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

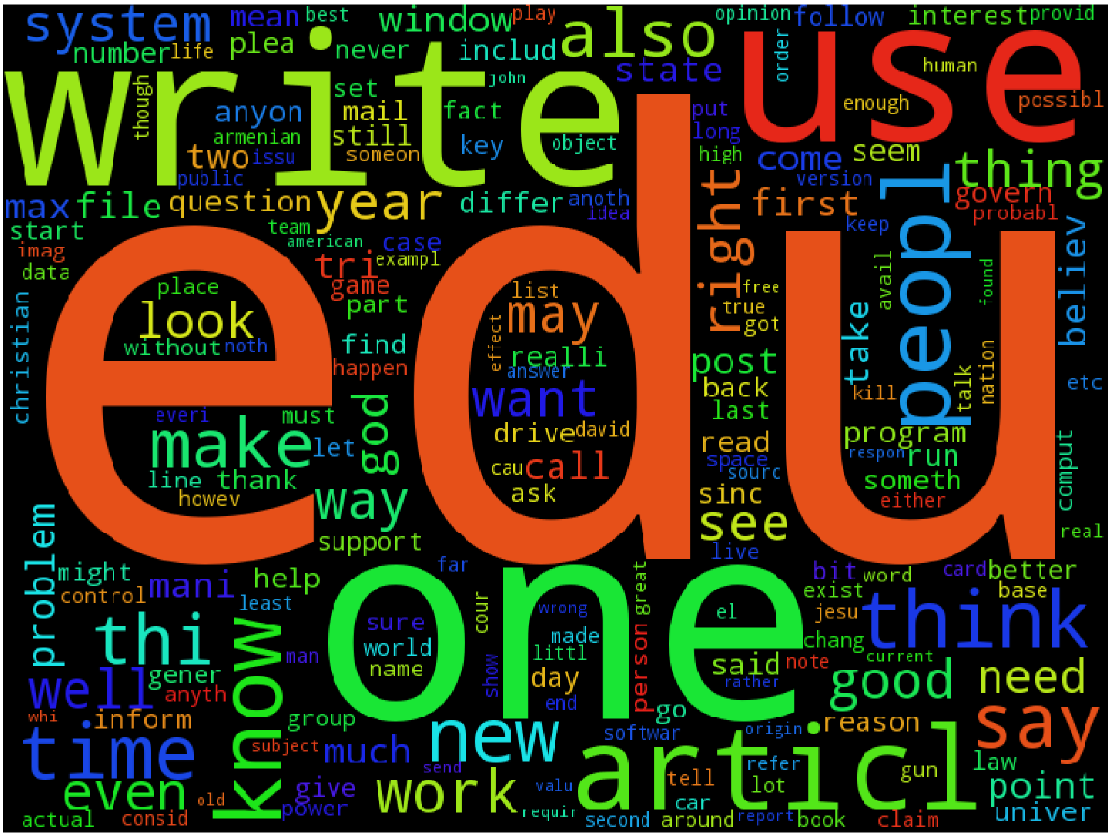
(All token are converted into lower case)

Stemming

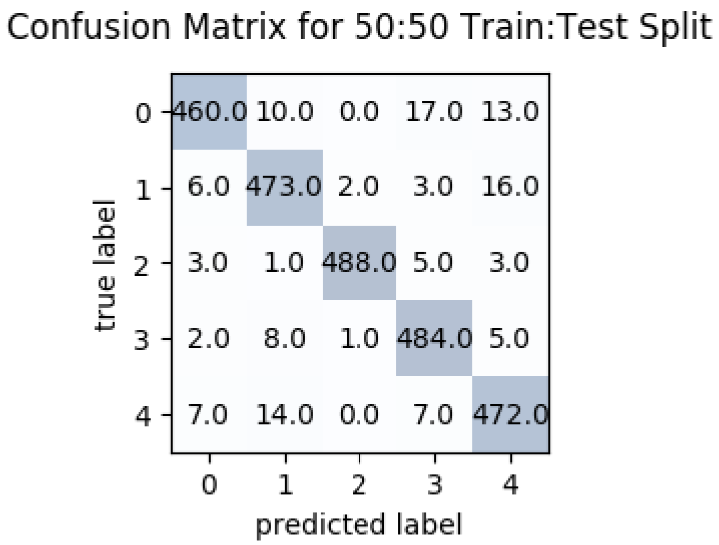
(Stemming is performed using Porter algorithm to get the root word)

