[CSE-508] Information Retrieval - Assignment 5

- *Shreyash Arya (2015097)*

Question 1

1. Bag of Words

For the bag of words part, the preprocessing has been done in a similar fashion as was done in the previous assignments.

Final Vocab Size = 68215
Purity = 0.49
RSS = 7236.88

2. Word2Vec

The documents are converted to the vectors using the Gensim package for Doc2vec package.

Purity = 0.85
RSS = 165283.932

From the above results, we can see that the Bag of Words approach does better clustering as there is a difference in the words present in all 5 classes documents.

Question 2

Word2Vec is used for this part for getting the document vectors.

Results (Accuracy in %):

	Naive Bayes	k=1	k=3	k=5
50:50	95.26	27.11	53.51	79.98
80:20	96.11	27.63	55.36	81.35
90:10	98.01	28.43	54.91	85.03

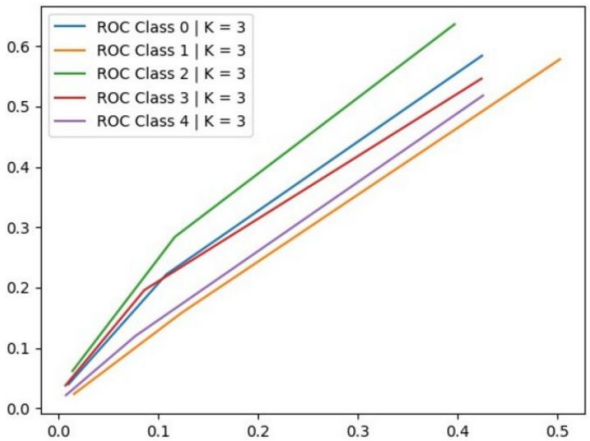
We can clearly see from the results that the Navie Bayes outperforms KNN with the highest accuracies. In KNN, as we increase the k values the data increases and reduces the chance of mis-classifying the class label.

Confusion Matrix and ROC Curve

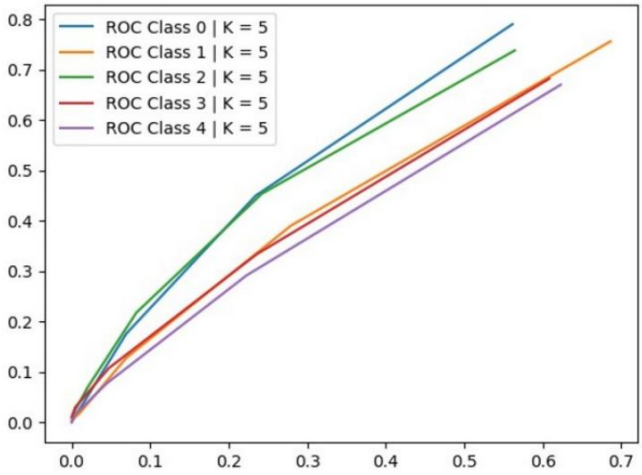
Ratio = 50:50

k=1
True vs Predicted
[[147. 97. 57. 90. 104.]
[120. 120. 111. 98. 98.]
[78. 90. 155. 80. 97.]
[72. 93. 89. 149. 82.]
[88. 93. 87. 84. 123.]]

k=3
True vs Predicted
[[235. 183. 111. 200. 151.]
[132. 146. 145. 122. 144.]
[58. 86. 155. 72. 97.]
[44. 44. 47. 60. 45.]
[31. 41. 42. 46. 63.]]



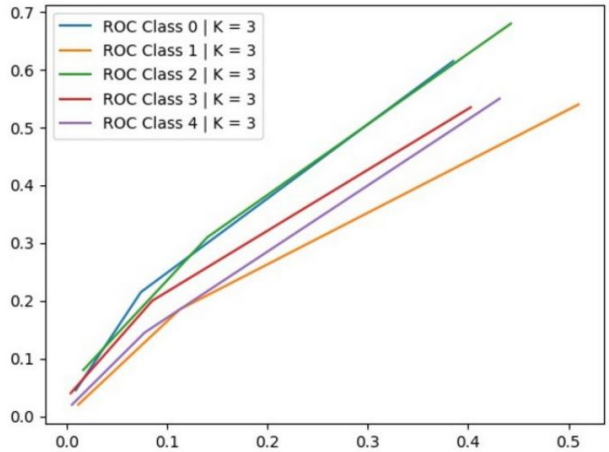
k=5
True vs Predicted
[[215. 155. 71. 152. 112.]
[126. 134. 129. 101. 138.]
[64. 92. 181. 102. 128.]
[56. 62. 70. 99. 58.]
[39. 57. 49. 46. 64.]]



