Introduction to

Information Retrieval

Review

Final Exam

- Exam date: 8 Dec (Wed) afternoon, Exact time to be announced
- Mock exam to be arranged on 1 Dec (Wed).
- Mark = 0.25 * ass + 0.25 * proj + 0.5 * exam
  - No double pass
  - Supplementary exam is only for those who cannot attend final exam.
  - Apply for UNSW Special Consideration (SC) with sufficient evidence and the SC team will make the final decision.

Final Exam - 2

- Time: 10 minutes reading time + 2 hr open-book exam + 20 minutes scanning+uploading+submission time.
- Very important for you to know how to scan, upload, and submit. Practice before-hand!
- Designed to test your understanding and familiarity of the core contents of the course.
- 100 (6-8 questions)
  - Similar to those in the assignment.

Special Note on the Final Exam

- We trust every student will uphold the academic integrity.
- Severe consequences for any misconduct in the final exam.

About the Final Exam . . .

- Read the instructions carefully.
- You can answer the questions in any order.

- Tip: Write down intermediate steps, so that we can give you partial marks even if the final answer is wrong.

Boolean Model

- incidence vector/matrix
- semantics of the query model (AND/OR/NOT, and other operators, e.g., /k, /S)
- inverted index, positional inverted index
- query processing methods for basic and advanced boolean queries (including phrase query, queries with /S operator, etc.)
- query optimization methods (list merge order, skip pointers)

Preprocessing

- typical preprocessing steps: tokenization, stopword removal, stemming/lemmatization,

Tolerant Retrieval

- Wildcard queries
- Permuterm index
- Bigram index
- Spelling correction
- Noise channel model

Index Construction

- Why we need dedicated algorithms to build the index?
- BSBI: Blocked sort-based indexing
- SPIMI: Single-pass in-memory indexing
- Dynamic indexing: Immediate merge, no merge, logarithmic merge

Index Compression

- Heap's law, Zipf's law
- Dictionary compression
  - Dictionary as a string
  - Front encoding
- posting lists compression
  - Elias encoding
  - Variable length encoding
- **Not required:** Shannon limit

Vector Space Model

- What is/why ranked retrieval?
- raw and normalized tf, idf
- cosine similarity
- tf-idf variants (using SMART notation): e.g., lnc.ltc
- basic query processing method: document-at-a-time vs term-at-a-time
- exact & approximate query optimization methods (heap-based top-k algorithm, MaxScore algorithms, etc.)
- **Not required:** Query processing methods based on advanced or tiered inverted indexes (e.g., high/low lists, impact-oriented lists, etc.)

Evaluation

- Existing method to prepare for the benchmark dataset, queries, and ground truth
- For unranked results: Precision, recall, F-measure
- For ranked results: precision-recall graph, 11-point interpolated
- precision, MAP, etc.
- **Not required:** Kappa (κ) measure for inter-judge (dis)agreement

Web Search Basics

- Estimation of relative sizes of two search engines.
- Near duplicate detection: the shingling method

Crawling

- Understand the requirements of crawlers
- Mercator scheme
- Not required: optimization for age

Thank you and good luck!