COL773 Machine Learning Assignment 2

Manoj Kumar (cs5180411@cse.iitd.ac.in)

December 7, 2020

1. Text Classification

(a) By using Naive Bayes algorithm to classify articles, we get the following observations on test data:

```
Parameters for naive bayes:
Phi :  [0.15011255029240642, 0.08140826216365785, 0.10977018800759808, 0.21985446985446985, 0.43885452968186783]
size of the training_dictionary:  331826
calculating theta
All Parameters calculated
Testing on test data
total tests :  133718
Testing Completed
correct predictions:  77293
incorrect predictions:  56425
```

Accuracy = 57.8%

(b) Test Set Accuracy by Randomly predicting and by majority prediction :

```
time python q1b.py train.json test.json
total tests :  133718
Randomly Prediction
correct Prediction:  26681
incorrect Prediction:  107037
Majority Prediction
correct Prediction:  58822
incorrect Prediction:  74896
```

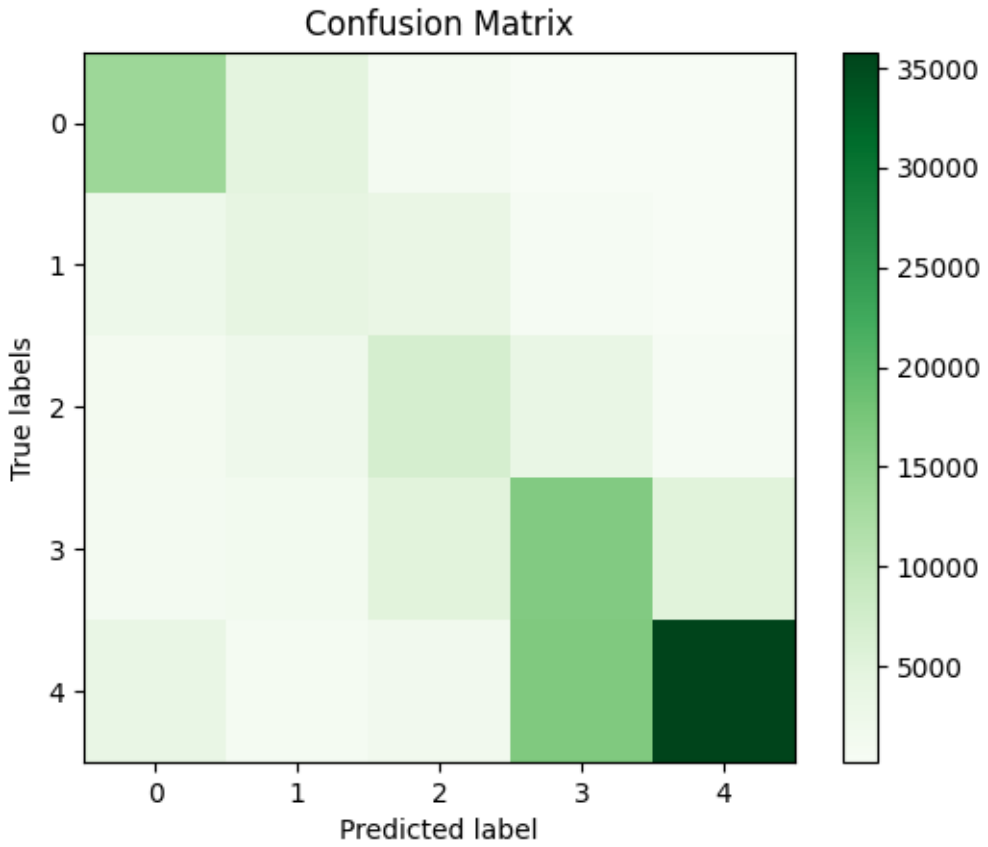
Accuracy in Random Prediction = **19.95**%

Accuracy in Majority Prediction = **43.99**%

Majority prediction is better than randomly prediction.