Number of Documents: 5000

WordCloud

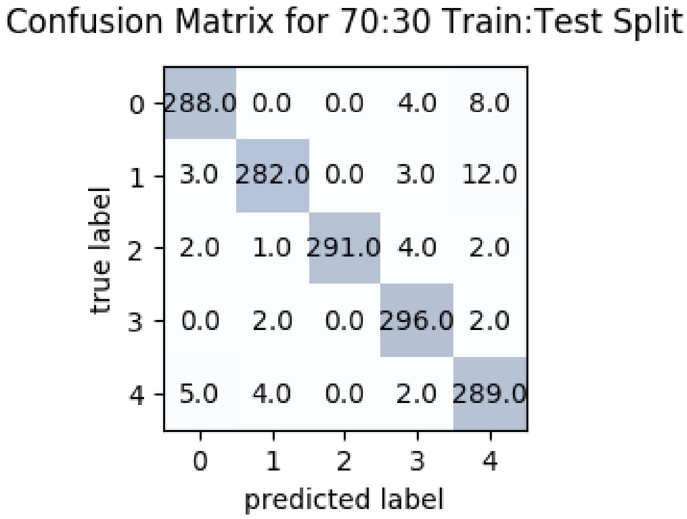


50:50 Train Test Split


Confusion Matrix for 50:50 Train:Test Split

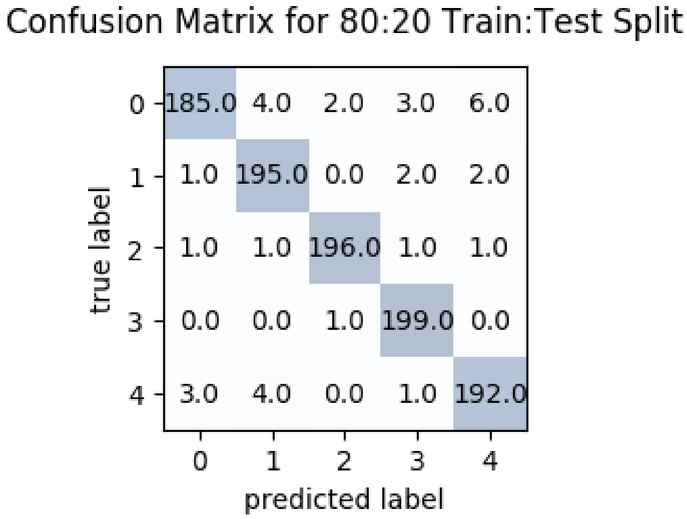
Accuracy= 95.08%

70:30 Train Test Split



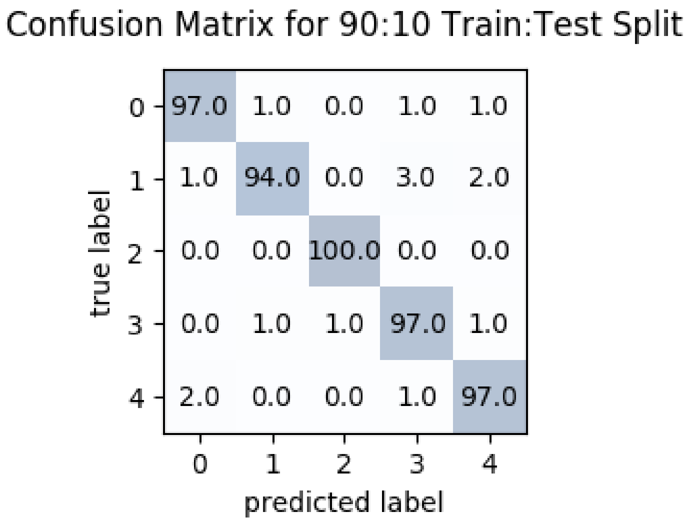
Accuracy=96.4%

80:20 Train Test Split



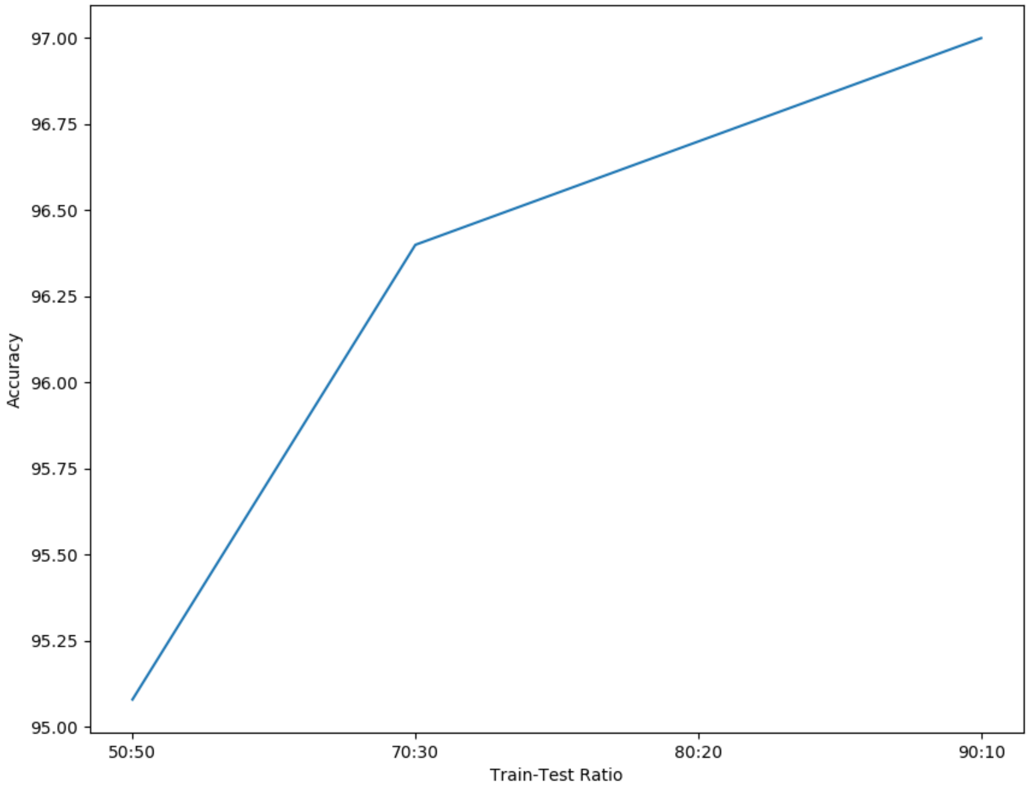
Accuracy= 96.7%

