Readme(Assignment 3)

(MT19111) Akshaj Patil

1:PreProcessing Steps:

- Case folding (all terms are converted to lower case).
- Stop words removal.
- Removal of punctuations.
- Numbers are converted to Words.
- Lemmatization.

Same pre-processing is done for data and query.

2:Assumptions:

- Meta data of the documents is considered as part of document.
- Query should be of at least 1 word.

3:Methodology:

Q1:]Here we preprocessed all the files. We first constructed inverted index of all the terms with docId as a key and term frequency of that doc as a value. Then we sort all docID w.r.t term-Frequencies. Then we divided inverted index into two lists 1:]High list and 2:]Low list. First 'r' docs will in high-list and remaining docs will be in low-list. Then sort high list and low list on the basis of their respective g(d) score which is provided in extra file.txt.

When User runs the code. It will ask user to enter query. When users enter query query will be preprocessed and will be converted into list of lemmatized words. Then user have to enter number of docs he want to retrieve. After that we will apply fast cosine algorithm to calculate score of each document. First we will traverse high-list of each term and calculate score, then check if number of docs retrived are greater than 'k' if yes then just display first 'k' docs if no then traverse low-list of every term and calculate score of docs. After that we will retrieve top 'k' docs whose score is high.

(Here g(d) score is normalized by max g(d) score.)

Q2:] Here we will first retrieve all the URL with qid:4. 0^{th} col is relevance score and and 75^{th} col(75:..) is sum of tf-idf. To calculate IDCG we have to sort all URL w.r.t 0^{th} col and then calculate DCG.

DCG(p)=Summation of[(pow(2,rel[i])-1)/math.log(i+1,2)] where I ranges from 1 to p.

I used this formula because when relevance score is not binary then above formula gives “stronger emphasis on relevant doc”.

NDCG(50)=DCG(50)/IDCG(50).

NDCG(overall)=DCG(overall)/IDCG(overall).

3-> Then we sort url on basis of 75^{th} value. And consider 0^{th} col for calculating precision and recall. If 0^{th} col value is non-zero then doc is relevant and if 0^{th} col value is zero then doc is not relevant. By this we calculated precision and recall at every point. And then plot Precision vs Recall curve.