(c) Confusion Matrix: We get the following confusing matrix by using the parameters of q1 and testing on test data:

```
Confusion Matrix:
[[13894, 2525, 1177, 1057, 3563],
[4693, 4089, 2300, 1332, 864],
[1138, 3472, 6934, 5146, 1708],
[227, 579, 3580, 16514, 16825],
[217, 173, 540, 5309, 35862]]
```



(d) When we apply stemming and stopword removal, we get the following observations:

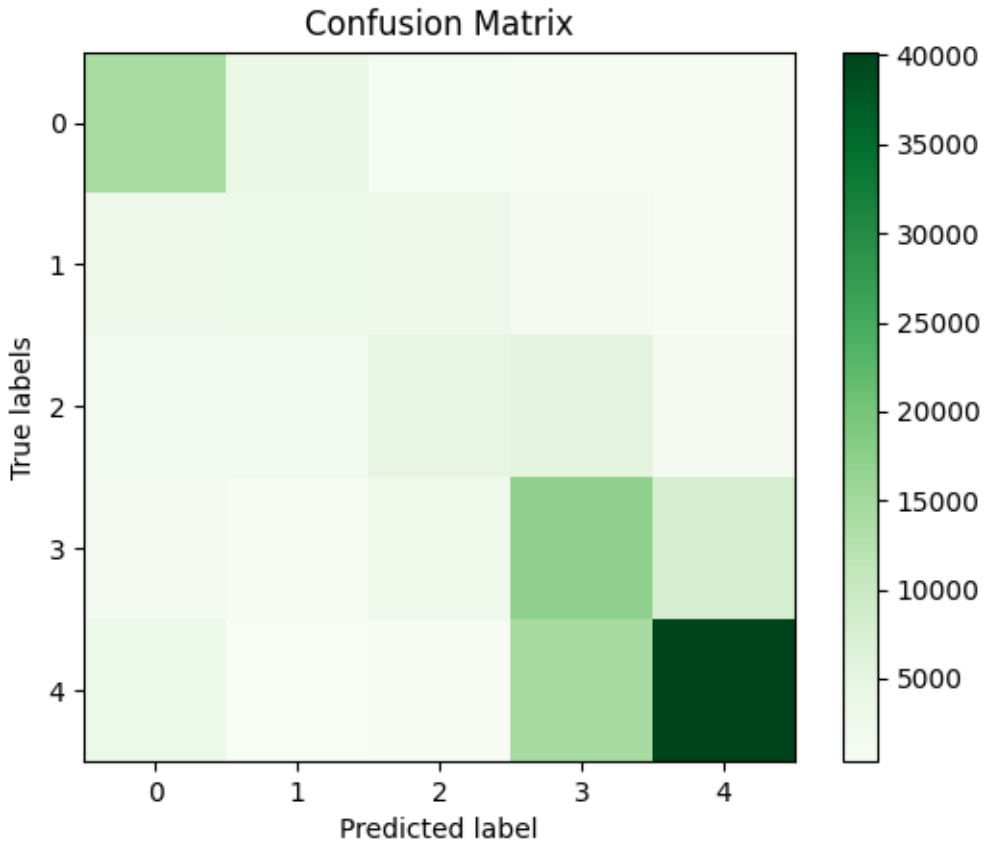
```
Stemming
Parameters for naive bayes:
Phi :  [0.15011255029240642, 0.08140826216365785, 0.10977018800759808,
0.21985446985446985, 0.43885452968186783]
size of the training_dictionary:  178637
calculating theta
All Parameters calculated
Testing on test data with stemming
Testing Completed
correct predictions:  79771
incorrect predictions:  53947
```

Accuracy: 59.66%

Confusion Matrix:

```
Confusion Matrix:
[[14348, 2954, 1471, 1226, 3203],
[3709, 3136, 1600, 644, 315],
[1098, 2981, 4819, 2405, 615],
[521, 1268, 5417, 17429, 14650],
```

```
[493, 499, 1224, 7654, 40039]]
```



Observations: Stemming a document increased the accuracy by 2%.

