Research Project 2:

Roll Your Own Mini Search Engine (30)

In this project, you are supposed to create your own mini search engine which can handle inquiries over "The Complete Works of William Shakespeare" (http://shakespeare.mit.edu/).

You may download the functions for handling stop words and stemming from the Internet, as long as you add the source in your reference list.

Your tasks are:

(1) Run a word count over the Shakespeare set and try to identify the stop words (also called the *noisy* words) – How and where do you draw the line between "interesting" and "noisy" words?

(2) Create your inverted index over the Shakespeare set with word stemming. The stop words identified in part (1) must not be included.

(3) Write a query program on top of your inverted file index, which will accept a user-specified word (or phrase) and return the IDs of the documents that contain that word.

(4) Run tests to show how the thresholds on query may affect the results.

Grading Policy:

The report of this assignment is due Sunday, March 26th, 2017 at 10:00pm.

- **Programming:** Write the programs for word counting **(1 pt.)**, index generation **(5 pts.)** and query processing **(3 pts.) with sufficient comments**.

- **Testing:** Design tests for the correctness of the inverted index **(2 pts.)** and thresholding for query **(2 pts.)**. Write analysis and comments **(3 pts.)**. ***Bonus: What if you have 500 000 files and 400 000 000 distinct words? Will your program still work?*** *(+2 pts.)*

- **Documentation:** Chapter 1 **(1 pt.)**, Chapter 2 **(2 pts.)**, and finally a complete report **(1 point for overall style of documentation)**.

The presentation **(10 pts.)** of this assignment is due Tuesday, March 28th, 2017 at 09:50am. All the contributors must be present at the classroom before 09:35am to have the computer ready and the speaker decided.

Peer review of the reports is due Thursday, March 30th, 2017 at 10:00pm.

Final grading sheets will be uploaded after the arbitration.