90:10 Train Test Split



	0	1	2	3	4
0	97.0	1.0	0.0	1.0	1.0
1	1.0	94.0	0.0	3.0	2.0
2	0.0	0.0	100.0	0.0	0.0
3	0.0	1.0	1.0	97.0	1.0
4	2.0	0.0	0.0	1.0	97.0

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

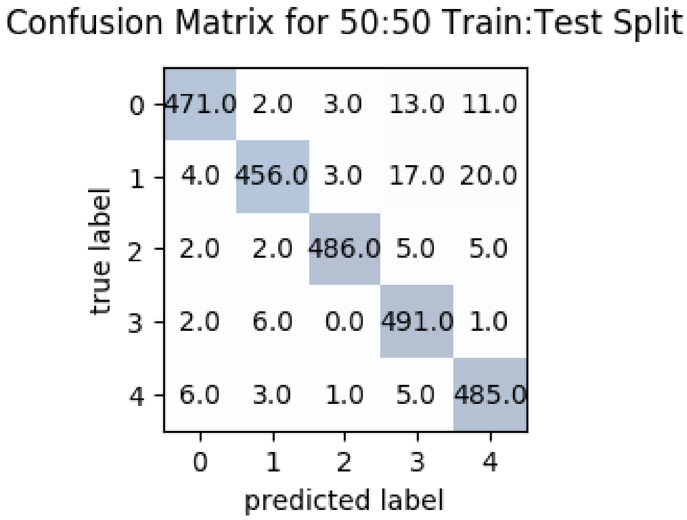
First the tf-idf of each term and document pair is computed and then top k terms from each documents are selected as features.

