CSC 483/583: MIDTERM REVIEW OUTLINE

MIHAI SURDEANU

At the exam:

- You can bring a 2-page sheet with your notes.
- You are allowed a simple, self-contained hand-calculator. Internet-connected devices are **not allowed** under any circumstances.

Topics to know for the midterm:

1. Lecture 1: Introduction and Boolean retrieval
   - A. Definition of information retrieval
   - B. Term-document incidence matrix: definition, how to build it
   - C. Inverted index: definition, how to build it, cost (runtime) of building it, why is it better than the incidence matrix?
   - D. Algorithm for intersection
   - E. Algorithms for other Boolean operators (see Homework #1)
   - F. Query optimization
2. Lecture 2: Term vocabulary and postings list
   - A. What is a document?
   - B. Token vs. term
   - C. Tokenization issues
   - D. Stop words, stemming, lemmatization
   - E. Skip pointers
   - F. Phrase queries, biword indexes
   - G. Positional indexes
   - H. Algorithm for proximity intersection
3. Lecture 3: Dictionaries and tolerant retrieval
   - A. Hashes vs. binary trees vs. B-trees

Date: Spring 2017.

   - B. Permuterm trees
   - C. k-gram index
   - D. Edit distance, including reading out operations
   - E. Spelling correction using k-gram indexes
   - F. Context sensitive spelling correction
   - G. Soundex algorithm

4. Lecture 4: Index construction
   - A. Single-pass in-memory indexing (SPIMI)
   - B. Note: Block sort-based indexing (BSBI), and the remaining topics after SPIMI in this chapter are not required for the midterm
5. Lecture 5: Index compression
   - A. Why compression?
   - B. Lossy vs. lossless compression
   - C. Heap's law
   - D. Zipf's law
   - E. Dictionary compression not required for the midterm
   - F. Postings compression: gap encoding with variable-length encoding, gamma codes
6. Lecture 6: vector space model
   - A. Feast of famine for Boolean queries
   - B. Jaccard coefficient: where else is this useful? Limitations
   - C. tf-idf
   - D. Vector space model
   - E. Cosine similarity
   - F. Different ways of encoding: term frequency, document frequency, normalization
7. Lecture 7: Complete search system
   - A. User studies for ranking
   - B. Tiered indexes
   - C. Zone indexes, proximity ranking, scoring functions with multiple components

D. Combinations of multiple scoring models, e.g., boolean and vector-space models, phrase-based and vector-space models

E. Query parser

F. Exact top K retrieval using min heap

G. Inexact top K retrieval: document at a time, term at a time, cluster pruning

8. Lecture 8: Evaluation

   A. Unranked evaluation: Precision, Recall, F score

   B. Accuracy. Why is Accuracy not a good measure?

   C. Ranked evaluation: P@1, precision-recall curve, mean average precision (MAP), mean reciprocal rank (MRR)

   D. Inter-annotator agreement: Kappa measure

   E. Real-world evaluations: A/B testing