Implementing a Prototype High Performance Web Search Engine

Objectives

1. Gain experience in building indexes for a Web-based search engine
2. Gain experience in implementing a prototype Web search engine and its ranking algorithm(s)

What to do?

Web search engines can provide us all kinds of information. It is a crucial issue for a Web search engine to provide accurate results for a given query. In this assignment, you are going to build an index for a set of given Web documents and return retrieved documents.

You will be given a collection of Web documents (compressed file with the size of 73.1M). This assignment consists of the following three parts.

1. Implementing an indexer that can create an index for the provided Web documents.
2. Searching these Web documents from the created index for 20 topics/queries.
3. Outputting your search result in a file. You should list topic Ids, document Ids, and the documents' ranking

Your ranking program should be able to rank the documents retrieved by your search program in the decreasing order of document's relevance value. The submitted ranking results should be formatted as follows:

401 Q0 WT24-B28-147 1 6.714665567764736 GroupID
401 Q0 WT24-B20-169 2 6.710866977409565 GroupID
401 Q0 WT23-B33-327 3 6.706060261387986 GroupID
401 Q0 WT06-B30-165 4 6.623599349833492 GroupID
401 Q0 WT08-B09-455 5 6.535713254328925 GroupID
402 Q0 WT18-B23-796 1 5.543729143867218 GroupID
402 Q0 WT08-B03-317 2 5.543563405653674 GroupID
402 Q0 WT22-B03-93 3 5.543294410357613 GroupID
402 Q0 WT13-B36-796 4 5.542368432072281 GroupID
402 Q0 WT16-B26-381 5 5.541382885371797 GroupID

where the first column is the topic number; the second column is always Q0; the third column is the retrieval document ID; the fourth column is the rank of the document retrieved; the fifth column shows the weighting score (integer or floating point) that generated the ranking, and the sixth column is your group ID such as "**g1**" stands for the group 1. Your submitted results will be evaluated by the standard TREC evaluation program.

What to submit?

You should submit the following items:

1. The programs for implementing a Web search engine and building an index from Web documents.
2. The assignment report that describes how to build your index, the method you use, and the design of your programs.
3. The programs for preprocessing the raw Web pages.
4. A readme.txt where you give a tutorial on how to compile and run your programs.
5. The files related to your index.
6. The file containing retrieved results for 20 topics.

How will you be graded?

The following will play a crucial role in your grade for this assignment.

1. Correctness of your programs for building the index and for doing the search.
2. Your assignment report that should include introduction, description of your implementation, analysis of the results.
3. Your retrieval performance in terms of document MAP and the number of relevant documents retrieved.
4. Clarity of your programs (comments!).
5. Ease of using Web-based interface for searching.
6. Time and space complexity of your programs.
7. Ease of using the README to test your programs and results.
8. Your collaborations with your team members.

The full mark for this assignment is twenty. Your programs, assignment report and supporting documents account for 15 marks. The group-peer marking from your team members accounts for 5 marks.