(e) **Feature Engineering:**

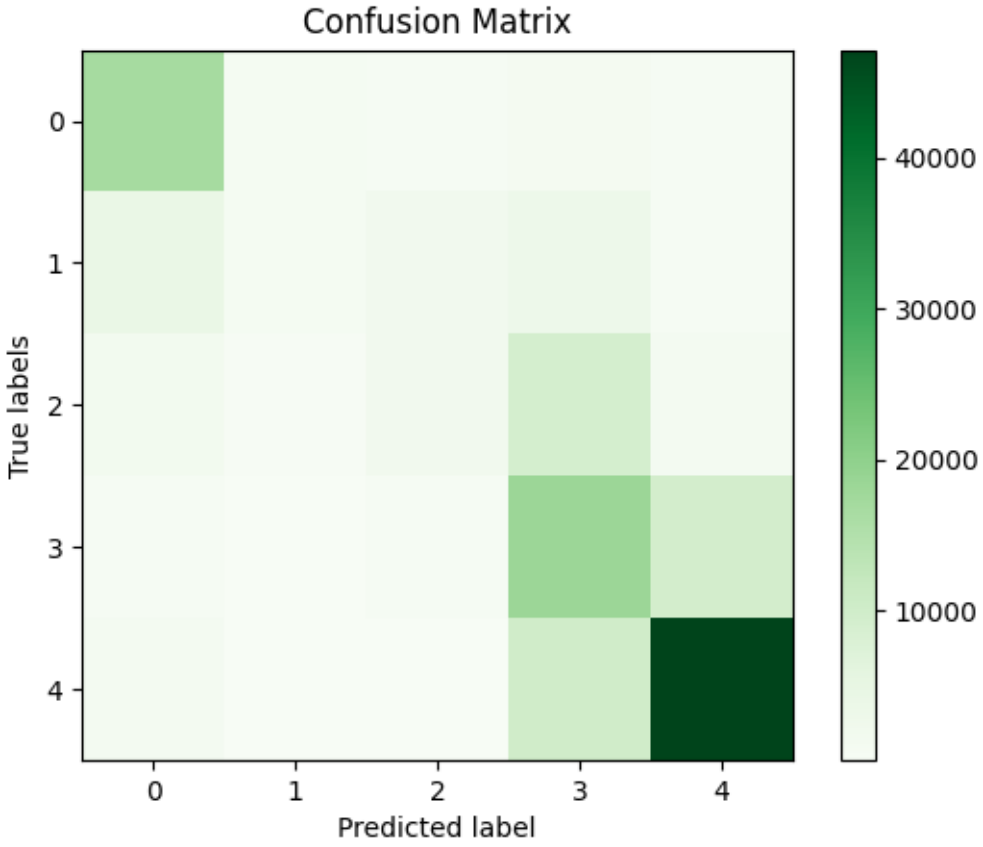
(i) **Bigrams without stemming:** When we apply bigrams without stemming on our training data as well as test data, we get the following observations:

```
Feature is : Bigrams
Parameters for naive bayes:
Phi :  [0.15011255029240642, 0.08140826216365785, 0.10977018800759808,
 0.21985446985446985, 0.43885452968186783]
size of the training_dictionary:  6411058
calculating theta
All Parameters calculated
Testing on test data with stemming and stopwords removing
Testing Completed
correct predictions:  85461
incorrect predictions:  48257
```

Accuracy : 63.9%

And the confusion matix as:

```
Confusion Matrix:
[[16874, 4215, 1706, 788, 1247],
[979, 928, 279, 74, 88],
[793, 2041, 2093, 486, 224],
[1067, 3154, 9260, 18441, 10138],
[456, 500, 1193, 9569, 47125]]
```



