Web Search Engines: Sample Final Exam

The take-home exam will be distributed in hard-copy and placed on the web at the end of the final class period on December 13. It is due at 3:00 PM, Thursday December 16. No exams can be accepted after that time.

The exam is open-book, open-notes, open-Web. You can consult with any resource you choose except another human being.

The exam may be submitted either in hard copy or eletronically. If you are submitting in hard copy, then slip it under the door of my office (WWH 429). For electronic submissions, I prefer plain text or Postscript, but will accept PDF or Word.

I will discuss the answers to the sample exam questions on Dec. 6. DO NOT SUBMIT ANSWERS TO THE SAMPLE EXAM.

Part I

(10 points each)

Problem 1:

A. How can the owner of a web site design a “spider trap”?

B. Is it possible for a crawler to distinguish between a spider trap and a legitimate web site with 100% certainty? Explain.

C. Describe one method that a crawler might use to avoid spider traps. What are the drawbacks of this method?

Problem 2: In the vector model of documents,

A. What are the dimensions of the vector space? What is the meaning of the coordinate of document D in dimension J?

B. In the TF-IDF model, how is the value of the coordinate of document D in coordinate J computed. (There are a number of different formulas that can be used; you may give any one of these.)

C. What is the reason for including the IDF factor, instead of using just TF?

D. Actual Web search engines that use a vector model presumably use a modified version of the formula in B. Name three considerations that might enter into the value of the coordinates of document D in coordinate J. (Note: These must be considerations that are dealt with at indexing time not at query time, and that involve both the word and the document, not just the word or just the document.

Problem 3: Describe the operations involved at query time in computing a “relevance score” for a document as an answer to a multi-word simple query (no quoted strings, no Boolean operators, no special features) in Google. Try to include in your answer all the consideration that Google is
known to use, or may be observed to use. For each of these, describe what data is used, when the
data is computed, and what data structure it is stored in.

**Problem 4:** Pages in the Web graph is divided into the categories SCC, IN, OUT, TENDRILS,
and DCC.

A. How are these categories related to the WCC?

B. Suppose that a crawler starts with a seed set consisting of a single page in the SCC. Describe the
categories of pages that the crawler can reach. (The form of your answer should be something
like “All of the pages in DCC and IN and some of the pages in TENDRILS”.)

C. Repeat part (B) for a seed set consisting of a single page in IN.

D. Repeat part (B) for a seed set consisting of a single page in DCC.

**Problem 5:** In computing PageRank, one solution to the problem of pages with no outlinks is to
posit that every page is considered to have an implicit self-link (that is, a link to itself).

A. Using that assumption, set up the system of linear equations that defines the PageRank for
the graph shown below. (You need not compute the solution.) Assume that the parameter
$E$, as defined in class project 2, (the probability of following an outlink rather than jumping
randomly) is equal to 0.6

$$
\begin{align*}
V & \rightarrow X \\
U & \rightarrow T \\
W & \rightarrow Y \\
T & \rightarrow V \\
X & \rightarrow W \\
Y & \rightarrow W
\end{align*}
$$

B. What does this approach (considering each node to have a self-link) not give good results?

**Problem 6:** I have never seen it in print, but I think that it is quite safe to assume that Google
keeps a log of the queries submitted; that the frequency of specific queries follows an inverse power-
law (Zipf) distribution; and that Google caches the answers to the most frequent queries in some
fast memory, and these are answered without need to refer to the main inverted index.

A. There is a feature of the Zipf distribution that implies that it will be reasonably easy to achieve
a fairly large cache hit rate. Explain.

B. There is a different feature of the Zipf distribution that implies that it will be very difficult to
achieve cache hit rates that are close to 100%. Explain.
C. How could knowing the shape of this distribution be useful in tuning the structure of the cache? I don’t mean ”knowing which queries are most frequent” which you obviously need; I mean knowing a fact like, “The Nth most frequent query has a frequency proportional to $1/N^{1.7}$.”

Part II (40 points)

Problem 7: (Open-ended. Your answer should be no more than about 1000 words.) Suppose that you wanted to automate the following task using web resources: Given a subject Q, and a starting date and an ending date, construct a chronology of the major events involving Q between the starting and ending date.

A. Propose a method for carrying out this task. Your method may use any existing online resources you choose. If you want, there can be a human in the loop, but, of course, some significant part of the task must be done by the computer.

B. Propose a learning method that would enable part (A) to become more effective over time.

C. How would you evaluate your program? What measure of quality would you use? How could you operationalize this measure? (You may use either human or automated evaluators.)