MSCI 541 - HW1
Due: noon, Feb 3, 2012, in the 541 drop box.

Data

We have uploaded the LA Times portion of the TREC volumes 4 and 5 collection to Learn/D2L under "Content". You must have signed a TREC agreement to have access to this data.

The data is gzip'd. Also uploaded is a program (CompressedInput.txt) for another course that shows you can read gzip'd files from your programs without ever uncompressing them on disk. The web contains lots of information on how to do this in your favorite language.

The data format is as described in class and as the uploaded READMEs also explain.

Tasks

Pick a document of reasonable length from the collection of LA Times documents. You document choice should be such that you feel confident that you understand what the document is about.

For this document:

1. Create a table of the top 10-20 words in the document given their TF in the document. Show what their TF is.
2. Create a table of the top 10-20 words in the document given their TFIDF scores. Show their TF and TFIDF. Your IDF must be based on the whole LA Times collection.

What to Turn In

1. Print all source code.
2. Print out your selected document.
3. Write a short report (approximately 2 pages) detailing:
    a. Methods. How did you tokenize the document? What choices did you make for tokenization? How did you compute TF and IDF?
    b. Results. Please use tables as appropriate.
    c. A discussion of your results. Do they make sense? Etc. In other words, interpret your results for the reader and tell the reader what you think is interesting.
    d. Any implementation issues encountered as part of completing this project. Was the implementation easy, hard, or just right?

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 written by classmates, but 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.