

Final Examination

165 minutes. Three pages of handwritten notes are allowed.

1. [5 pts] When choosing the name for a new company, one thing to consider is how easy it will be to keep the company's homepage at the top of the search results for a query on the company name. Which of the following would be best on this criterion, and why?
 - El Paso Marketing
 - Hernandez Marketing
 - HrndzMktg
 - Shark Bop Marketing
 - Shark Bop
2. [1] In the vector model of documents, what are the dimensions of the vector space? (Ernest Davis)
3. [3] Give three reasons why a search engine might present a page with low inferred quality (e.g. low PageRank) above one with higher quality.
4. [2] One could store a search engine's inverted index in a database. Explain why this would or would not be a good idea.

5. [17 pts] Suppose you are tasked to create a metasearch engine for scanning high school student essays and flagging likely instances of plagiarism from Web documents.
- a. [2] Sketch or describe the user interface.
 - b. [5] Assuming that essays are typically 2000 words but that a query can be at most 20 words, explain how the system can formulate queries useful for finding likely source documents.
 - c. [10] Sketch or describe how you would build such a system, explaining at least the hardware and software requirements and the behavior of any modules you would need to write.

8. [12] Suppose that you want to automate the following task using web resources: given a subject Q, return all UTEP courses that are likely to teach Q. Note that Q may be mentioned in a course title, in a course description in the catalog, or on a course web page or syllabus. Assume that metasearch is not an option. (Ernest Davis)
- a. [10] Explain how you would build a system to support this task.
 - b. [2] Explain how you would evaluate this system.

9. [4] Give two reasons why audio search is much harder than text search.
10. [4] For a search engine based solely on the vector space model, would you expect the query “new new” to give the same results as the query “new”? Why or why not?
11. Imagine that you are augmenting a simple corporate intranet search engine to give a preference for pages which have been recently updated. Suppose that “recency” is modeled as $4 - \log_{10} t$, where t is the time in days since the last update.
- [2] Give an equation showing one plausible way to combine this recency metric with vector space similarity to give an overall score suitable for generating rankings.
 - [4] Suppose your co-worker proposed a rival equation. Explain the procedure you would follow to determine which to use.

12. For each of the following operations:

- a. [8] Specify whether it is done pre-query or at query time or both.
 - a. applying a stoplist
 - b. inferring the user's intent or information need
 - c. query rewriting
 - d. spam removal
 - e. term weighting
 - f. compression of documents
 - g. generating snippets (aka summaries) to include in the search results
 - h. computing match using the vector space model
 - i. computing PageRank
 - j. clustering
 - k. indexing
 - l. retrieving documents from the web
 - m. stemming
 - n. formatting results
 - o. spelling correction
 - p. determining which advertisements to display
- b. [4] Identify 4 of these where personalization could best be applied.
- c. [2] For one of these explain briefly how personalization can be done.

13. [4] Some search engines include on the initial results page an estimate of the total number of results, e.g. "1-10 of about 22". These estimates are rough, and can be off by 50% or more. Explain how such estimates can be efficiently obtained.

14. Baeza-Yates and Ribeiro-Neto suggest that a way to find synonyms is to find pairs of words which appear together in many documents.
- a. [8] Give pseudocode to do this. Assume that the inverted indices are available.
 - b. [2] State the computational complexity of your algorithm.
 - c. [2] Mention one way this could be speeded up.

15. Another way to find synonyms is to look for words which tend to appear in similar local contexts. For example, if a corpus includes the phrases “complaints of barking, nuisances on sidewalks, and even a recent mauling suggest that our town has a canine problem” and “dogs a public nuisance and calling for stricter enforcement of the leash law and zero tolerance for nighttime barking”, we can infer, from the shared words “barking” and “nuisance”, that “dog” and “canine” are synonyms (or near synonyms).

We can formalize this by defining the “neighbor document” for a word to be the concatenation of all phrases containing that word in the entire corpus. The problem of finding synonyms for a word w then reduces to the problem of finding words whose neighbor documents are similar to the neighbor document for w .

- a. [2] Express this idea with an equation, showing the summation explicitly.
- b. [2] Explain how this is like the standard information retrieval problem, of finding documents similar to a query.
- c. [2] Explain how this differs from the standard information retrieval problem.
- d. [6] Explain how you could use a standard IR package to do this, noting any needed modifications to the algorithms or data structures.