Training is performed using the new vocabulary (features).

K=10

50:50 Train Test Split

Vocabulary Size: 20972



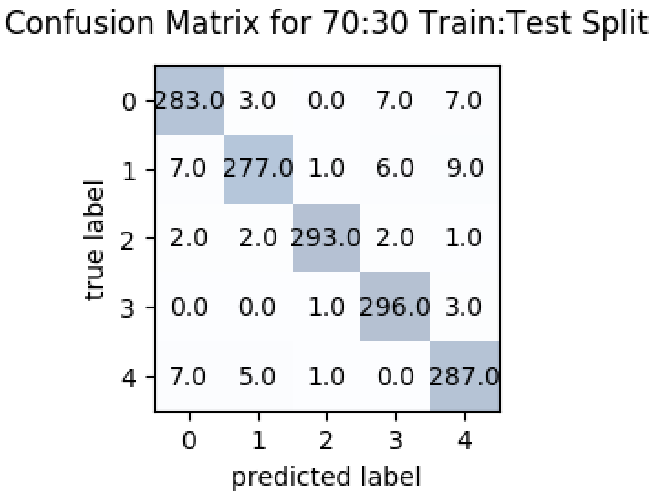
Confusion Matrix for 50:50 Train:Test Split

true label \ predicted label	0	1	2	3	4
0	471.0	2.0	3.0	13.0	11.0
1	4.0	456.0	3.0	17.0	20.0
2	2.0	2.0	486.0	5.0	5.0
3	2.0	6.0	0.0	491.0	1.0
4	6.0	3.0	1.0	5.0	485.0

Accuracy= 95.56%

70:30 Train Test Split

Vocabulary Size: 26173



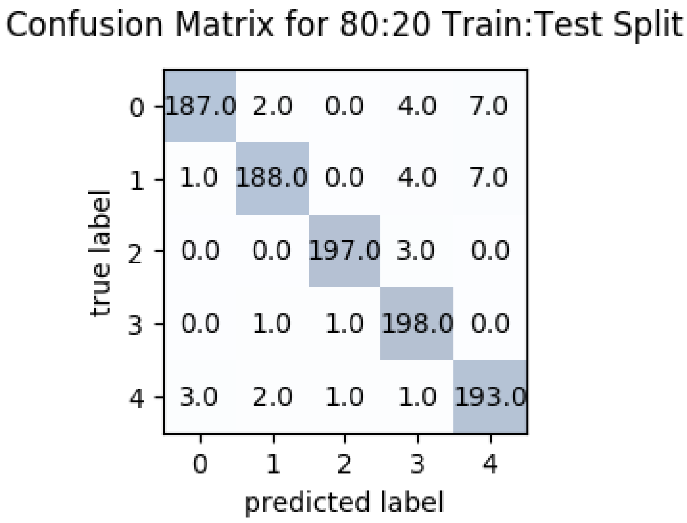
Confusion Matrix for 70:30 Train:Test Split

true label \ predicted label	0	1	2	3	4
0	283.0	3.0	0.0	7.0	7.0
1	7.0	277.0	1.0	6.0	9.0
2	2.0	2.0	293.0	2.0	1.0
3	0.0	0.0	1.0	296.0	3.0
4	7.0	5.0	1.0	0.0	287.0

