MSCI 541 - HW3
Due: noon, Friday, Mar 23, 2012, in the 541 drop box.

Data

Same data as HW1 and HW2. Plus these two files:
http://mansci.uwaterloo.ca/~msmucker/teaching/541/topics.401-450.txt

http://mansci.uwaterloo.ca/~msmucker/teaching/541/LA-only.trec8-401.450.rel.txt

Tasks

1. For the 50 search topics in the topics.401-450.txt file, extract the topics' titles and treat each title as the query that a user would enter into a search engine. (You may do this by hand. No need to write a program.) For example, the query for topic 401 should be:

   foreign minorities, Germany
2. Implement either BM25 or the language modeling approach to retrieval. If you select language modeling, you should select the smoothing method of your choice.
3. Perform and time retrieval for each of the 50 topics and retrieve the 1000 top ranked documents for each topic. Construct a single results file for the given 50 topics. You result file must have the following format:

   topicID 0 docno rank score runTag

   where the above fields are separated by a single space, and the rows are sorted in ascending order by the topicID (primary key) and the score (secondary key). The columns are:

   topicID: an integer value in the range [401,450]
   0: the number zero, ignored but required
   docno: the document's docno, no extra whitespace
   rank: the rank of the document from 1 to min(1000, the number you retrieve if it is less than 1000). Rank 1 is the best document.
   score: a numeric value such that if score_i > score_k, then docno_i should be ranked higher (nearer the top of the list, i.e. a smaller rank) than docno_k. **Note: Evaluation practice is that rank is ignored and only score is used to determine the order of the ranking.**
   runTag: a unique string of less than 12 alphanumeric characters that uniquely identifies your run. You should select a runtag that is your nexus username.

   In summary, only 4 columns are used: topicID, docno, score, and runTag, but you must include all 6.
4. Spot check your results by computing the average precision at rank 10 given the LA-only.trec8-401-450.rel.txt file. This file is known as a qrels file. In this file, there are 4 whitespace separated columns:

   topicID ignore docno judgment

   The columns are:

topicID: an integer value in the range [401,450]
ignore: a column to ignore.
docno: the document's docno
judgment: the NIST assessor judgment. If judgment is 0, the document was judged non-relevant to the search topic. If the judgment is > 0, then the document was judged to be relevant to the search topic.

You should be finding some relevant documents in the top 10 for some of the topics. Topic 436 is an easy topic (Railway accidents), for example. Some other topics only have 1 relevant in the LATimes sub-collection. If you want, you may use trec_eval to compute retrieval metrics for you: http://trec.nist.gov/trec_eval/ but it is only designed to be easy to compile and use on unix systems.

What to Turn In

1. Print all source code.
2. Write a short report (approximately 2-3 pages) detailing:
   a. Introduction. Brief introduction about the assignment.
   b. Methods. Required: retrieval method employed and all of its parameter settings. Did you implement document-at-a-time or term-at-a-time retrieval? Please explain any other important details needed to produce your results. For example, you should touch again on tokenization.
   c. Results. Report your average precision at 10 results and the average time per topic to produce results.
   d. A discussion of your results. Where does your retrieval method work well and where does it seem to fail? Some form of failure analysis would be good.
   e. Any implementation issues encountered as part of completing this project. Was the implementation easy, hard, or just right?
   f. Conclusion. Brief conclusion about the assignment.

Marking

I will mark your report based on the quality of the writing, analysis, and correctness of your results. Please describe your methods clearly.

Use of Existing Code Etc.

You may program in your language of choice. You must work alone on this assignment except that you may openly discuss design choices and the homework with anyone. You may use any code that you are legally allowed to use for non-commercial purposes excluding code:
1. written by classmates, and
2. complete retrieval suites of software such as Lucene, Lemur, Indri, Terrier, etc.
and you **must carefully acknowledge all code that you have not written yourself.** **You may not use code written by any other 541/720 student.** Please see the course outline for general academic honesty guidelines.