**OUTPUT FOR NAÏVE BAYES**:

NAIVE BAYES WITH STOP WORDS INCLUDED:

THE ACCURACY OF NAIVE BAYES OVER SPAM RECORDS IS: 98.46153846153847

THE ACCURACY OF NAIVE BAYES OVER HAM RECORDS IS: 96.55172413793103

NAIVE BAYES WITH STOP WORDS REMOVED:

THE ACCURACY OF NAIVE BAYES OVER SPAM RECORDS IS: 98.46153846153847

THE ACCURACY OF NAIVE BAYES OVER HAM RECORDS IS: 95.40229885057471

The accuracy for Naïve Bayes does not improve after removing the Stop Words. This could be because Naïve Bayes is a comparative probability classifier. So, removing stop words removes them from both the probabilities (for spam and ham). Hence, the comparative result of the two probabilities is the same. Thus, there should not be any obvious increase in the accuracy by removing Stop Words. I have theorized that there is a small decrease in my accuracy for HAM records because Ham depends more on Stop Words as compared to Spam(which is garbage and rarely uses grammatically correct English and spelling).

**OUTPUT FOR LOGISTIC REGRESSION:**

STOP LIST ENABLED:**false** lemda = 0.1

THE ACCURACY OF LOGISTIC REGRESSION OVER SPAM RECORDS IS: 86.15384615384616

THE ACCURACY OF LOGISTIC REGRESSIOn OVER HAM RECORDS IS: 93.96551724137932

STOP LIST ENABLED:**true** lemda = 0.1

THE ACCURACY OF LOGISTIC REGRESSION OVER SPAM RECORDS IS: 90.0

THE ACCURACY OF LOGISTIC REGRESSIOn OVER HAM RECORDS IS: 96.26436781609195

There is an increase in accuracy if Stop words are removed from Logistic Regression model. This could be attributed to the fact that the weights for stop words(more common words) would be large and would require regularization. When these words are removed the need for regularization lowers and better accuracy can be achieved for smaller values of lamda.

**LOGISITC REGRESSION**

The hard limit on iterations is 10. I used 20 for lower values of lamda as the convergence for lower values takes more iterations than usual. While submitting, I have set the number of iterations performed as 10.

Learning rate(eenta) for the algorithm is 0.01(the default learning rate proved enough to converge to the values in 10 iterations)

Values of lamda from 0.0 to 4.0 in steps of 0.4 report the following values:

THE OUTPUT IN REFERENCE IN MENTIONED IN THE FINAL OUTPUT TEXT FILE.

As you can see, after lemda = 3.7, the accuracy starts fluctuating. Hence, lemda=3.0 is the largest value of regularization we can afford. However, when we remove the stop words from the list of vocabulary, the accuracy fluctuates after 2.5.