Test 2

75 minutes. Two pages of handwritten notes are allowed.

1. [16 points] Metasearch
   a. [10] Diagram a simple metasearch system (either the one that you built or one that you could easily build).
   b. [4] If you had a 100GB disk and wanted to speed up your system by caching, what would you cache?
   c. [2] What if you had 10 terabytes?
2. [3] What is the “rank merging problem” (aka the “rank aggregation problem”). Why is it hard?

3. [3] Is it better to crawl depth-first or breadth-first? Why?

4. [1] Some of the papers we read referred to the “relevance” of results. Is this the same thing as precision? as recall? neither?

5. [6] It has been suggested that the least frequent terms of the query should be processed first, as they carry the most meaning and probably contribute most to the final ranking.
   a. Does this make sense mathematically? If so, why? If not, is there a better measure?
   b. Does this make sense in the context of an implementation like Google’s? Why or why not?
6. [16] Suppose you have been appointed quality assurance manager for a search engine specialized for finding jokes. Briefly describe the testsets and testing procedures for evaluate the quality of: a. the crawler, b. the data structures, c. the ranking algorithm, and d. the user interface.
7. [2] In Y!Q why are the context terms not simply added to the query?

8. [2] In Google the inverted index is divided into barrels handled by separate servers. How is the index divided? (Ernest Davis)


10. [11] It is important that when a crawler downloads a page, it can quickly check whether it has seen the content before. (Ernest Davis)
    a. [3] How can this check be implemented efficiently?
    b. [2] It common for two pages P and Q to differ only in their HTML tags and white space. Describe how your method in a. can be modified to check whether two pages are identical in this sense.
    c. [2] Describe an application in which such P and Q should not be treated as identical.
    d. [4] 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?
11. When, if ever, does it make sense to use a page quality metric (e.g. PageRank) during the crawl? Why?

12. The queries “platform independent” and “independent platform” should retrieve different sets of pages. There are various ways to implement such “phrase search”.
   a. Recall that Google uses data structure called a “hitlist”: for term $t$ in document $d$, this lists all the places where $t$ occurs in $d$. Give pseudocode for implementing phrase search using hitlists.
   
   b. An alternative way to implement phrase search is to treat phrases like “platform independent” as terms, adding them to the lexicon. Give two advantages or disadvantages of this approach compared to the method you gave in part a.