Homework Assignment 2

Weerdhawal Chowgule

February 25, 2018

1 Data-Preprocessing

In the previous assignment we created a Document-Term matrix (DTM) but that did not contain the tag of which category it belonged to. So, I followed the Preprocessing procedure once again and added the categories tag. Also, the sparsity was considered as 0.99, but as the matrix was to huge and it was taking a lot of time for the 5-fold cross validation. Hence, I changed the sparsity to 0.85.

2 Feature Selection & Data Description Preprocessing

The DTM consists of 18828 documents and 270 features (i.e. the words). It contains a column named 'mainCATEGORIES' which contains the category tag of the respective document. The documents are arranged in a continuous order i.e. Consider there are 5000 documents across 4 categories then the first 1200 are from one category then the next 1300 from another category and so on.

Feature selection was performed, and top 100 words were selected. A list of top features were found in the last assignment, I took the column(feature) names and extracted only those columns from the main DTM and a file named 'checkfinal.csv' is obtained.

3 Packages and Libraries

The program has been scripted in Python 3.6 and a varied list of modules were used to generate the results

- **SciPy** : Used for mathematical computations
- **NumPy** : Used for mathematical computations
- **SciKit-Learn** : Machine learning module for python

- **Matplotlib** : Used for visualizing the data
- **Seaborn** : Used for visualizing the data
- **PyPDF2** : Used to save the Decision Trees and to merge the files

4 Knn & Decision Tree with 5-fold Cross validation

Knn and Decision Tree algorithms were applied on both the files (main and top 100) data using sklearn. In Knn algorithm for each fold I calculated the best K value in the range of 1 to 20 and applied it to the Knn algorithm. There were no alterations made for the Decision Tree algorithm, the original decision tree algorithm was applied. The out put of the is a Pdf file of the Decision tree and is attached in the folder submitted.

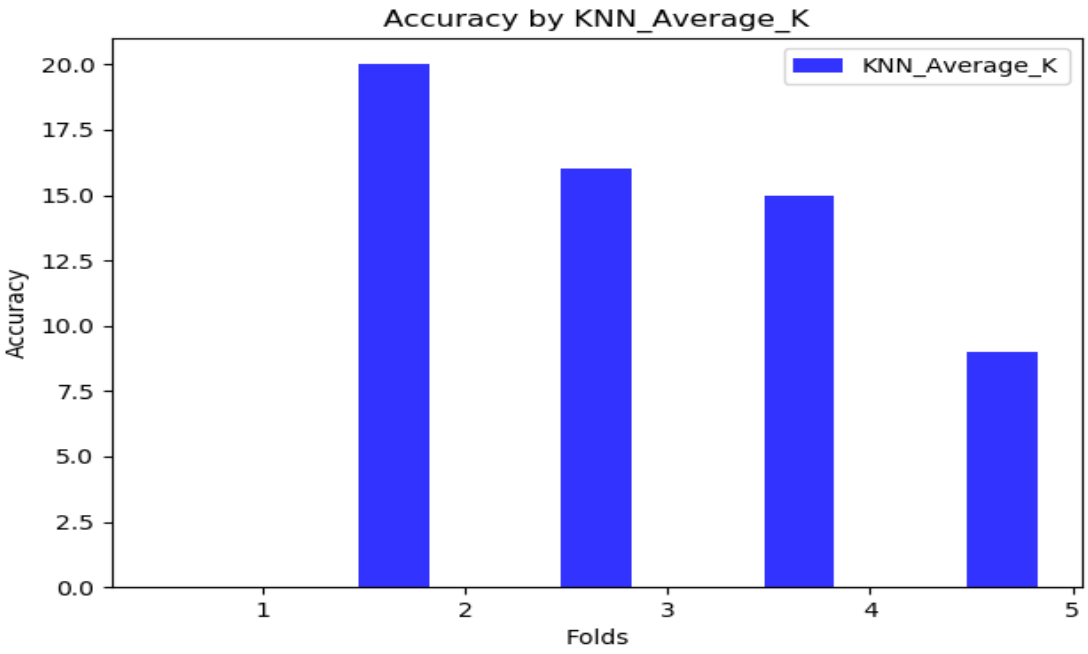

Figure 1: KNN Accuracy V/s Folds (Without Feature selection)

The Figures 1 and 2 represent the accuracies of Knn and Decision Tree Algorithm during the 5-fold Cross Validation without feature selection. We see that maximum accuracy of Knn algorithm is around 20% and the average of the accuracies(5-Fold) being 12% where as for the Decision Tree Algorithm maximum is 14% and the average of accuracies(5-Fold) being 9.6%.

