CSCI 4030 CSCI/ DASC 6030: Information Extraction and Retrieval

Programming Assignment: 03

12 February 2022

1 Assignment Goal

Implement the Vector Space Information Retrieval model. Demonstrate your solution on the Cranfield corpus.

2 Cranfield corpus

The Cranfield corpus is available on MS Teams. You may also download the Cranfield corpus from `http://ir.dcs.gla.ac.uk/resources/test_collections/cran/`. It has the following components:

1. cran.all - A collection of 1400 documents
2. cran.qry - The queries posed on the corpus
3. cranqrel - The relevance assessments (the level of relevance of a document to a query)
4. readme - Explanation about relevance judgements

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 .
```

- .I 1 is the document/information identifier, which is 1.
- The lines that follow:
  - .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.

3 Solution Steps

Following are high-level solution steps. You may need to make several decisions in each step related to low-level implementation details. Think about alternatives, articulate their pros and cons, reason about their algorithmic correctness and efficiency, and make informed decisions. It is strongly encouraged that hold one or two brainstorming sessions with your team members to strategize a solution before you delve into code-level details.

This is an involved project. It is essential that all team members actively contribute to the project. Get started early. Spend time upfront on algorithm and data structure choices.

1. Build two **zonal** indexes – one for the title and another for the abstract. Use the $tf \times idf$ weighting to construct weighted vectors.

2. Develop a simple user interface (command-line based interface is just fine) to prompt the user for a query/information need. The user should use queries from the Cranfield collection. To make query specification simpler, the system may prompt the user to enter a 3-digit integer which corresponds to a query/information need (see the **cran.qry** file). Allow the user to specify **boost values** to indicate the relative weights for the title and abstract. For example, the boost values pair $(0.7, 03)$ indicates that a weight of 0.7 be assigned to the title match and 0.3 for the abstract match.

3. Compute the similarity of the query with the corpus documents. You may use a brute-force approach or use heuristics to fetch only the top k documents.

4. Show the titles of ranked documents that are relevant to the query. Optionally, you may also show the abstract of the top k matches.

5. Design test cases and execute them. Document test case execution results. Assess how well the retrieval performance of your system compares with the expert-provided answers (see the **cranqrel** file).

4 Submission

Submit your **source code files**, instructions for compiling and running your program, and results from execution of your test cases. Upload all documents to Canvas. Only one submission per team is required.

5 Seeking help

Please post your questions to the "Programming Assignments" channel on MS Teams.