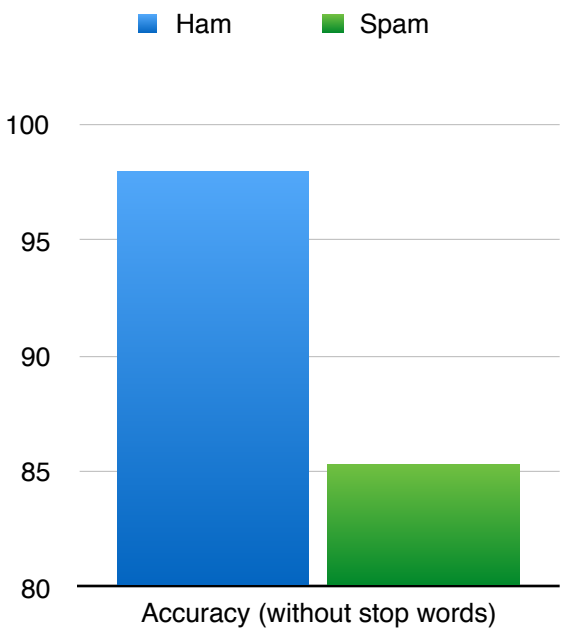
Submitted by :
Himanshu Kandwal (hxk154230)

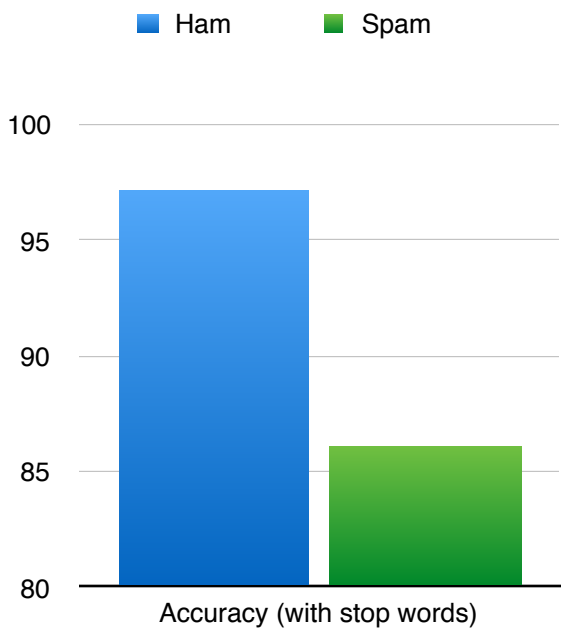
Naïve Bayes Classification

Setup:

Data Set	SPAM	HAM
Training Data	123	340
Test Data	130	348

Accuracy	Stopwords	SPAM	HAM	Total
Test Data	no	85.384	97.988	94.56
Test Data	yes	86.153	97.126	94.14





As we can see there is no big change after removal of stop words. This is because Naive Bayes algorithm is based on count of tokens and conditional probability corresponding to that. There are not many stop words in the training set so there is not much impact of stop words in our number of features.

Logistic Regression Classification

Setup:

Data Set	SPAM	HAM
Training Data	123	340
Test Data	130	348

Repetitions	Learning rate	Lambda	Stopwords	Spam Accuracy	Ham Accuracy	Total Accuracy
100	0.01	0	no	86.92	94.82	92.67
100	0.01	0	yes	86.15	96.55	93.72
100	0.01	0.01	no	86.15	95.11	92.67
100	0.01	0.01	yes	86.92	96.26	93.72
100	0.01	1	no	86.15	94.82	92.46
100	0.01	1	yes	86.92	96.83	94.14
100	0.01	5	no	85.38	95.11	92.46
100	0.01	5	yes	80	97.7	92.88
100	0.5	1	no	86.15	94.82	92.46
100	0.5	1	yes	86.92	96.83	94.14

There is an increase in the Ham and Spam Accuracy after removal of Stop words. The reason is reduction in features has removed noise from data. Stop words frequency count was big in numbers. The best practice in Text classification is to always remove stop words from training Data. Removal of Stop words always help in improving accuracy.

We can see that sometime there is sudden decrease in the spam Accuracy after increasing value of lambda. This is the case when after penalizing weights the decision boundary has included spam data on ham side. This can be the case of under fitting as the decision boundary is not proper.