Accuracy= 95.73%

80:20 Train Test Split

Vocabulary Size: 28128



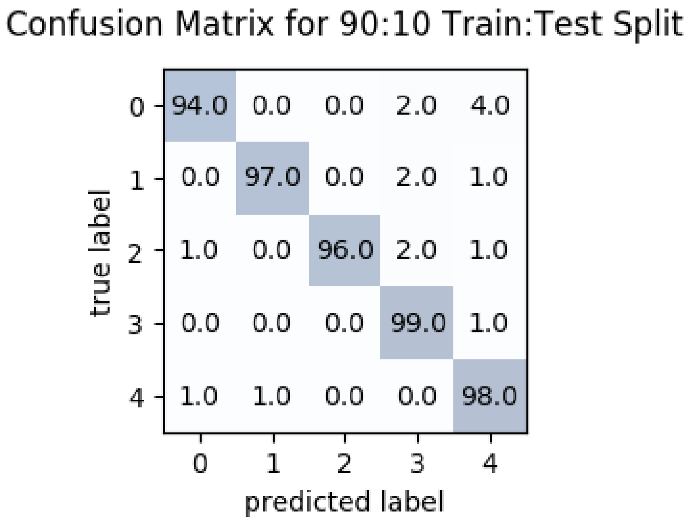
Confusion Matrix for 80:20 Train:Test Split

	0	1	2	3	4
0	187.0	2.0	0.0	4.0	7.0
1	1.0	188.0	0.0	4.0	7.0
2	0.0	0.0	197.0	3.0	0.0
3	0.0	1.0	1.0	198.0	0.0
4	3.0	2.0	1.0	1.0	193.0

Accuracy= 96.3%

90:10 Train Test Split

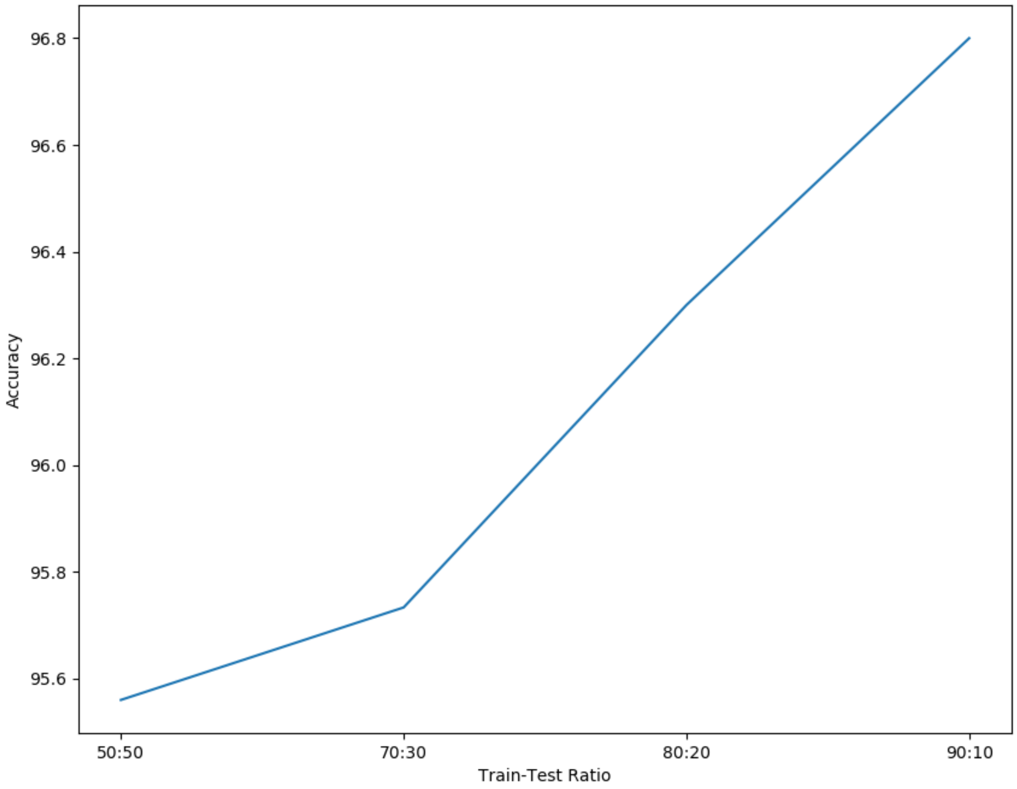
Vocabulary Size: 29930



Confusion Matrix for 90:10 Train:Test Split

	0	1	2	3	4
0	94.0	0.0	0.0	2.0	4.0
1	0.0	97.0	0.0	2.0	1.0
2	1.0	0.0	96.0	2.0	1.0
3	0.0	0.0	0.0	99.0	1.0
4	1.0	1.0	0.0	0.0	98.0

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.