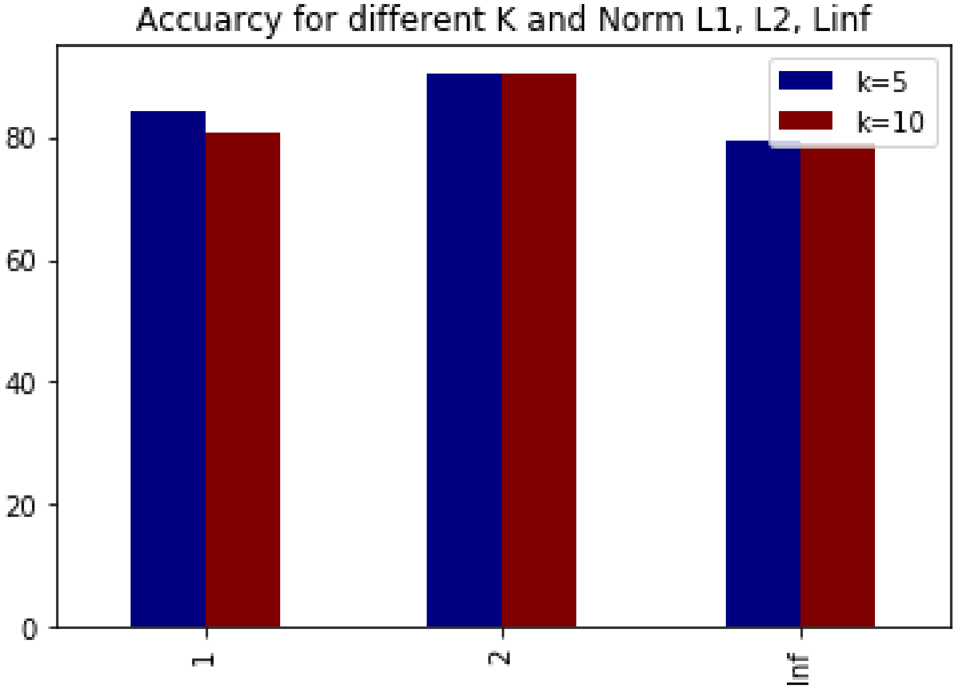
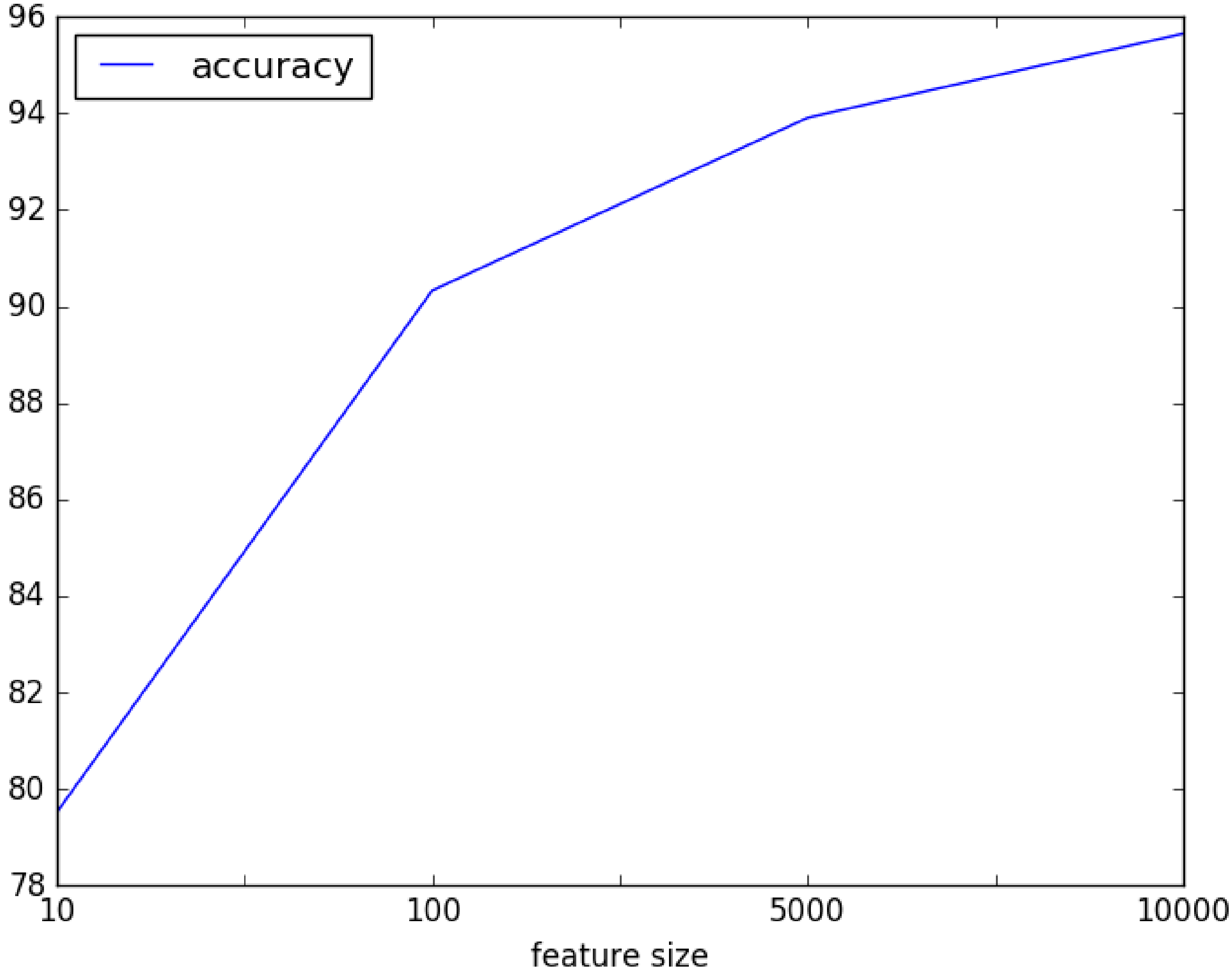
Problem 6

