CSE508: Information Retrieval
Assignment 2

Team:
Yukti Goswami (MT21109)
Saurabh Pandey(MT21077)

Q1
DATA - Humor, Hist, Media, Food

PREPROCESSING

Performed the same readme file as the question 2 of IR Assignment-1.

METHODOLOGY

- Calculated Union and Intersection of document and query to find the Jaccard coefficient.
- Calculated the Document Frequency using all 5 techniques
- Found TF-IDF using all five techniques
- Designed matrix for each TF-IDF
- Found top k ranked documents using the TF-IDF matrices

Q2
DATA - Microsoft learning to Rank Dataset

METHODOLOGY

- Selected data for qid=4
- Found number of files possible by calculating accumulating the factorial the length of each unique value doc
- Calculated DCG and NDCG
- Plotted the Precision-Recall curve

Q3
DATA - 20_newsgroup dataset

PREPROCESSING

- Lowercase
- Removed stopwords
- Lemmatization

METHODOLOGY

- Implemented method to randomly split data in any ratio required
- Calculated class frequencies
- Applied Gaussian Naive Bayes model
- Calculated Confusion matrix for different train test split ratios.