Information Retrieval

ASSIGNMENT 3

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 and Features of each classes have been calculated based upon highest TF-IDF combined in each class.
- Feature Vector has been defined by combining the best defining each class.

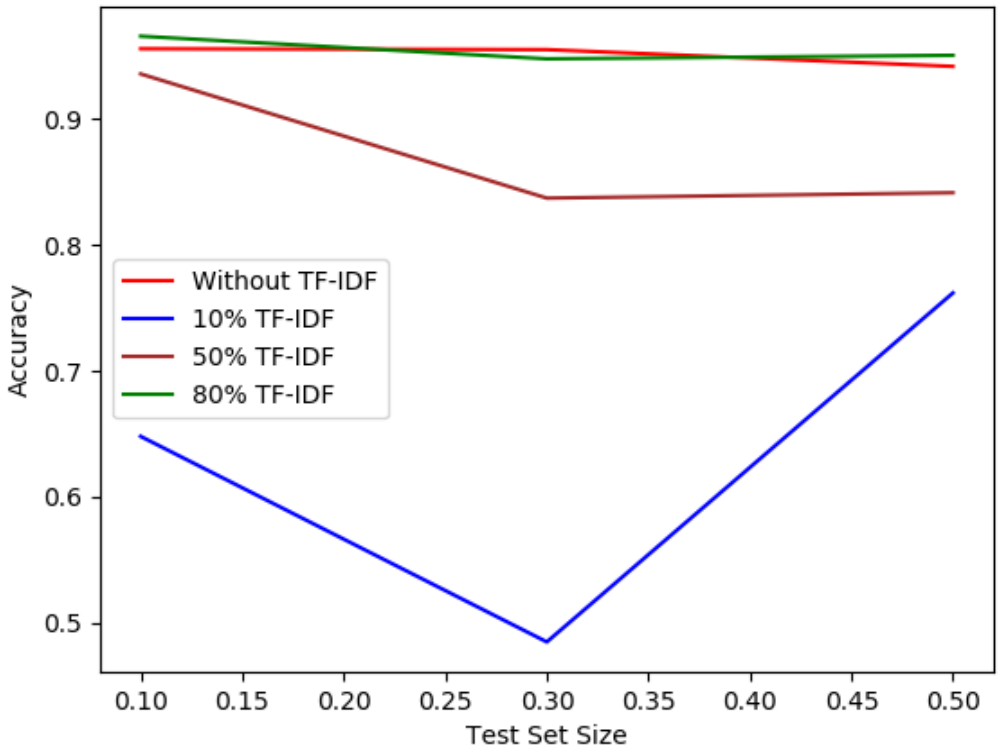
METHODOLOGY

- Preprocessing of Data
- Feature Extraction/ Selection
- Training of Naïve Bayes Classifier
- 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



- It can be observed that the TF-IDF Extraction near to 100% is equivalent to one with any type of feature extraction.
- It can also be seen that the TF-IDF based unigram extraction is not a effective feature extraction technique, for the given dataset.

CONFUSION MATRIX OF VARIOUS CASES CAN BE FOUND IN THE PLOTS FOLDER

REFERENCES

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