CSC 483/583: MIDTERM REVIEW OUTLINE

MIHAI SURDEANU

Before we begin:

- The exam will be open book. You are welcome to bring any textbooks, notes, etc.

- 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. Algorithm 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

Date: Fall 2015.

    A. Hashes vs. binary trees vs. B-trees

    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. Block sort-based indexing (BSBI)

    B. Single-pass in-memory indexing (SPIMI)

    C. Note: 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: as single string, string with blocking, front coding

    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. Note: The remaining topics in this chapter (i.e., exact and inexact top K retrieval) are not required for the midterm