**ASSIGNMENT 3**

**Assumptions**

1. Header is assumed to be everything above the first blank line.
2. Stop words present in the NLTK corpus are considered.

**Pre-processing Steps**

**Removal of Header**

(All the lines before the first blank line are removed)

**Removal of Punctuation marks, comma, etc**

(They are removed through regular expression)

**Tokenization**

(Tokens are formed using word\_tokenize and special symbols are removed)

**Removal of Stop Words**

(Stop words are removed using NLTK stop words)

**Normalization**

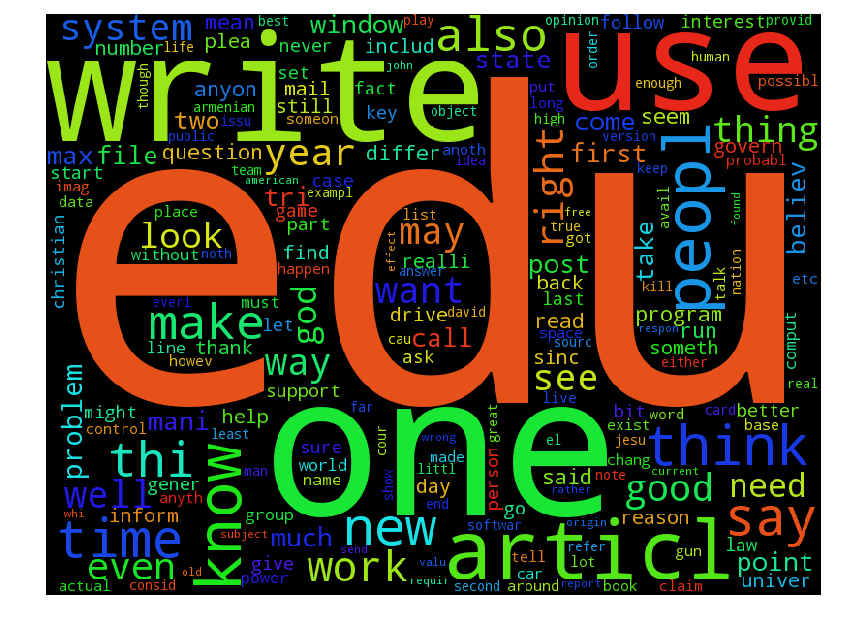
(All token are converted into lower case)

**Stemming**

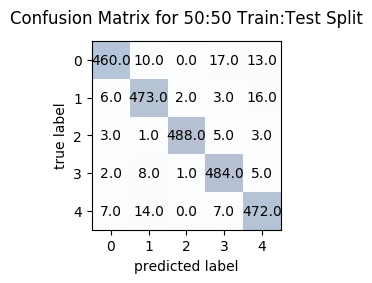
(Stemming is performed using Porter algorithm to get the root word)

