Akanksha Dewabngan
MT19049

README

Que1:

Methodology:

In this first we unzip the data file of 20_newsgroup.zip.Then perform the preprocessing steps as mention in next section.After that I created a term dictionary with posting list contains document ids as keys in posting list which contain tf values of that particular word in mention document.And then I created a list of unique words which are keys to my dictionary.After that i uploaded that gd score file and then normalise it values between o and 1.As i created a term posting list dictionary so i make a pickle file to use it again and again because code will take time to run.after that i appended the gd score to the term dictionary.After that sorted dictionary on the basis of tf. Then heuristic for choosing r is mention further.Then split the term dictionary in high and low list on the basis of chosen r and create new dictionary for that. After that ort the dictionary of low or high list in the basis of gd score.And also calculate idf.Then cosine function is created in which document to query similarity is taken out and then add gd score to get top K required document by user first scan the high list and if returned document id less then K then scan for low list and do the same as high list for it and return the required K document to the user.

R :mean apoprach
Mean is a approch taken for r.
Average length of posting list will be there on high list and reduce chance of going in low list.

Mean is 21

,the numbers of access to low list will be reduced.

Preprocessing:

In preprocessing step first convert the paragraph in upper to lower then remove punctuation then remove stopwords, then tokenize it and apply lemmatizer then apply stemming too,after that number to word conversion

Assumptions:

- Normalise the gd score of file.txt file by min-max normalization formula so that the gd score will be between o and 1.

```
gd_new=(data -min_key) / (max_key - min_key)
```

- Use myfile.pkl which contain the term dictionary.

Que 2:

(A)

Methodology:

In this first file is uploaded and then perform little preprocessing over data, then make data frame of quid 4 urls and then sort first column of relevance of decreasing order and then find out Maximum DCG of that points .There are perm(),permuation() for caluculating permuation of all possible points. the n calculate IDCG for the same.

Preprocessing:

Split and put in table format after splitting on the basis of next line and space and remove the spaces if it will come in the list and blank list.

In this only datas of qid 4 wer taken out

(B)

Methodology:

In this first taken out urs of qid 4 then taken out top 50 for part I taken out IDCG and DCG for same to calculate nDCG and fir whole data calculate IDCG and DCG for that and calculate NDCG for the same.

(C)

Methodology:

In this precision recall curve are plot in which first sort the qid4 urls of o column and sorted on the basis of col 75 which are tf and then find out Precison and recall of 103 points and on all 103 points plot it on the graph.

Que3:

PDF file is uploaded for this question.