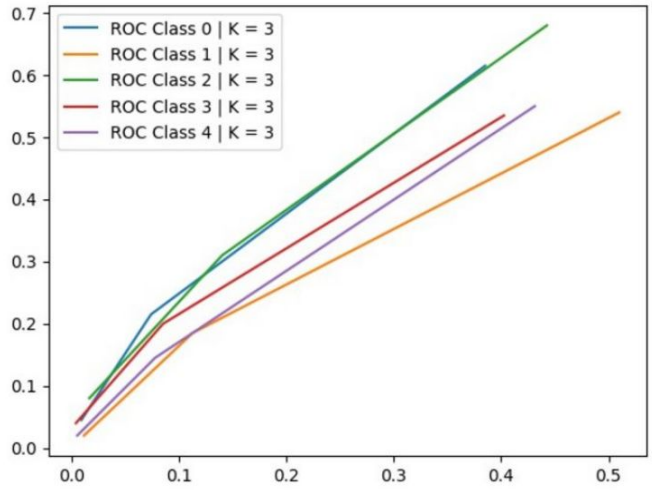
Ratio = 80:20

k=1
True vs Predicted
[[92. 70. 39. 59. 52.]
[61. 75. 52. 50. 57.]
[46. 59. 100. 63. 75.]
[40. 52. 55. 79. 50.]
[61. 44. 54. 49. 66.]]

k=3
True vs Predicted
[[146. 110. 78. 105. 110.]
[66. 100. 77. 68. 82.]
[41. 44. 98. 35. 51.]
[20. 27. 27. 60. 20.]
[25. 19. 20. 32. 37.]]



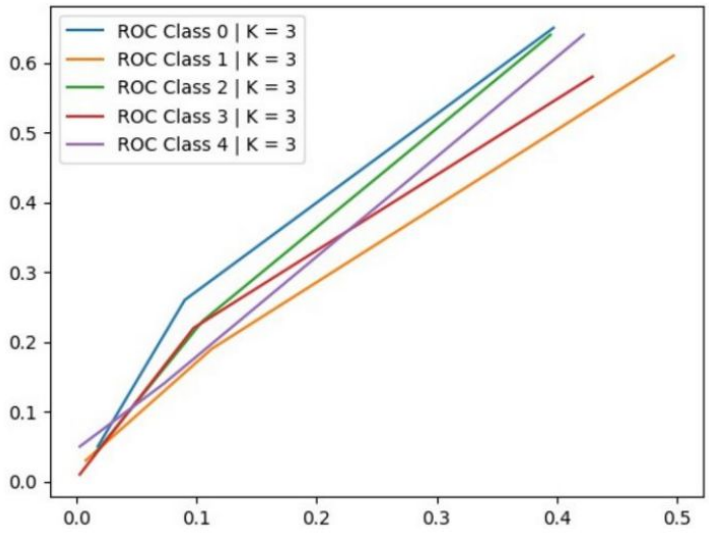
k=5
True vs Predicted
[[126. 79. 51. 97. 73.]
 [60. 80. 65. 68. 67.]
 [40. 51. 121. 71. 82.]
 [40. 34. 40. 64. 35.]
 [28. 36. 26. 23. 43.]]



Ratio = 90:10

k=1
True vs Predicted
[[34. 21. 4. 17. 13.]
 [17. 25. 23. 15. 21.]
 [18. 19. 36. 24. 30.]
 [15. 20. 14. 28. 18.]
 [16. 15. 23. 16. 18.]]

k=3
True vs Predicted
[[47. 35. 24. 31. 23.]
 [23. 30. 27. 28. 31.]
 [12. 18. 30. 17. 22.]
 [6. 5. 8. 14. 11.]
 [12. 12. 1. 10. 13.]]



k=5

True vs Predicted

[[40. 25. 21. 30. 17.]
[22. 34. 28. 26. 24.]
[13. 19. 27. 16. 28.]
[14. 12. 12. 22. 13.]
[11. 10. 12. 6. 18.]]