LAB 5

Based on lecture 11

This lab is depends on the collection from Lab 2 and Lab 3, and the queries of Lab 3. The lab focuses on applying Psuedo Relevance Feedback (PRF) as an example of query expansion methods for IR.

TFIDF

- Please be sure that all the steps in lab 3 already done, and that you have index you built for the collection (trec) from lab 2 ready.
- Based on the ranked results you retrieved in Lab 3, please build a module that does the following:
  - Read the results file and extracts the numbers of top *n_d* ranked documents.
  - Read the preprocessed document from the collection (if you don't have the collection saved after preprocessing, then read the document from the collection then apply preprocessing to it).
  - Append the content of all the *n_d* documents together.
  - For every term in the appeneded documents, calcualte the tfidf score. Use the formula *tf.log(N/df)*.
  - Sort the terms by tfidf score. Report the top *n_t* terms.
- You need to apply this for each of the queries in queries.
- For simplicity, start with *n_d* = 1, and *n_t* = 5, which means using only the top ranked document, and find the top 5 terms having the highest tfidf score within it.
- Create a file called Qm.1.5.txt in the following format:

```
1 incom tax reduc + tax spend labour conserv rate
2 peac middl east + baker peac israel soviet arab
```

  where the terms after the "+" are the *m* terms you extracted as expansion for each query. All terms are supposed to be preprocessed (since your inverted index is preprocessed). However, if you haven't applied preprocessed, it is still OK. The objective hear is to see how PRF can suggest useful terms for expansion.
- Repeat (as much as you can) for different values of *n_d* and *n_t*. Try *n* = {1,3,5}, and *m* = {1,5,10}. Each time the file to print will be named: Qm.*n_d*.*n_t*.txt
- You are encouraged to discuss the output you have with your colleagues and to compare the results lists you get for each query. Feel free to post it on Piazza.
- Of course you might be interested to run the new queries to search the collection and observe how the results will change.

TIPS

- In case you haven't finished this before, you can find the list of ranked documents for Lab 3 queries in the following results file.
  Note: this contains the top 10 results only per query, which what you will only need for this lab.
  Also, if for some reason you don't have the index built for this collection, you can just build a quick script to count the df of terms in the collection, so you can use when calculating the tfidf.
- The output for the first two queries when *n_d* = 1 and *n_t* = 5 is displayed in the examples above. If you get the same set of 5 words for these two queries, it should indicate that you are moving in the right direction.
  Note: if you are using different preprocessing, it might be a bit different.

Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, Scotland, UK
Tel: +44 131 651 5661, Fax: +44 131 651 1426, E-mail: school-office@inf.ed.ac.uk
Please contact our webadmin with any comments or corrections. **Logging and Cookies**