Analysis

Information Retrieval
Assignment 3

Submitted by:
Himanshu Aggarwal
MT17015

Naive Bayes

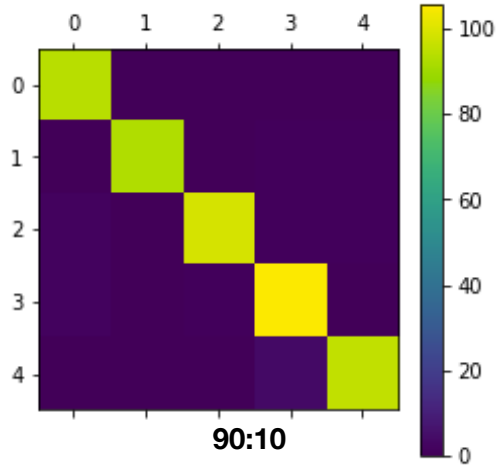
Below table shows accuracy of the Naive Bayes model on various splits:

Splits (Train:Test)	Accuracy
90:10	0.974051896208
80:20	0.976023976024
70:30	0.97601598934
50:50	0.972411035586

For optimal feature selection TF-IDF score is used. However, no significant change in accuracy is seen after applying this.

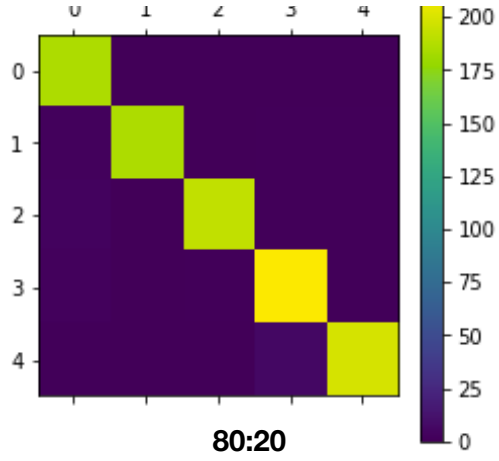
Following are the confusion matrices obtained in each split.

```
[[ 94   0   0   0   0]
 [  0  93   0   1   1]
 [  2   0  99   1   1]
 [  2   0   1 106   0]
 [  0   0   0   4  96]]
```



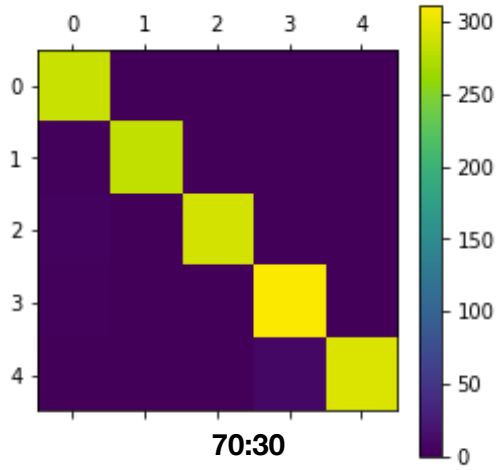
90:10

```
[[186   0   0   0   0]
 [  3 186   0   1   1]
 [  4   0 193   1   1]
 [  3   0   1 213   1]
 [  1   0   0   7 199]]
```



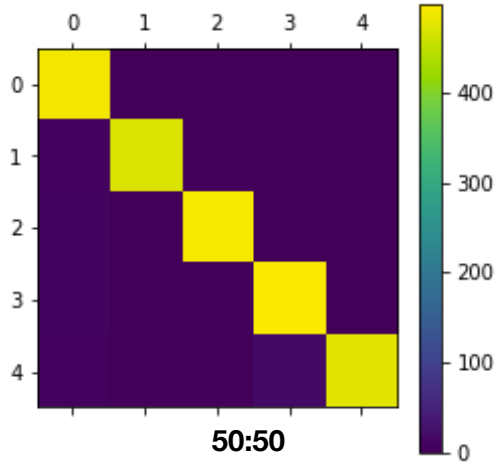
80:20

```
[[286   0   0   1   0]
 [  4 282   0   2   2]
 [  6   1 291   1   1]
 [  4   0   1 312   1]
 [  1   0   1  10 294]]
```



70:30

```
[[491   0   0   1   1]
 [  9 471   0   2   1]
 [ 11   1 494   2   2]
 [ 10   0   1 499   1]
 [  8   0   1  18 477]]
```



50:50

TF-IDF Score

The tf-idf score for a term in corresponding to a document has been computed using the formula:

tf-idf = tf X idf

where

tf = log(frequency) + 1 , if frequency>0

0 , otherwise

idf = log(number of document / df) + 1

where, df = number of documents where the term appears atlas once