CSE508 : Information Retrieval
Assignment 3

Deadline : 20th March'18, 2359 hrs

Total: 100 marks + 10 marks bonus

Instructions

- Assignment is to be attempted individually. Please keep the discussions on an abstract level
- Language allowed : Python
- For Plagiarism, institute policy will be followed
- You need to submit ReadMe, code files and analysis.pdf
- You folder should be renamed in the NameRollNo_HW3 format before zipping

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.

- 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 analyse the accuracy scores

Bonus: Implement Tf-idf scoring technique for efficient feature selection