Tenth week summary
April 14, 2008

All material ***including*** today will be covered on exam out Monday, April 21 and due Wednesday, April 23, 2008.

Emphasis will be on materials not covered on first exam.

1

Major topics (highlights) at midterm review March 12

- 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

2

Major topics (highlights) since midterm review March 12

- Indexing algorithms, continued
  - Compressing index structures
- Using user characteristics and feedback
  - Search refinement
  - Collaborative filtering
- Other analysis of information objects
  - Clustering
  - Detecting near-duplicate documents
- Semi-structured information
  - XML
- Crawling the Web

3

Reading assigned to date in *Introduction to Information Retrieval*

- 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 – all sections **including 7.1.6 (uses clustering)**
- Chapter 8 – all sections

Bold indicates assigned ***since*** first exam

4

More reading assigned to date in *Introduction to Information Retrieval*

- **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

Bold indicates assigned ***since*** first exam

5

More reading assigned to date in *Introduction to Information Retrieval*

- 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

Bold indicates assigned ***since*** first exam

6