CSE508 : Information Retrieval
Assignment 3

Instructions

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

Dataset: Use the 20newsgroups dataset

Question-1

Use **Tf-Idf based vector space document retrieval** to get top 10 documents based on a cosine similarity between query and document vector.
Further, provide a client for a user to give Relevance Feedback (tell which all docs are relevant and which are irrelevant). Using Rocchio algorithm, optimize the query. Keep doing this process of taking user feedback and updating the query vector until the user quits the program.
In the report, show how the query vector changes after applying the Rocchio algorithm. Show a 2D TSNE plot of the vectors to demonstrate the difference. You can use SKlearn's inbuilt functions to make this plot.
For this question, use only **comp.graphics** and **rec.motorcycles** documents.

Question-2

Use the data file provided here. This has been taken from Microsoft learning to rank dataset, which can be found here. Read about the dataset, and what all it contains.
Assume a model that simply ranks URLs on the basis of the value of feature 75 (sum of TFIDF on the whole document) i.e. the higher the value, the more relevant the URL.
Assuming any non zero relevance judgment value to be relevant, plot a Precision-Recall curve for query "qid:4".