CSCI 4130 Information Retrieval

Project 3

1. Develop a ranked IR system based on the tf-idf weighting.
2. Use the Cranfield corpus for the system development.
3. Evaluate the effectiveness of the ranked system by MAP measurement.

About the Cranfield collection:

1. Was a pioneering test collection in allowing precise quantitative measures of information retrieval effectiveness.
2. It has the following components:
    a. Cran.all: a collection of 1400 documents
    b. Cran.qry: the queries posed on the corpus
    c. Cranqrel: the relevance assessments(the level of relevance of a document to a query)
    d. Readme: explanation about relevance judgements
3. The following is the structure of the Cranfield corpus documents:

```
.I 1
.T
experimental investigation of the aerodynamics of a
wing in a slipstream .
.A
brenckman,m.
.B
j. ae. scs. 25, 1958, 324.
.W
experimental investigation of the aerodynamics of a
wing in a slipstream .
  an experimental study of a wing in a propeller slipstream was
made in order to determine the spanwise distribution of the lift
increase due to slipstream at different angles of attack of the wing
and at different free stream to slipstream velocity ratios .  the
results were intended in part as an evaluation basis for different
```

theoretical treatments of this problem .

the comparative span loading curves, together with supporting evidence, showed that a substantial part of the lift increment produced by the slipstream was due to a /destalling/ or boundary-layer-control effect . the integrated remaining lift increment, after subtracting this destalling lift, was found to agree well with a potential flow theory .

an empirical evaluation of the destalling effects was made for the specific configuration of the experiment .

The lines that follow:

.I 1: the document/information identifier, which is 1

.T: comprise the document title
.A: denote authors of the document
.B: denote the journal in which the document was published
.W: denote the abstract of the document/journal paper

4. High-level solution steps
    a. Extract each single document and save in a single file
    b. For the whole collection, maintain a list of all the terms, together with their document frequency
    c. For each document, maintain a list of terms used in that document, and term frequency for each term. Then calculate tf-idf for each term in the documents.
    d. Use the provided queries to test your system, return top 10 documents for each query (collect results from running at least 20 queries)
    e. Check the ranked results provided in cranqrel to calculate the MAP of your system
    f. You should develop a simple user interface(command-line based interface is just fine) to prompt the user for a query/information need.

5. What to submit:
    a. Java program,
    b. Instructions for compiling and running your program
    c. You test report on executing some queries
    d. Only one submission per team is required

6. Bonus: (20 points)
   Write the program to test the queries and calculate MAP automatically.

7. This is an involved project. It is essential that all team members actively contribute to the project. Please start early and spend some time on algorithms and data structure choices.