Readme

Question 1 :

Data set Used: 20newsgroup

The python file consist of code to create Inverted List of terms for the corpus mentioned above. For each term in the inverted list documents are sorted on the basis of static quality score. Along with this each term also consist of High and Low lists of documents for high and low values of Term frequency. This code file also present analysis on the selection on the size of the High list (denoted by 'r').

Pre-Processing Steps :

- All the articles were broken down into tokens using **RegexpTokenizer**.
- All the punctuations were removed.
- **Porter steamer** was used to perform stemming.
- All the stop words were left as it is.

Description of Functions used:

1) Def TARextract() : This function is used to extract TAR file of the dataset
2) Def Static_Dict_Create() : This function is used to create the dictionary of the static quality score of the various documents present in the corpus.
3) Def Inverted_Index_Create() : This function is used to create the inverted index for the terms in the dictionary for the entire corpus. The posting list is sorted by the static quality score.
4) Def Analysis_r() : This function is used to perform the analysis on the value of the size of the High List.
5) Def Net_Score() : This function is used to generate net score for a query document pair.

Question 2 :

Data set Used: https://drive.google.com/file/d/1aG_sOmDqN2cIx0ChUdxfGjdSVZAt7LGA/view

This python code performs analysis on the evaluation parameters like maxDCG , nDCG at various levels for **Microsoft's Learning to Rank Dataset.** This code also generate precision recall curve for the URLs on the basis of feature number 75.

Functions Used :

1) Def Data_Generator() : This function is used to generate data in the form of Lists of List from the given data set.
2) Def DCG_MAX() : This function is used to find the value of maximum DCG for QID:4 and finding all the permutations which will give maximum DCG value.
3) Def nDCG(k) : This function is used to find nDCG at 50 and for entire dataset.
4) Def P_R() : This function draws precision recall curve.