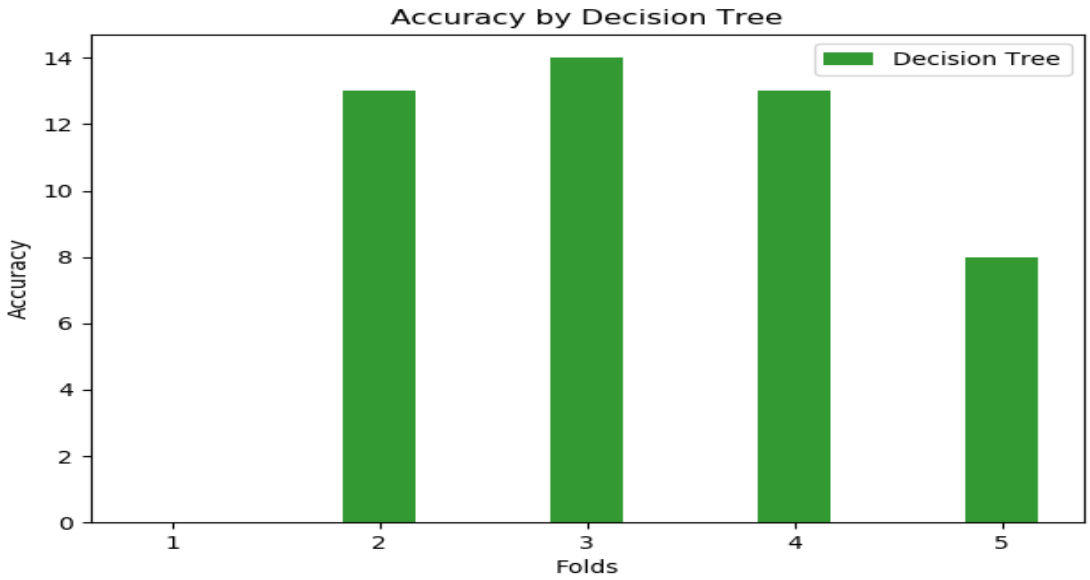

Figure 2: Decision Tree Accuracy V/s Folds (Without Feature selection)

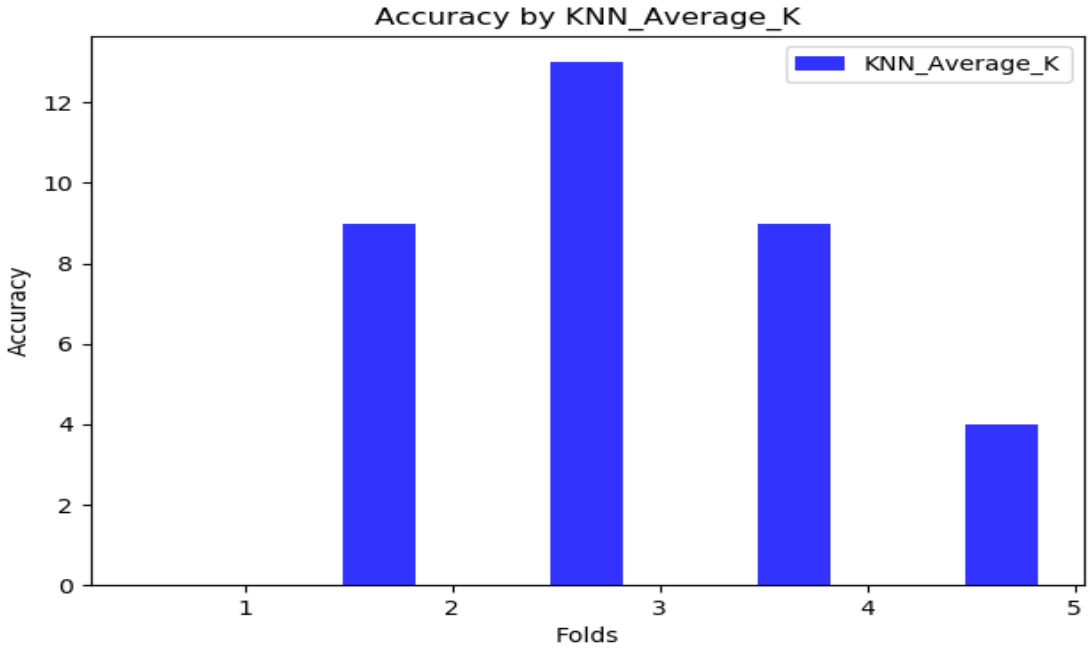

Figure 3: KNN Accuracy V/s Folds (With Feature selection)

