Information Retrieval

ASSIGNMENT 4

Prabhat Kumar

MT17036

DATASET

Dataset consists of 5000 text documents, divided into 5 classes – 'comp.graphics','rec.sport,hockey','sci.med','sci.space','talk.politics.misc'.

ASSUMPTIONS

- Extraction of features is done by TF-IDF values where IDF has been calculated for Training data collection
- Rocchio's Centroid has been calculated on TF-IDF Feature Vector

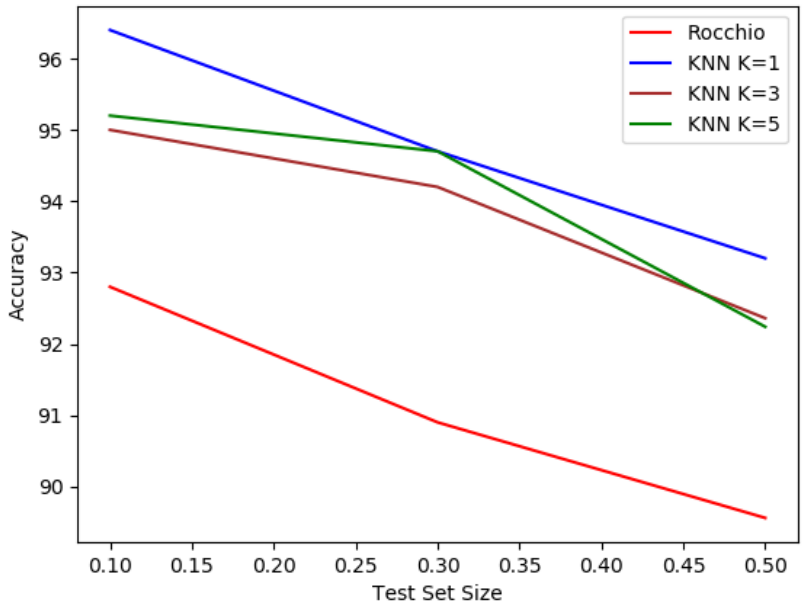
METHODOLOGY

- Preprocessing of Data
- Training of Rocchio's Algorithm
- Training of KNN for K=1,3,5
- Analysis of Results

PREPROCESSING

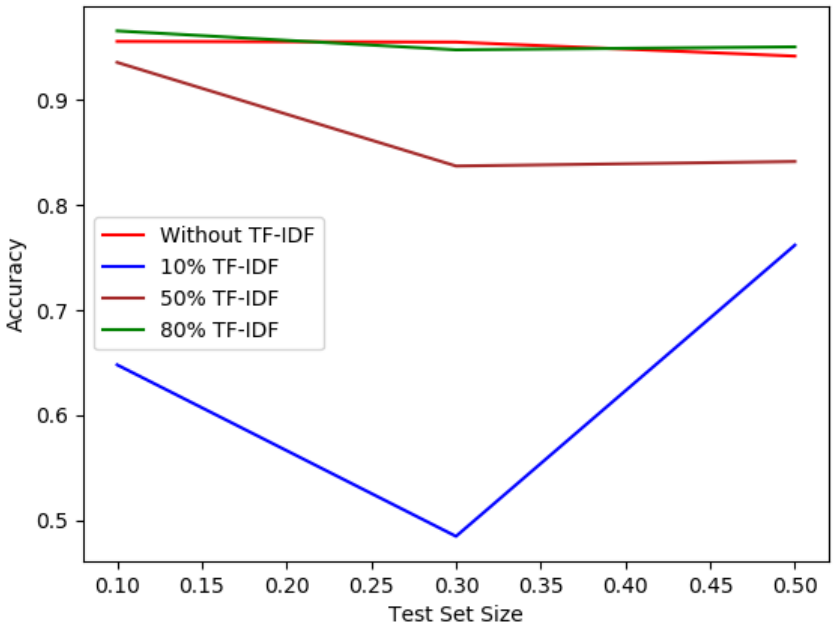
- As data was initially grouped into classes by folders, each data point consists of path to the text file and ground truth or label.
- Headers of each file is removed.
- Punctuation Removal was performed followed by tokenization.
- Each token is then Stemmed using Porter Stemmer.

OBSERVATIONS AND INFERENCES



Rocchio's Algorithm and KNN

Accuracy on a Scale of 0-100



Naïve Bayes

Accuracy on a Scale of 0-1

- Accuracy achieved using Rocchio's Algorithm hovers around 90% for all test train splits.
- Similar results observed for Naïve Bayes for either no feature selection or by complete feature selection using TF-IDF as a metric.
- Accuracy achieved by KNN increases as Training Size increases or in other words decreases as Test Size increases. K value of KNN does not have a significant effect on results achieved by KNN, where the reason for such result cannot be generalized for the dataset provided.

CONFUSION MATRIX PLOTS HAVE BEEN STORED IN PLOT FOLDER IN ROOT DIRECTORY

REFERENCES

- NLTK
- Introduction to Information Retrieval (English, Christopher D. Manning)