We first consider how the classifiers perform for their respective hyperparameters.
For knn we check the performance with respect to two parameters: k and norm
1. k: two values of k are checked: 5 and 10.
Not a lot of change is seen over 5 and 10. We observe that euclidean distance gives the best results and 5 or 10 don't necessarily make a difference. Perhaps the lack of difference between the two chosen values of k is owing to how small the difference is.



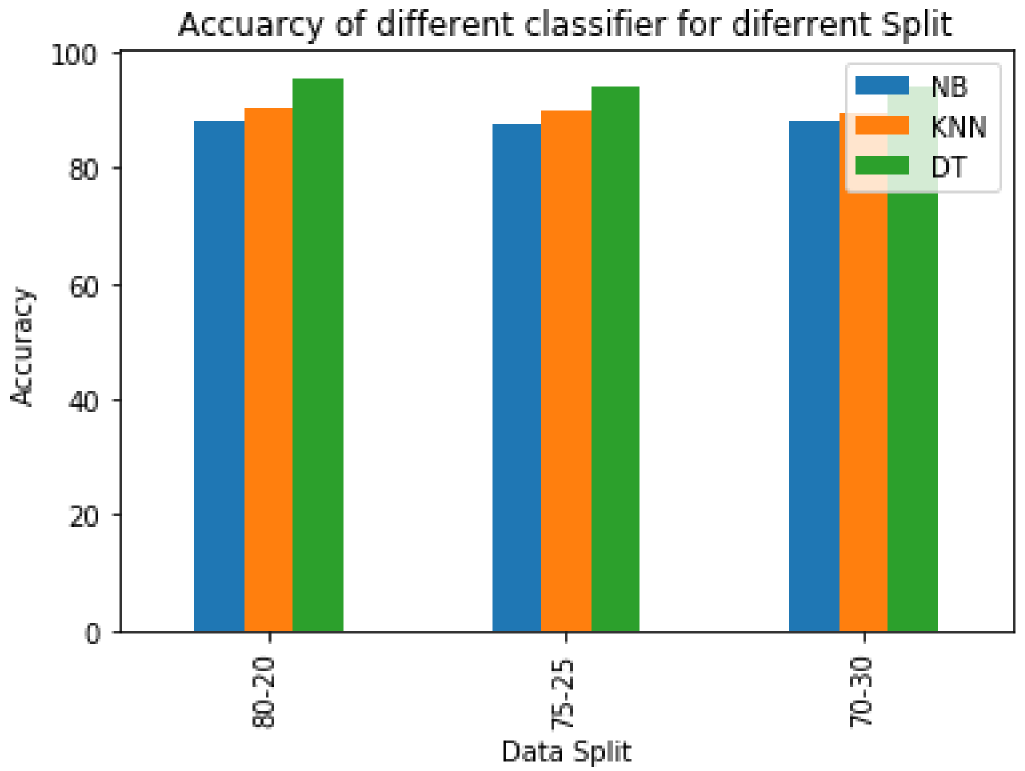
2. norm: We observe that changing the norm affects the accuracy a lot. Primarily, L2 norm is seen to be performing the best.

For decision trees, we observe that change in number of words being used as features does impact the accuracy substantially. Within our tested number of attributes, as we increase the number of attributes, the accuracy improved. We went up to 10,000 attributes.



Naive Bayes Naive Bayes, which was theoretically expected to give optimal accuracy seems to be lagging in our tests. We expect our cleaning methods to be the source of this discrepancy. Naive Bayes performance on various data size splits is shown in the graph comparing different classifiers.

Overall, we observe that the decision tree with 10k words would be the optimal classifier for the given dataset. Given below is a graph of performance of various classifiers on different train:test splits. The splits involve 80 train, 20 test; 75 train 25 test; 70 train, 30 test.


Accuarcy of different classifier for diferrent Split
Accuracy
0
20
40
60
80
100
80-20
75-25
70-30
Data Split
NB
KNN
DT