The Figures 3 and 4 represent the accuracies of Knn and Decision Tree Algorithm during 5-fold Cross validation with feature selection. We see that

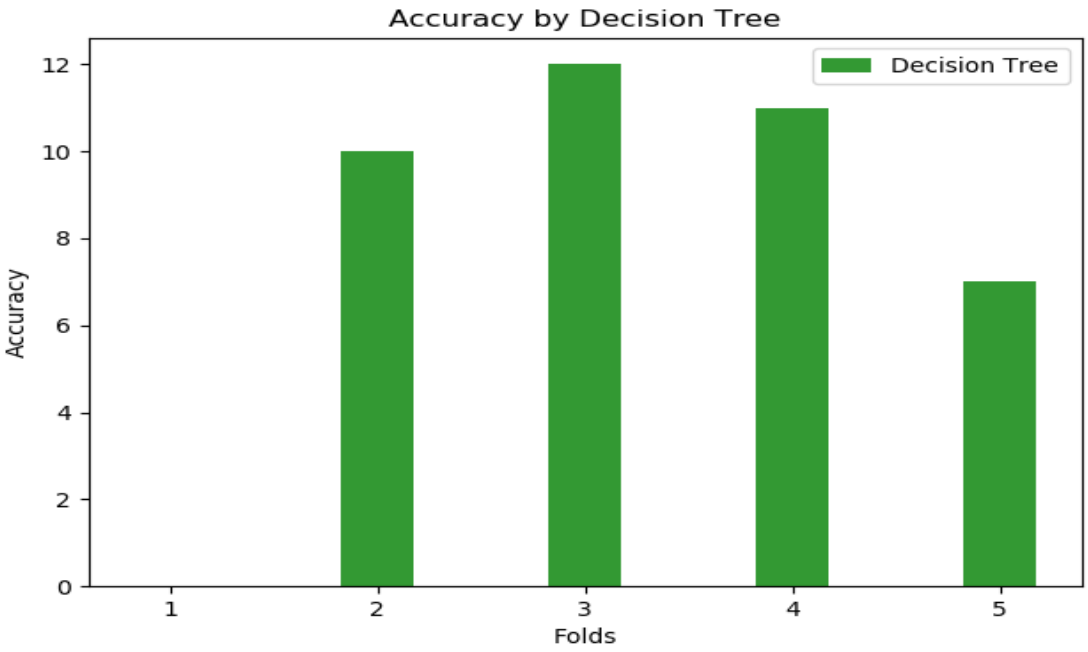

Figure 4: Decision Tree Accuracy V/s Folds (With Feature selection)