(ii) **Bigrams and Stemming together:**

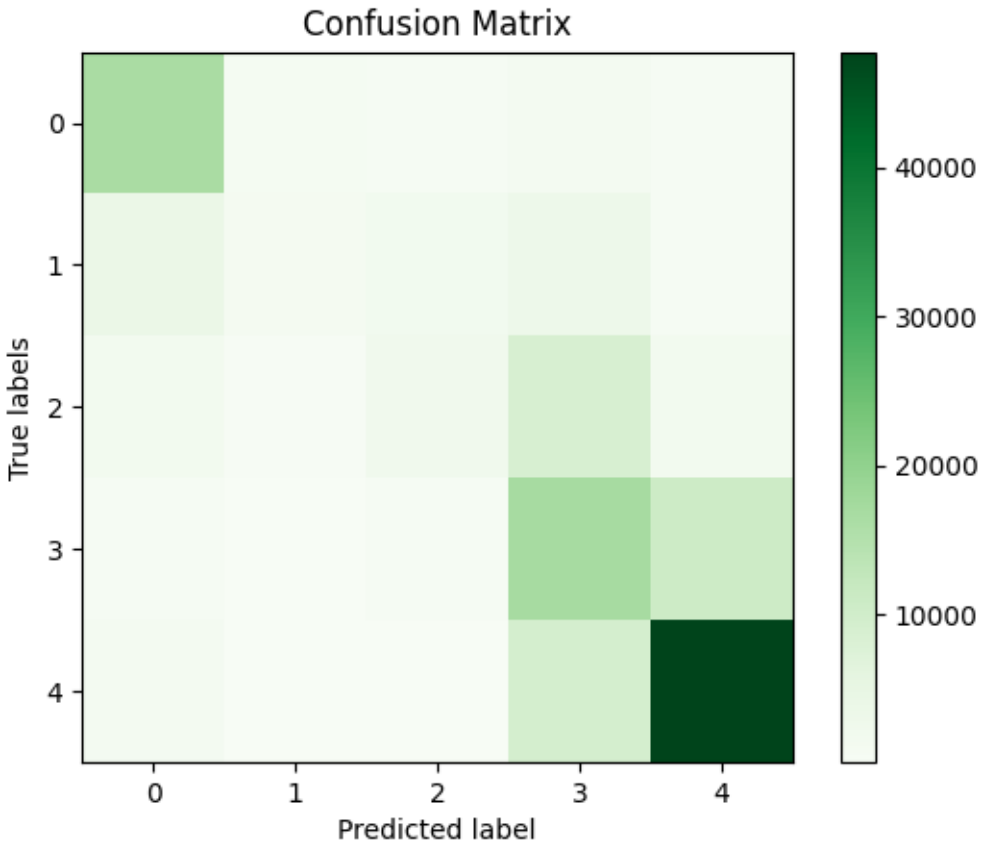
When we stemmed the words and then added two consecutive words (bigrams) we get the following observations:

```
Feature is : Bigrams with stemming
Parameters for naive bayes:
Phi :  [0.15011255029240642, 0.08140826216365785, 0.10977018800759808,
0.21985446985446985, 0.43885452968186783]
size of the training_dictionary:  6062080
calculating theta
All Parameters calculated
Testing on test data with stemming and stopwords removing
Testing Completed
correct predictions:  84568
incorrect predictions:  49150
```

Accuracy: 63.24%

And the confusion matrix is:

```
Confusion Matrix:
[[16513, 4010, 1658, 813, 1324],
[958, 1067, 309, 89, 92],
[791, 1825, 2181, 630, 267],
[1225, 3179, 8662, 17031, 9363],
[682, 757, 1721, 10795, 47776]]
```



Observations: We can observe that feature bigram performs better without stemming. Although, stemming is good if we do not join 2 consecutive words.

Thanks
Manoj Kumar
2018CS50411