**Number of Documents:** 5000

**WordCloud**

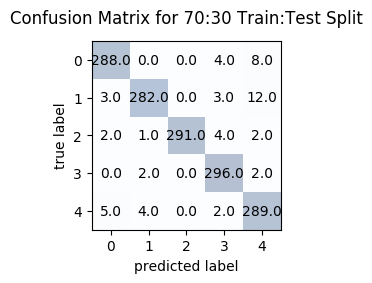


**50:50 Train Test Split**



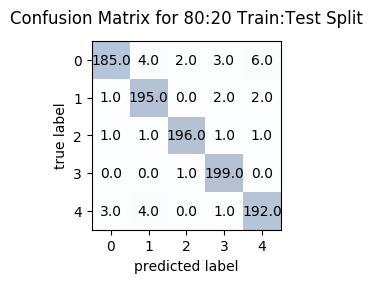
**Accuracy= 95.08%**

**70:30 Train Test Split**



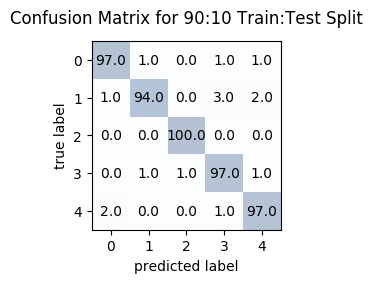
**Accuracy=96.4%**

**80:20 Train Test Split**

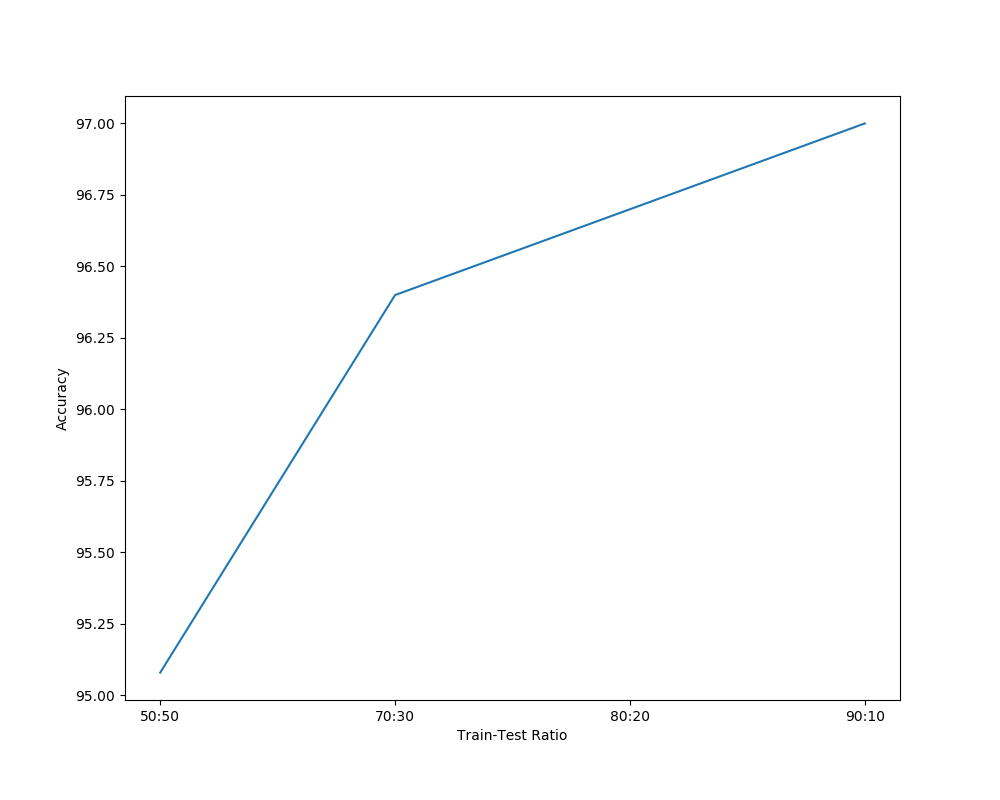


**Accuracy=** **96.7%**

**90:10 Train Test Split**



**Accuracy=** **97.0%**



**Accuracy vs Train-Test Ratio**

From the graph, we can infer that the accuracy of the Naïve Bayes model increases with the increase in the Train Ratio (Train Data) (decrease in Test Ratio).

**Feature Selection using TF-IDF Score**

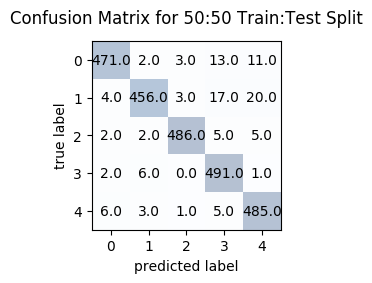
First the tf-idf of each term and document pair is computed and then top k terms from each documents are selected as features.

Training is performed using the new vocabulary (features).