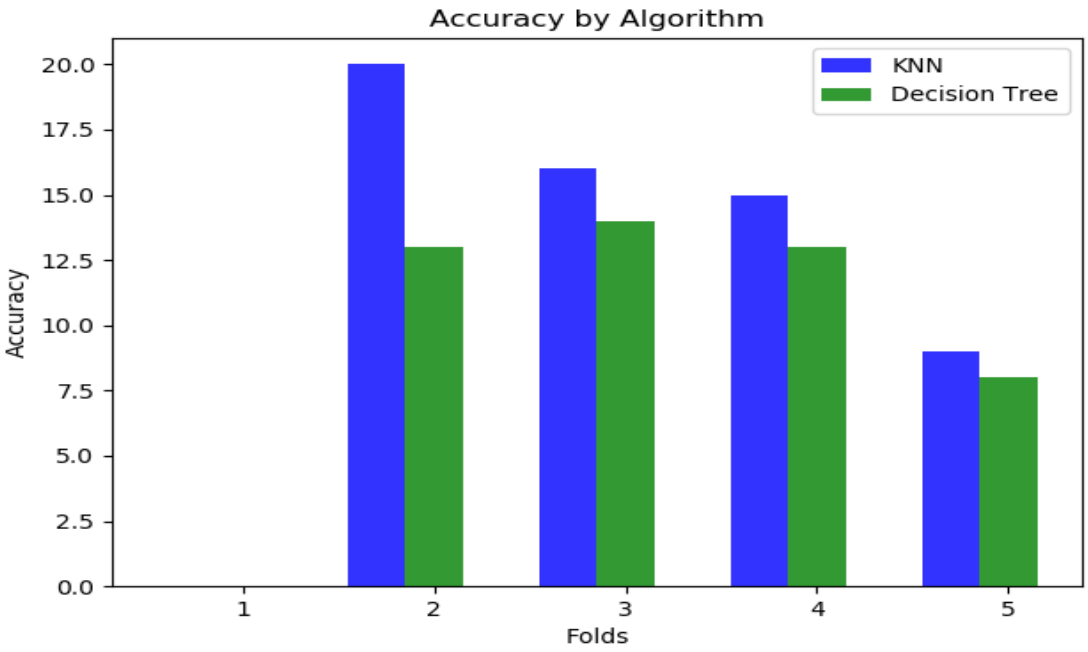

Figure 5: Accuracy V/s Folds (Without Feature selection)

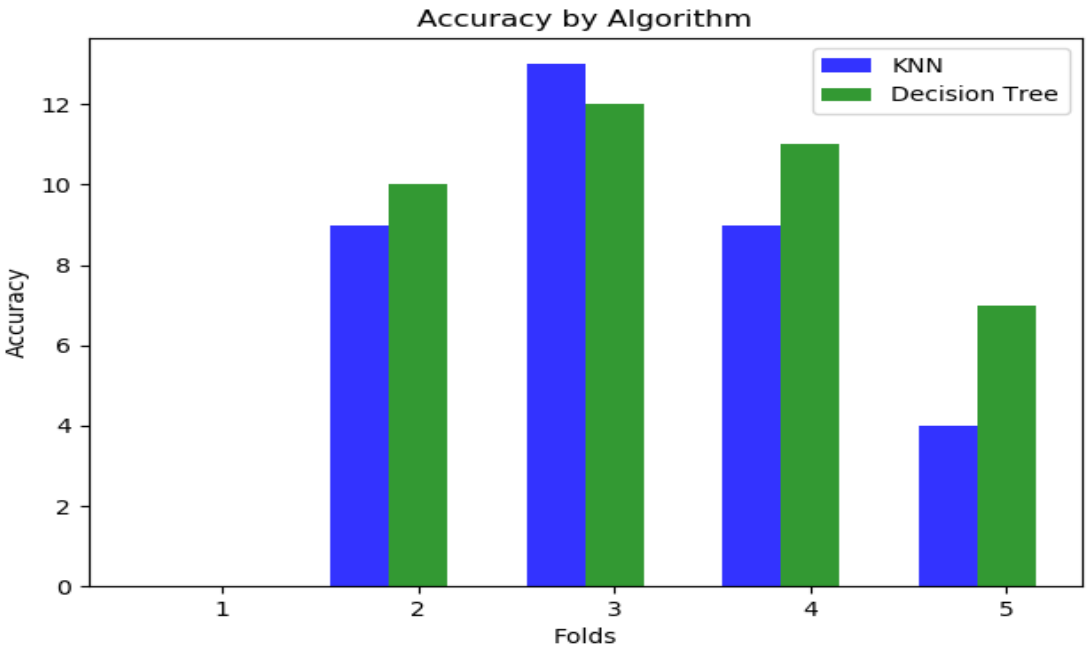

Figure 6: Accuracy V/s Folds (With Feature selection)

the maximum accuracy of Knn is 13% and the average being 7% where as for Decision tree maximum accuracy is 12% but the average being 8%.

As we can see from the comparison graphs in Figures 5 and 6 we can observe that using 5-fold cross validation is a good technique, as the data set is divided in 5 parts and training and testing is performed considering each part, which gives different accuracies for the same data set.

From the above observations we can say that the selecting a few features from a comparatively larger datasets would give less accuracy. Hence concluding that Feature Selection is not a good criteria to be used for Knn and Decision Tree Algorithms. Also we can see that Knn Algorithm performs better than that of Decision Tree Algorithm both the times i.e. before and after feature selection.

The figures 7 and 8 are screen shots of the last parts of the execution of the programs. We can see the calculated f-score, precision and recall from the last phase of the 5-fold cross validation. Due to not precise distribution of the data in the fscore obtained there are only a few values of only certain categories. The graph in Figure 9 is a plot of Accuracy Vs Value of K in the Knn Algorithm. I performed it on the DTM after feature selection by choosing the test and train from the highest of the 5 fold. We can see that the maximum is 13% and has a kind of linear fall as K increases and becomes

