This 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 electronically. 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.

Part I
10 points each

Problem 1:

A. In using the vector model to compare the similarity of two documents, why is it desirable to normalize the vectors to unit length?

B. Let \( \vec{D} \) and \( \vec{E} \) be the normalized vectors corresponding to documents \( D \) and \( E \). What is the minimum possible value of \( \vec{D} \cdot \vec{E} \)? If \( \vec{D} \cdot \vec{E} \) is equal to this minimum, what can you say about \( D \) and \( E \)?

C. What is the maximum possible value of \( \vec{D} \cdot \vec{E} \)? If \( \vec{D} \cdot \vec{E} \) is equal to this maximum, what can you say about \( D \) and \( E \)?

Problem 2:

A. In Google, the inverted index is divided into separate parallel databases handled by separate servers. How is the index divided?

B. Describe the high-level steps in using the parallel database to answer queries.

C. Besides the parallelism involved in (A), there are two other types of parallelism involved in the query engine (not including parallelism in the crawler). Describe them briefly.

Problem 3: Pages in the Web graph is divided into the categories SCC, IN, OUT, TENDRILS, and DCC. Suppose that page \( P \) is in OUT, page \( Q \) is in IN, and that page \( W \) is in TENDRILS. Suppose that you now add a link from \( P \) to \( Q \) and a link from \( P \) to \( W \). For each of the following statements, state whether it is definitely true, possibly true, or definitely false:

A. Page \( P \) is now in IN.

B. Page \( P \) is now in SCC.

C. Page \( P \) remains in OUT.
D. Page Q remains in IN.
E. Page Q is now in SCC.
F. Page Q is now in TENDRILS.
G. Page W is now in OUT.
G. Page W remains in TENDRILS.
H. Some pages other than P, Q, or W have been moved into SCC.
I. The diameter of SCC has increased.

**Problem 4:** In the standard PageRank algorithm, every page starts with the same “inherent” values, and then acquires more value through its inlinks.

A. It has been suggested that a link from page P to page Q should be considered as more “significant” if P and Q are from different domains than if they are from the same domain. Give two arguments in favor of this.

B. In Brin and Page’s original PageRank paper, they suggest that one could define alternative notions of PageRank by assigning different “inherent” values to different pages. At an extreme, one could assign all the inherent value to a single starting page START, and have all other pages derive their values via chains of inlinks from START. In terms of the stochastic model, this would correspond to a modified version where, if you flip “heads”, you follow a random outlink; if you flip “tails”, you always jump to START, rather than jumping at random in the web. (They tried the experiment with START = John McCarthy’s home page.)

Now, suppose that you combine this model with the link-weighting scheme in part (A), interpreting “more significant” as “twice as significant”. Using this new model, set up the system of linear equations for PageRank in the graph shown below. Assume that the probability of flipping “heads”, $E=0.6$. 

![Graph](image-url)
**Problem 5:** Consider an experiment that is doing statistical studies over a sample of 1 million web pages. For any page $P$, let $I(P)$ be the number of inlinks to $P$. Let $P_N$ be the page with the $N$th largest value of $I(P)$ (ties broken arbitrarily). Suppose that $I(P)$ follows the inverse power-law distribution, $I(P) = \lceil 200,000,000/(N+20) \rceil^2$, where $\lceil X \rceil$ is the largest integer less than or equal to $X$. Thus, the page $P_1$ with the greatest number of inlinks has $\lceil 200,000,000/21 \rceil^2 = 334,479$ inlinks.

A. How many pages have no inlinks? What is the median value of $I(P)$?

B. How many links are there in total? What is the average number of links per page?

C. For any link $L$, let $S(L)$ (the “siblings” of $L$) be the set of links whose head is the same as $L$. Consider $L$ itself to be an element of $S(L)$. For instance, in the graph in problem 3. if $L$ is the link from $A:U \rightarrow C:W$ then $S(L)$ is the set $\{ A:U \rightarrow C:W, A:X \rightarrow C:W, B:T \rightarrow C:W, B:V \rightarrow C:W \}$. What is the average size of $S(L)$, averaged over all links in the experimental sample?

D. Explain, in general terms, how the use of link-based evaluation strategies in popular search engines such as Google tends to exacerbate the very uneven distribution of inlinks among web pages.

**Problem 6:** It is important that when a crawler downloads a page, it can quickly check whether it has seen the content before.

A. How can this check be implemented efficiently?

B. It is reasonably frequent that two pages $P$ and $Q$ differ only in their HTML tags and white space. Describe how the method in (A) can be modified to check whether two pages are identical in this sense. Describe an application in which $P$ and $Q$ can be treated as identical (that is, if it has dealt with $P$, it can ignore $Q$.) Describe an application in which $P$ and $Q$ should not be treated as identical.

C. Suppose that crawler for a general purpose search engine has downloaded URL $P$ and has discovered that its content is exactly identical, including HTML tags and white space, to URL $Q$, which has already been processed. What does the crawler now do with $P$?
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 general task using web resources:
Given a subject Q, return all university courses that are likely to teach Q.

A. Propose a method for carrying out this task. Your method may use any existing online
resources you choose.

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.)

The following considerations might be significant:

- The topic Q might be in the title, a course description in an online catalogue, or course web
  pages such as a syllabus or class notes.

- The course may well teach Q if it mentions a super-category. For instance, a course on “Russian
  Music” is likely to teach Tschaikowsky, even if no details beyond the course name can be found.

- It would be useful if the courses were listed in decreasing order of the centrality of Q or the
  likelihood that the course will teach Q. For instance, a course that has “Tschaikowsky” in the
  title should rank high; one that mentions “Tschaikowsky” as one of many subjects should rank
  in the middle; a course on “Russian music” that does not mention Tschaikowsky should rank
  lowest.

I am by no means saying that you have to address all these issues in your answer; still less, that
you may not prefer to focus on other issues. I’m just pointing out some directions in which such a
project could go.