THE UNIVERSITY of EDINBURGH
School of informatics

University Homepage
School Homepage
School Contacts
School Search

LAB 3

Based on lecture 7

This lab is a continuation to previous labs (1 and 2), and it should help in completing you assignment.

TFIDF

- Please be sure that all the steps in lab2 to be already done, and have the collection (trec) from lab 2 ready.
- Build a new module within your code to run ranked retrieval, and the similarity measure to be used is:
  - TFIDF: use formula in lecture 7, slide 13 with title "TFIDF term weighting".
  - TFIDF with different SMART notation (optional)
- Run the following queries, and report the list of retrieved documents for each query. Create to output file:
  - tfidf.results: contains the retrieval results using tfidf.
- The query file is formated where each query is given on a separate line. The first field is the query number, followed by the query itself, as follows:

  ```
  1 income tax reduction
  2 peace in the Middle East
  ```

  There are 10 queries in the file.
- Results files format should be as follows:

  ```
  1,710,0.6234
  1,213,0.3678
  2,103,0.9761
  ```

  This means that for query "1" you retrieved two documents with numbers "710" and "213". Document "710" is a better match to the query than "213" (similarity 0.6234 vs. 0.3678). For query "2" you retrieved one document ("103"). Print results for queries in order. All results for query 1 is sorted by score, then results for query 2, 3, ... 10.
- You are encouraged to discuss the output you have with your colleagues and to compare the results lists you get for each query
- Notes:
  - It is expected to have preprocessing applied to queries and documents.
  - This is ranked retrieval not Boolean search, so you should retrieve all documents that contain at least of the terms of the query.
  - It might worth trying running the search twice: with and without stopping (the default should be with using stopping).
  - Number of results per query would vary. Some queries will have only few relevant documents, others can have 10's (100's if stopping is not applied). So please list all results you get in the tfidf.results format.
  - It is **highly encouraged** to discuss your results with friends and even share on Piazza.

NOTE

Once you complete lab2 and lab3, then your CW1 should be almost done. All what you need to do later is to run your system on the new collection and queries to be released, then write the report.