CSE508 : Information Retrieval

Assignment 4

Deadline : 6th April'18, 2359 hrs

Total: 100 marks

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

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.

You need to implement

1) Rocchio Classification Algorithm
2) KNN classification (vary k=1,3,5)

Perform the above steps on 50:50, 80:20 and 90:10 training and testing split and analyse the accuracy scores

Compare and Analyse these two methods with previously implemented Naive Bayes Algorithm.