Assignment 2:

Naïve Bayes and logistic regression

Accuracy of Naïve Bayes algorithm including stop words: 97.6923 %

Accuracy of naïve Bayes algorithm excluding stop words: 99.2307 %

Accuracy of Logistic regression with rate of learning = 1 and lambda = 0.3 algorithm including stop words: 73.01255 %

Accuracy of Logistic regression algorithm with rate of learning = 1 and lambda = 0.3 excluding stop words: 75.941%

Why does accuracy improve after eliminating stop words?

As we can see from the results, accuracy of the program increased after removing stop words.

Stop words include words such as “a” , “an”, “about” , “after” and symbols like “!” , “,” etc.

These words do not predict the class of a file. Treating these words as feature words would result in poor performance in text classification. These are common words which can occur in both the classes. So we cannot classify the mail using these words. If these words and symbols are not considered in bag of words we can estimate more precisely.