Assignment 2
Text Classification

Rituparna Datta
1505091

October 16, 2020

1 On Validation set of 2200 documents, 11 classes

K-NN	K=1	K=3	K=5
Hamming Distance	28.955%	31.727%	30.955%
Euclidean Distance	51.364%	50.136%	51.409%
Cosine Similarity	78.7%	80.25%	81.6%

2 Naive Bayes on Validation set of 2200 documents using 10 smoothing factor

α	Accuracy
1.5	74.909%
1	76.227%
0.5	78.54%
0.2	80.318%
0.1	81.59%
0.07	82.5%
0.05	82.59%
0.01	83.36%
0.005	83.68%
0.0001	82.8636%

As from the accuracy analysis, given in the matrices, it can be concluded that, Cosine Similarity using TF-IDF weight for K=5 value provides the best accuracy value among the rest k-nearest neighbour methods. And from Naive Bayes analysis, for alpha = 0.005 gives the best accuracy.
Now, I would choose the approaches which performed best, and run the test data of 5500 documents of 11 topics, for 50 iterations. In each iteration, 10 documents from each topic would be chosen and then run in both NB (alpha = 0.005) and KNN (k=5) methods and saved the accuracy in text files.

3 KNN and Naive Bayes Comparative Analysis

3.1 50 iterations on Test Data

iteration	Naive Bayes	Cosine Simulation
1	86.36363636363636	85.45454545454545
2	87.27272727272727	84.54545454545455
3	82.72727272727273	84.54545454545455
4	84.54545454545455	86.36363636363636
5	81.81818181818183	83.63636363636363
6	86.36363636363636	87.27272727272727
7	91.81818181818183	85.45454545454545
8	89.0909090909091	81.81818181818183
9	79.0909090909091	80.0
10	82.72727272727273	80.0
11	85.45454545454545	82.72727272727273
12	90.0	80.0
13	73.63636363636363	72.72727272727273
14	82.72727272727273	82.72727272727273
15	82.72727272727273	83.63636363636363
16	73.63636363636363	72.72727272727273
17	80.0	77.27272727272727
18	80.9090909090909	78.18181818181819
19	83.63636363636363	77.27272727272727
20	78.18181818181819	80.9090909090909
21	75.45454545454545	77.27272727272727
22	82.72727272727273	88.18181818181819
23	84.54545454545455	81.81818181818183
24	85.45454545454545	84.54545454545455
25	81.81818181818183	81.81818181818183
26	76.36363636363637	80.9090909090909
27	85.45454545454545	83.63636363636363
28	79.0909090909091	77.27272727272727
29	78.18181818181819	81.81818181818183
30	82.72727272727273	80.9090909090909
31	84.54545454545455	80.0
32	81.81818181818183	83.63636363636363
33	81.81818181818183	86.36363636363636
34	87.27272727272727	80.0
35	85.45454545454545	73.63636363636363
36	79.0909090909091	77.27272727272727
37	79.0909090909091	78.18181818181819
38	83.63636363636363	77.27272727272727
39	90.9090909090909	89.0909090909091
40	87.27272727272727	81.81818181818183
41	82.72727272727273	80.9090909090909
42	86.36363636363636	83.63636363636363
43	80.0	80.0
44	84.54545454545455	79.0909090909091
45	87.27272727272727	82.72727272727273
46	87.27272727272727	82.72727272727273
47	84.54545454545455	83.63636363636363
48	79.0909090909091	77.27272727272727
49	84.54545454545455	85.45454545454545
50	88.18181818181819	87.27272727272727

After running 50 iterations, the average value of cosine similarity, 5-nearest neighbours is **81.51%** and the average value of Naive Bayes using alpha = 0.005, is **83.20%**.
Now, for t-stat analysis, I've imported **stats from scipy** and then run **ttest_rel()** function on 50 accuracy values of both Naive Bayes and Cosine Similarity and found **statistic = 3.3134598364347445**, **p-value = 0.0017378416481416729**

3.2 Statistical Significance and P-value

1. A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

2. Significance level 0.01 means statistically significant as $P < 0.01$ (less than one in a 100 chance of being wrong).

3. Significance level 0.005 means statistically significant as $P < 0.005$ (less than one in a 200 chance of being wrong).

From the observation of t-stat of Naive Bayes and Cosine similarity, p-value≈0.0017 which indicates it is statistically significant. And for the mentioned significance values : 0.05, 0.01, 0.005, it rejects the null hypothesis that the means are equal as p_value is less than 3 of them.