K-NN on validation set

Distance Metric	k=1	k=3	k=5
Hamming	36.36363636363637	38.59090909090909	37.31818181818182
Euclidean	56.31818181818182	57.40909090909091	55.0
Cosine similarity using TF-IDF vectors	81.0909090909091	83.22727272727273	83.5

Naive Bayes on validation set

Smoothing factor,α	Accuracy (%)
0.0001	90.68181818181819
0.001	91.54545454545455
0.01	92.22727272727273
0.025	92.22727272727273
0.05	92.22727272727273
0.065	92.31818181818181
0.075	92.27272727272727
0.1	92.18181818181819
0.5	91.9090909090909
1	91.45454545454545
10	85.54545454545455
100	79.5909090909091

Comparing K-NN and Naïve Bayes on test set:

Iteration	K-NN (TF-IDF for K=5)	Naïve Bayes ($\alpha = 0.065$)
1	89.0909090909091	91.81818181818181
2	88.18181818181819	95.45454545454545
3	84.54545454545455	96.36363636363636
4	81.81818181818181	90.0
5	87.27272727272727	95.45454545454545
6	88.18181818181819	92.72727272727273
7	87.27272727272727	95.45454545454545
8	87.27272727272727	91.81818181818181
9	81.81818181818181	91.81818181818181
10	82.72727272727273	91.81818181818181
11	83.63636363636364	93.63636363636364
12	85.45454545454545	95.45454545454545
13	76.36363636363636	89.0909090909091
14	80.9090909090909	90.9090909090909
15	80.0	90.9090909090909
16	77.27272727272727	89.0909090909091
17	80.0	90.9090909090909
18	79.0909090909091	92.72727272727273
19	76.36363636363636	90.9090909090909
20	79.0909090909091	88.18181818181819
21	80.0	84.54545454545455
22	87.27272727272727	97.27272727272727
23	85.45454545454545	93.63636363636364
24	81.81818181818181	94.54545454545455
25	82.72727272727273	90.9090909090909
26	84.54545454545455	89.0909090909091
27	82.72727272727273	88.18181818181819
28	78.18181818181819	90.0
29	81.81818181818181	89.0909090909091
30	80.9090909090909	94.54545454545455
31	84.54545454545455	90.0
32	80.9090909090909	92.72727272727273
33	85.45454545454545	92.72727272727273
34	82.72727272727273	94.54545454545455
35	81.81818181818181	89.0909090909091
36	80.0	91.81818181818181
37	80.9090909090909	89.0909090909091
38	79.0909090909091	91.81818181818181
39	87.27272727272727	96.36363636363636
40	89.0909090909091	96.36363636363636
41	77.27272727272727	89.0909090909091
42	88.18181818181819	96.36363636363636
43	80.0	90.0
44	83.63636363636364	90.0
45	81.81818181818181	92.72727272727273
46	85.45454545454545	93.63636363636364
47	82.72727272727273	90.9090909090909
48	80.0	89.0909090909091
49	82.72727272727273	88.18181818181819
50	82.72727272727273	95.45454545454545

Summary of test set accuracies:

K-NN on test set:(using Cosine similarity on TF-IDF vector and k=5)
Min: 76.36363636363636
Max: 89.0909090909091
Mean: 82.76363636363637
Std Dev: 3.4013123378829406

Naive Bayes on test set:(using 0.065 as the smoothing factor)
Min: 84.54545454545455
Max: 97.27272727272727
Mean: 91.92727272727272
Std Dev: 2.785084238380596

Result of t statistics :

If the p-value is smaller than the threshold, then we reject the null hypothesis of equal averages. Small p-values are associated with large t-statistics. Following result was calculated by ***stats.ttest_rel (test_set_accuracies_knn, test_set_accuracies_nb)*** from **scipy** library. Relative t-test was used because we wanted to compare on each test set pairwise (KNN and Naïve Bayes).

T_statistic*= -23.120636250963607,* ***pvalue****= 5.1705019279572745e-28*

T_statistic value (as it is negative) tells us Naïve Bayes is better than K-NN. This p-value is smaller than 0.005. So we reject the null hypothesis of equal averages at significance level of 0.005.Same goes for significance levels of 0.01 and 0.05. Such a small P-value also states that Naïve Bayes is consistently better than K-NN.

Summary:

Performance:

Naïve Bayes is better than K-NN looking at the averages of the accuracies run on 50 iterations of test set. Naïve Bayes takes probability of all the words of test set in training space into account. K-NN is really bad at prediction in hamming and Euclidean distance measures. It increases significantly in cosine similarity with TF-IDF weights.

Running time:

Naïve Bayes took around 2s on validation set whereas KNN took almost ~50s. Naïve Bayes is faster because KNN needs more real time computation than Naïve Bayes.