CS533 Information retrieval

Project part 3

Due: December 5, 2017, 11:59PM on Blackboard

This program will perform the retrieval. First, it will read the dictionary, postings file and the document table from disk and store them in memory.

Then it will start to perform the query processing. It should be able to read from the screen many queries, and will terminate only after you print the command EXIT. For each query, it will compute and write on the screen a list of retrieved documents. The query and the results will be written also to an output file.

For simplicity the query q can be:

An AND query that starts with the word AND followed by a list of one or more words. For example (AND action war). The AND query can also contain AND NOT followed by one or more words. For example: AND mystery AND NOT horror. The AND must appear before the AND NOT.

An OR query that starts with the word OR followed by a list of two or more words. For example (OR action romance).

Change all query words to lower case before performing retrieval. Remove stop words.

The one word query (AND thriller) should retrieve all documents that have the word thriller in the review. For example the document 0002.HTML should be retrieved since the word thriller is contained in its text.

The two word query (AND Jeff Goldblum) should retrieve all movie reviews that mention Jeff and Goldblum. 0003.HTML should be retrieved since it contains the name of the actor. Note that even though the name appears in parenthesis in the document it should still be retrieved.

The queries (AND beyond therapy love story) should retrieve all documents with the all these words. Document 0003.HTML should be retrieved.

The query (OR mystery thriller action) should retrieve all documents that contain any of these three words in their text. Both 0002.HTML and 0004.HTML will be retrieved.

The output of the program will be a file called *output.txt*. For each query:

Line 1 will contain the query.

If there are no results for the query the program will print NO RESULTs in the next line. Otherwise it will print the results sorted first by rate and then by file name. (All movie reviews with a positive rate will appear first.)

Each result will contain the filename, title, reviewer, rate and the snippet.

Submission directions

Same as for program 1.

Plagiarism Policy

All your code will be checked for similarity to other submissions using Moss. Programmers have an uncanny knack of reproducing the same code that they have seen before. So you are advised not to look at each other's code.