CS657A: Information Retrieval

Assignment 1 (120 marks)

Due on: 28th February, 2022, 11:00pm

The documents are hosted as `english-corpora.zip` available from https://www.cse.iitk.ac.in/users/arnabb/ir/english/. Each document is a plain text in English. The filename of the document is its `DocId`.

1. (10 marks) Perform tokenization and stemming. You may use whitespace tokenization and stemming algorithms. You are free to use or experiment with other methods. You can also use standard libraries. Please clearly specify if you have used such methods. You may need to perform some text cleaning, such as removal of non-ascii characters, etc. You may or may not remove stopwords.

2. Implement various IR systems using the processed documents. (You cannot use library functions.) Ensure that the implementations for all of these are efficient, since this is an IR system. (Each query should finish within 3 minutes.)
   (a) (15 marks) Implement a simple Boolean retrieval system.
   (b) (20 marks) Implement a system from the Tf-Idf family with appropriate forms for the functions and tuned parameters. A query is matched using cosine similarity.
   (c) (20 marks) Implement a system from the BM25 family with appropriate forms for the functions and tuned parameters.

3. (20 marks) Submit a set of 20 queries in the *QRels* format. In other words, for each query, mention the document ids that are relevant to it. Mention at least 10 relevant documents in ranked order. The QRels format is
   ⟨`QueryId, Iteration, DocId, Relevance`⟩
   where `QueryId` is the query id, `Iteration` is akin to a version number and can be set to `1` by default, `DocId` is the document id (for this assignment, this is the filename) and `Relevance` is a binary digit where `1` denotes that it is *relevant* and `0` otherwise.

   You need to return a *ranked* list with the top-10 relevant documents.

   Queries can be a set of words or a sentence.

4. (20 marks) Your systems will be evaluated using random 40 queries. For each query, a set of top-5 documents will be retrieved. The *mean average precision (mAP)* over all the queries will be the score.

   All the three systems will be evaluated, and the best score will be considered. So,

   $$\text{score} = \max\{\text{mAP(Boolean)}, \text{mAP(Tf-Idf)}, \text{mAP(BM25)}\}$$

5. (15 marks) The submission MUST contain a README file and a Makefile. The program should take as input a file that contains a set of queries, each in one single line. The first field of a query line is the query id (it will run from `Q01` to `Q20`). The second field is a free text. The fields are separated by a TAB character. The code must have documentation with appropriate comments.

Instructions

Submit the assignment as one zip file `rollno-assignment1.zip` in the course portal (hello.iitk.ac.in) within the deadline.
Submit your set of queries and QRels (Q3) as `rollno-qrels.zip`.
Submit the code and running system (Q4) as `rollno-ir-systems.zip`.
The evaluation will be done *automatically*. So, if your code fails to compile or run properly, you will get ZERO for that question.
The programs MUST run in the Linux operating system.