Information Retrieval : Implementation Project

Subject :

The goal of the project is for you to understand the fundamentals of search engines by creating a search/retrieval system for the furnished document base :
Minimum requisites for your system are:

- Should work on the furnished documents and at least on the queries corpus (you can also obtain from several sources online).
- Should parse and index documents using inverted file
- Should make use of stemming and stop lists (however, you can existing tools for this part).
- You should implement matching based on one of the methods discussed in class.
- Form of queries is up to you from the simplest (i.e. simple terms queries) to more sophisticated ones (i.e. wildcar queries, phrasal queries, or tolerant retrieval) .
- Should propose an interface to permit users to friendly query your system

- May add extensions to your system such as :
  - Enlarging the set of queries that could be asked to it,
  - Propose ranked results
  - Introduce similarity measures between queries and documents
  - …

Deliverable

Report :
Submit a comprehensive final report (not more than 20 pages) that discusses the search engine tool and how it works. The indexing, the query process, and, should the cas arise, how the search engine calculates relevance should be discussed in detail.

Presentation :
You will give a presentation for 15 minutes on your search project

- Present an introductory presentation on your search engine topic
- Your conceptual choices : type of the choosen model(s), implementation, types of queries etc.
- A live demonstration of the search engine should be performed.

Some links that may help you :

http://tartarus.org/~martin/PorterStemmer/ (Stemmer)
http://www.web-mining.fr/methodes/stop-words (Stopwords lists)
http://www.algolist.net/Data_structures/Dictionary_%28ADT%29 (Dictionary data structures)
http://www.gutenberg.org/catalog/ (corpora)
http://www.umiacs.umd.edu/~jimmylin/downloads/brill-javadoc/edu/mit/csail/brill/BrillTagger.html (Brill's tagger)