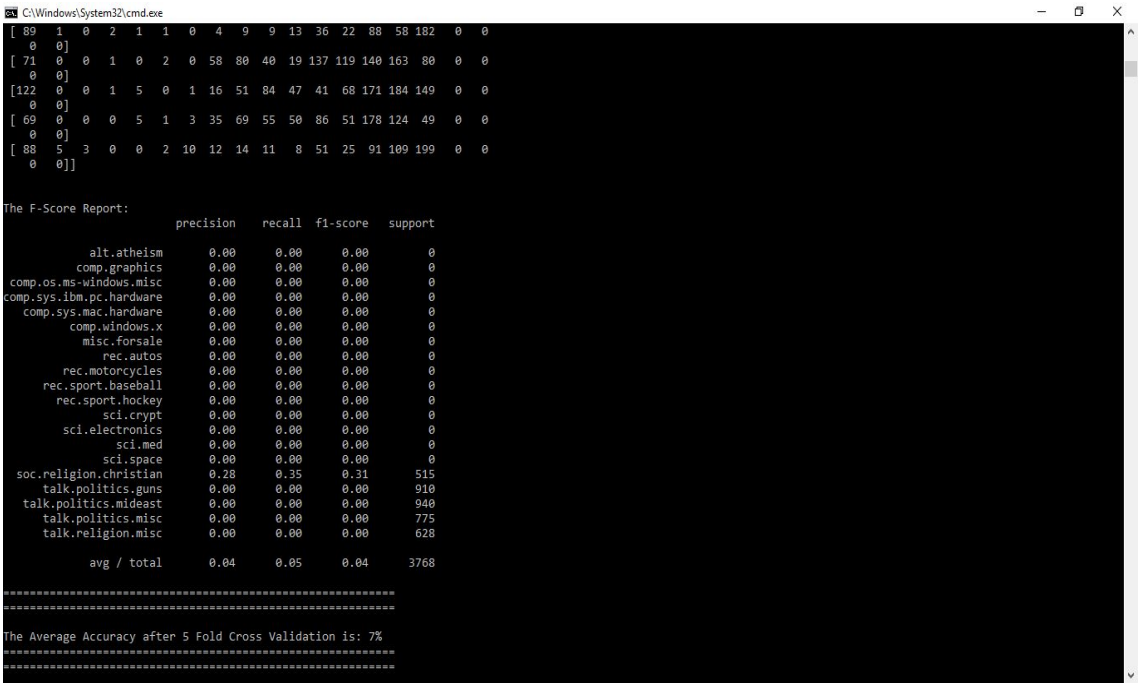

Figure 7: KNN Average Accuracy and f-score(With Feature selection)

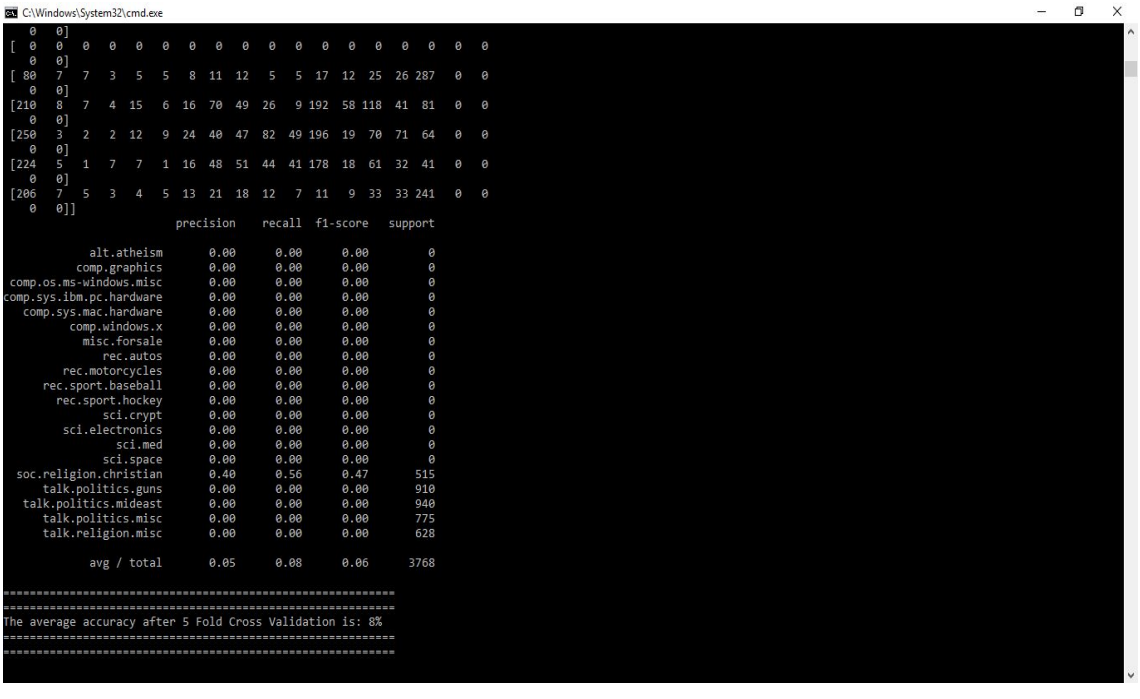

Figure 8: Decision tree Average Accuracy and f-score (With Feature selection)

