IR Assignment 3

Shalini Bhardwaj(MT19045)

Ques1: Static Quality Score

1.Preprocessing

Dataset:20Newsgroup

Preprocessing steps: Nltk library is used for preprocessing
1. To Lowercase: query is transformed to lower case
2. Punctuation is removed.
3. Stop words are removed
4. Tokenization : converted to tokens
5. Lemmatization : converting to root dictionary word

Both query and dataset are preprocessed.

Files from all folders are kept in one place and then read.

2.Methodology

1.A dictionary is built by extracting gdscore from file.Its keys are doc and value is its gd score.

2.All the above preprocessing steps are performed initially.

3.After lemmetization document frequency is calculated for each term of document.Then two dictionary of inverted index and gdscore with respect to each term is built.

4.Idf is calvulated for each term and stored in dictionary.

5.Inverted index and gd lsit dictionary are sorted by tf and gd score respectively.

6.High and low list are built from sorted inverted index dictionary.
r is chosen by taking mean of length of posting list of all terms.It comes out to be nearly 20.Therefore r=20 is number of docs to be kept in high list.Rest docs are kept in low list.

7.At this moment high list is sorted by tf values.Now those docs in high list and lowlist are further sorted by gd score from gd list.

8.Now query and number of docs to retieved is taken as input from user and preprocessing steps are done on it.
9.Now for each query term tf idf is calculated and stored in a separate dictionary.For this we used fast cosine.
10.If we get number of docs(k) from high list tf score only ,we don't go to lowlist.Otherwise we go to lowlist also.Then final retrieval of docs is done with max score.

11. we have used g(d) score normalized by max g(d) score

Assumption:

Metadata is considered in documents.

Q2 NDCG

1.Preprocessing

1.All the urls with quid 4 are extracted from dataset in a file named newQ2.txt
2.each url is converted to list of list .
3. column 0 and column 76 are kept separately as col 76 col contain sum of tfidf.

2.Methodology

1.Original url are stored with variable res and new sorted file is stored as X.

2. For max dcg value sorted urls are calculated with formula:
Sum of (2^relscore-1)/log(i+1) where i is position of document.
As this formula gives more emphasis to relscore when scoreis not binary .Hence this is used.
3.This sorted file is stored in a file named output file.

4.Permutations of files with max dcg is calculated using counting files where url with rel score 0 are not present +0 is present once+0 is present twice …..so on till 0 is present all 59 times. All its permutations are calculated which comes out to be 5.407e+121

5.Similarly ndcg is calculated by dcg/max dcg for first 50 urls then for whole data given

6.For ndcg at 50 dcg is calculated at given sequence but for max dcg it is sorted.

7. rel score are sorted based on tfidf value in new extracted list it is named as globa.

8.Precision and recall are calculated for quid=4 and stored in a list as each document is retrieved. Consider 0th col for calculating precision and recall

9.Precision recall graph is constructed.

Assumption:

Text file containing output has list of list of sorted rel score docs.