COS 435   Spring 2009
Summary for 2nd Exam

All material *through* April 16 will be covered on exam out Tuesday, April 28 and
due Thursday, April 30, 2009.

Emphasis will be on materials not covered on first exam

Major topics (highlights) up to midterm exam

- Models
    - Modeling queries and documents
        - Satisfying, scoring and ranking w.r.t query
    - Vector models
        - LSI
    - Link analysis for ranking
- Evaluating retrieval systems
- Algorithms and data structures
    - Index structures
    - Evaluating queries using indexes
    - Building indexes

Major topics (highlights) since midterm exam

- Indexing algorithms, continued
    - Compressing index structures
    - Distributed computing for index construction and query evaluation
- Using user characteristics and feedback
    - Search refinement
    - Recommendation methods
- Other analysis of information objects
    - Clustering
    - Detecting near-duplicate documents
- Semi-structured information
    - XML
- Crawling the Web
    - Characteristics of the changing Web

A summary of reading assigned to date in *Introduction to Information Retrieval. Schedule and Assignments* Web page is the definitive source of reading assignments. Bold indicates assigned since March 3.

Chapter 1 – all sections
Chapter 2 – all sections
Chapter 3 – section 3.1: structures for dictionaries
      Rest on wildcard queries and spelling correction
Chapter 4 – all sections
Chapter 5 – all sections
Chapter 6 - all sections except:
      6.1 zone scoring
      6.4.4 advanced dealing with document length in scoring
Chapter 7 – Section 7.1, except 7.1.6, and 7.2
Chapter 8 – all sections
Chapter 9 – all sections
Chapter 10 – all sections
Chapter 16: all sections except:
      16.3 Evaluation of clustering
      16.5 Model-based clustering
Chapter 17: all sections except
      17.5 Optimality of HAC
      17.7 Cluster labeling
Chapter 18 – all sections
Chapter 19 –all sections except:
      19.5 estimating relative sizes of search engine indexes
      (includes 19.6 finding near-duplicate pages)
Chapter 20 all sections except:
      20.4 Connectivity servers
Chapter 21 - all sections