**K=10**

**50:50 Train Test Split**

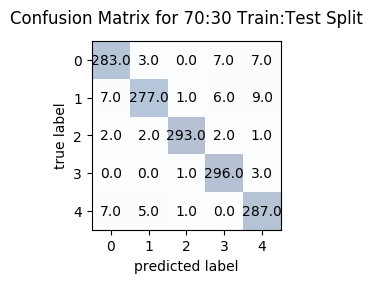
Vocabulary Size: 20972



**Accuracy=** **95.56%**

**70:30 Train Test Split**

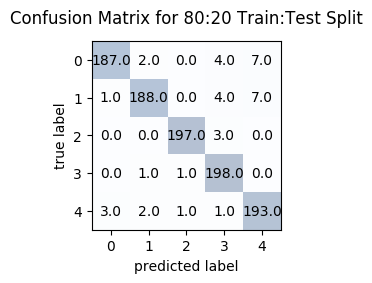
Vocabulary Size: 26173



**Accuracy=** **95.73%**

**80:20 Train Test Split**

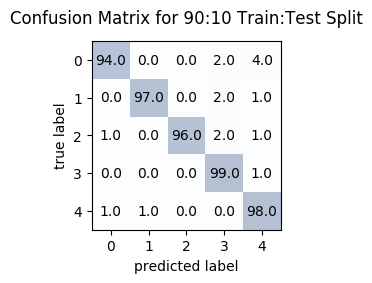
Vocabulary Size: 28128



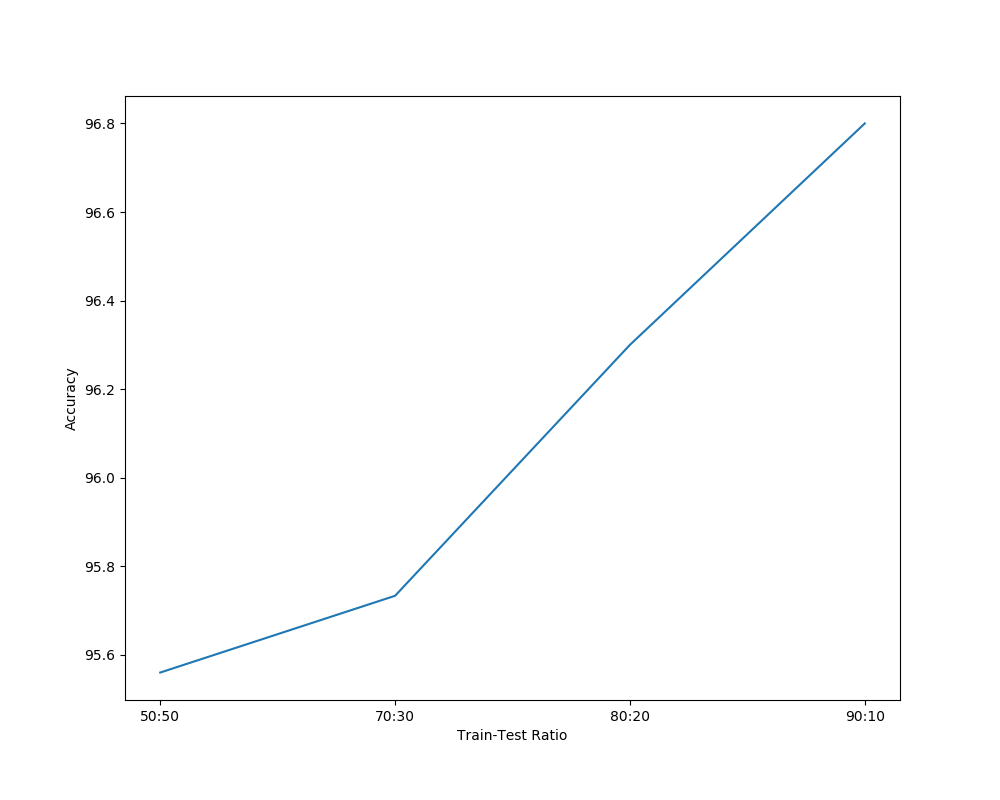
**Accuracy=** **96.3%**

**90:10 Train Test Split**

Vocabulary Size: 29930



**Accuracy=** **96.8%**



**Accuracy vs Train-Test Ratio**

Even though the number of terms (features) are reduced by almost 10,000, the accuracy on test documents is still greater than **95%.** The accuracy increases with the increase in train data size.