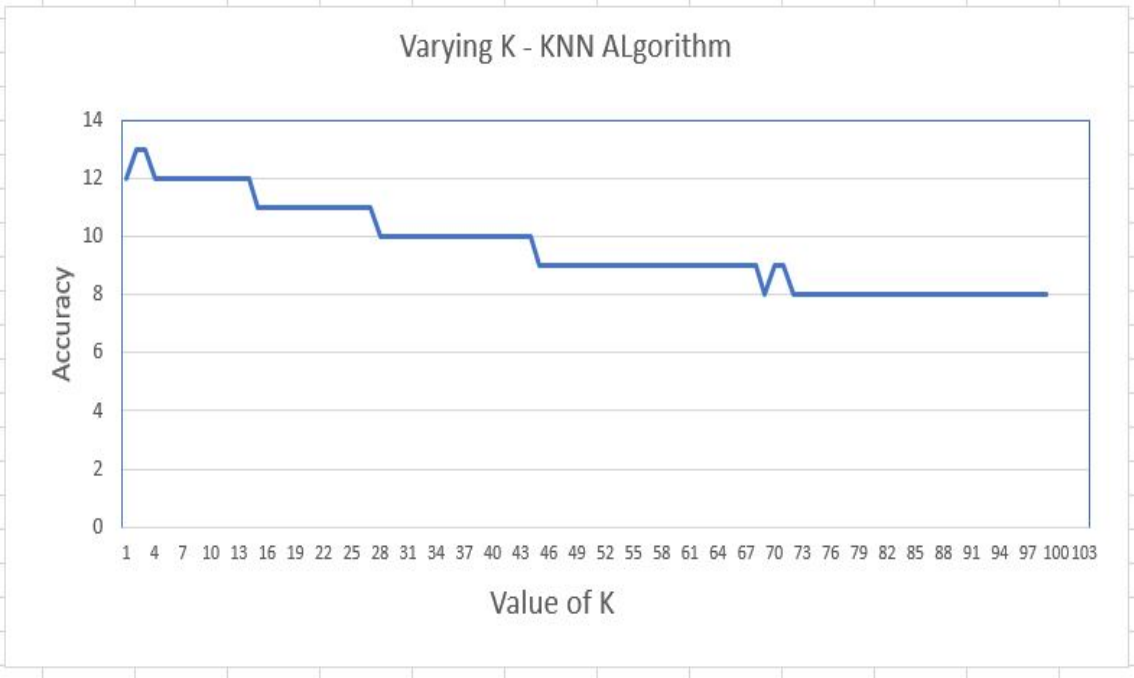

Figure 9: Varying K -KNN Algorithm(With Feature selection)

constant at 8% and the best being K=2. But according to theory there has to be increase in the accuracy, my assumption to answer is that as I selected the DTM after feature selection due to data loss there is some adverse effect on the algorithm.

5 Randomized Sampling

As I saw above that the algorithms produced very low accuracies. I preformed Knn algorithm on the main DTM by making the system assign random training and testing sets, this was done by passing 70% of samples for training and 30% for testing. Accuracy of around 78% was obtained.

Another observation that can be noted here is that the while training these models it is better to select the documents across different categories then there is a chance of getting better accuracy.

NOTE: *Preprocessing and this part of the code is done using R. Please check "r project\knn_hw2\hw2.r" for reference code.*

6 Conclusion

From the above results there are many conclusions that can be drawn depending on the methods used to perform them,

- Cross validation ensures better accuracy and gives better predictions.
- Performing Average K rather than setting K to constant value i.e performing prediction for a range of values of K and choosing the best and then using that.
- Knn Algorithm works better when the whole DTM is available when compared to Decision tree.
- Decision Tree Algorithm works better after Feature selection when compared the KNN.
- Randomized sampling must be done for better results.

Lastly, both algorithms have their merits and demerits it actually depends on the type dataset and the approach a researcher might use.