CSCI 4030 CSCI/ DASC 6030: Information Extraction and Retrieval

Programming Assignment: 03 Documentation

14 March 2022

Abelson Abueg

This program implements the Vector Space Information Retrieval model for searching relevant documents, demonstrated on the Cranfield corpus.

The program will read through the the cran.all.1400 corpus and ask the user for a query ID from the cran.qry file. The program will then calculate the Cosine Similarity Scores for each document in relation to the query to then create a list of top k results for relevant documents, where k is the number of results that the user wishes to preview.

How to Compile:

```
javac -O -cp ".\opennlp-tools-1.9.1.jar" .\VectorSpaceModelIR.java
```

How to Run and their Parameters:

```
java -cp ".\opennlp-tools-1.9.1.jar" VectorSpaceModelIR .\cranfield-corpus\cran.all.1400 .\cranfield-corpus\cran.qry
```

Program Assessment:

The program runs really well performance-wise (runtime will be printed in the terminal) and does a decent job at scoring the documents. But from the cranqrel relevance assessment file, the program does a good job to get highly relevant documents but it also retrieves irrelevant as being highly relevant. I felt like more could be done to it beyond cosine similarity scoring to improve the search results, but I think the program is good enough for this assignment, especially with the performance of the code being very efficient.

Generic Test Cases:

- If argument length != 2. you will get a usage message on how to correctly run the file.
- If either or both arguments do not lead to the required file, you will get an error message saying that it is not a filepath to either or both files. Otherwise the program will run.
- If the program is running, any invalid inputs will give an invalid input message and the program will let the user try again. Otherwise, the program will run smoothly.