CSE508: Information Retrieval
Assignment 4

Instructions

- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism, institute policy will befollowed
- You need to submit ReadMe, code files and analysis.pdf

Naive Bayes Algorithm You need to implement Naive Bayes Algorithm on your own for the following question. Library usage is not allowed apart from data pre-processing steps.
Download 20_newsgroup dataset from
https://drive.google.com/file/d/1VA4a-wveTVXEy0J_NNv8oZ_YG2smxvPL/view
You need to pick documents of comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space [5 classes] for text classification.

Q1) Perform the data pre-processing steps - Split your dataset randomly into 70:30 train:test ratio for each class - Train your Naive Bayes Classifier on the training data - Test your classifier on testing data and report the confusion matrix and overall accuracy - Perform the above steps on 50:50, 80:20 and 90:10 training and testing split and analyze the accuracy scores.

Q2) Implement Tf-idf scoring technique for efficient feature selection and perform NB classification over that for 70:30. Compare its result with Q1 results.