CSE508 : Information Retrieval

Assignment 3

Max Marks: 80

Instructions

- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism, institute policy will be followed
- You need to submit README.pdf, Code files and Analysis.pdf
- You are allowed to use libraries such as NLTK for data preprocessing.
- Mention methodology, preprocessing steps and assumptions you may have in README.pdf.
- Mention your outputs, analysis done (if any) in Analysis.pdf
- Submit code, readme and analysis files in ZIP format with the following name: **A3_<roll_no>.zip**
- The deadline for this assignment won't be extended.

Q1. Consider the 20newsgroup dataset for this question. Download this file containing the number of favourable reviews for each of the documents in the corpus. The first column of this file denotes the doc no. and the second column denotes the corresponding number of favourable reviews.The doc no. 1-1000 belong to alt.athiesm, the next 1000(i.e. 1001-2000) belong to comp.graphics and so on. **(35 Marks)**

You need to implement static quality ordering on the above corpus. You have to extract a static quality score for each document from the number of favourable reviews as provided above. Maintain a global champion list for each term sorted by a common ordering (i.e. static quality score). This champion list for a term ‘t’ is made up of two posting lists consisting of a disjointed set of documents. The first list is the high list containing documents with highest tf values for ‘t’ and the second list is the low list containing other documents having term ‘t’. You need to define a suitable heuristic to select the minimal efficient value of ‘r’ where ‘r’ is the number of documents to be contained in the high list.

You need to return K documents (where K will be given at runtime). During query processing, only the high lists of query terms are scanned first to compute net scores and return as output. If less than K documents are returned then go for low lists. Report proper analysis for all the values of ‘r’ that you tried and the reason behind choosing that value of ‘r’.

Q2. Use the data file provided here. This has been taken from Microsoft learning to rank dataset, which can be found here. Read about the dataset carefully, and what all it contains.

1. Make a file having URLs in order of max DCG. State how many such files could be made.

2. Compute nDCG
    a. At 50
    b. For the whole dataset

For the above two questions, results have to be produced for qid:4 only and consider the relevance judgement labels as the relevance score.

3. Assume a model that simply ranks URLs on the basis of the value of feature 75 (sum of TF-IDF on the whole document) i.e. the higher the value, the more relevant the URL. Assume any non zero relevance judgment value to be relevant. Plot a Precision-Recall curve for query “qid:4”. **(30 marks)**

Q3.

1. Explain the relationship between ROC curve and PR curve.
2. Prove that a curve dominates in ROC space if and only if it dominates in PR space.
3. It is incorrect to interpolate between points in PR space. When and why does this happen? How will you tackle this problem?

You can take reference from the following paper :
https://www.biostat.wisc.edu/~page/rocpr.pdf
Include your answers in the analysis file. **(15 marks)**