CS574 Language Technologies
Fall 2002
Assignment 1

Due at the beginning of class, on Monday, September 30

Goals for the assignment

1. Start becoming familiar with one research-oriented IR system — The Lemur Toolkit.
2. Gain experience with IR evaluation methods (using Lemur) and approaches for presenting evaluation results.
3. Gain a deeper knowledge of term-weighting strategies for ad-hoc retrieval.

The assignment is fairly open-ended. (From your perspective, this should be a good thing since there will be a lot of flexibility in how you interpret each question and in the methods that you use to address each question.) The assignment will require that you delve a bit into the documentation for Lemur, understand and modify its basic shell scripts, possibly look at the texts in the CACM collection, interpret results, etc. The material covered in class will provide the necessary general background, but it won't provide "the answers."

Caveat: The on-line version of this assignment is the official one. Please check it occasionally for additional notes, clarifications, etc.

Part 1: Setup

1. Download and install the Lemur Toolkit, which is available at: http://www.cs.cmu.edu/ lemur.

   Instructions for installing and running Lemur are available via the DOCUMENTATION link.

   Be sure to complete steps 1–4 of the installation/running instructions. Note, however, that:

   - We will only make use of the basic indexing code (not the position indexing code).
   - In step 4 ("Testing the Toolkit on Sample Data"), note that the link that is supposed to provide sample output with which to compare your own is not working. Instead, compare your output to the sample output from version 1.0 of Lemur: http://www.cs.cmu.edu/ lemur/1.1/sample-output/

   Steps 5 ("Using the API") and 6 ("Modifying the Toolkit") shouldn't be necessary for this assignment.

2. **Collaboration:** Collaboration is encouraged for just this part of the homework. Please do try to help each other out to get Lemur up and running on your platform of choice.

3. **What to turn in:** Send e-mail to cardie@cs.cornell.edu when you have completed this part of the assignment. The e-mail should indicate the platform under which you installed Lemur as well as briefly describe any major problems that you ran into with the installation. In addition, list the names of any classmates that you worked with for this part of the assignment.

Part 2: Evaluation in IR

1. Spend some time understanding what information is being presented in the output files produced by the test_basic_index shell script.

2. Among the system variations tested via test_basic_index is a "simple TFIDF" run and a "simple Okapi" run. Based on the outputs of those runs, construct an argument as to which system is better. Your argument should be supported quantitatively using whatever evaluation metrics, tables, and graphs you think are necessary. Some additional questions you might consider in your analysis of the results are the following: Is one evaluation metric enough to decide which system is best? Is one system always better than the other? If not, under what circumstances is one system better than the other? Are there specialized retrieval tasks for which the reverse might be true? Do you need to look more closely at the queries and documents in the CACM collection to address any of the above issues?

3. **What to turn in:** For this part of the assignment, turn in the type-written evaluation analysis.

Part 3: Term Weighting

1. Experiment with changing the parameters for the "simple TFIDF" system and write-up the results of your evaluation (as above). Your evaluation need only compare the "simple" version of the system to **one** other version of the system. Before running the evaluation and doing the analysis of results, be sure to think about and write down any hypotheses regarding (a) what you expect to see happen to the results based on the specific parameter change and (b) why. The write-up should include both the hypotheses and the analysis of results.

   Perusing the Lemur documentation for specific information on the TFIDF parameter settings should be helpful here.

2. **What to turn in:** For this part of the assignment, turn in the type-written evaluation analysis (including the initial hypotheses).