COMP6714 Review

Wei Wang

weiw AT cse.unsw.edu.au

School of Computer Science and Engineering
Universities of New South Wales

November 11, 2020
Course Logistics

- **THE formula:**
  
  \[
  mark = \begin{cases} 
  0.25 \cdot (ass1 + proj1) + 0.50 \cdot \text{exam} & , \text{if } exam \geq 40 \\
  39FL & , \text{otherwise.}
  \end{cases}
  \]

- Exam date: Exact time to be announced, 2 Dec (Wed) afternoon.
- Pre-exam consultations:
  - TBA
  - TBA
- Sample exam papers to be released soon.
- Course survey or private messages to me on the forum.

(1) The final exam mark is important and you must achieve at least 40! (2) Supplementary exam is **only** for those who cannot attend final exam. (3) Apply for UNSW Special Consideration (SC) with sufficient evidence and the SC team will make the final decision.
About the Final Exam

- **Time**: 10 minutes reading time + 2 hr open-book exam + 15 minutes scanning+uploading+submission time.
  - Very important for you to know how to scan, upload, and submit. Practice before-hand!! We will launch a practice session before hand.
- Designed to test your understanding and familiarity of the core contents of the course.
- 100 (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.
- Some of the “Advanced” Methods/algorithms/systems are not required, unless explicitly mentioned here.

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

**Disclaimer:** We will go through the main contents of each lecture. However, note that it is by no means exhaustive.
Boolean Model

- incidence vector
- 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)
- **Not required**: next-word index
Preprocessing

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

- What is/why ranked retrieval?
- raw and normalized tf, idf
- cosine similarity
- tf-idf variants (using SMART notation): e.g., Inc.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 and WAND 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**: NDCG, Kappa ($\kappa$) measure for inter-judge (dis)agreement
Probabilistic Model and Language Model

- Probability ranking principle (intuitively, how to rank documents and when to stop)
- Derivation of the ranking formula of the probabilistic model
- The BM25 method
- Query-likelihood unigram language model with Jelinek-Mercer smoothing.
Web Search Basics

- Difference between Web search and Information Retrieval.
- Estimation of relative sizes of two search engines.
- Near duplicate detection: the shingling method.
- **Not required**: the SimHash method.
Crawling

- Understand the requirements and the current architecture of crawlers (e.g., the Mercator architecture).
- **Not required**: optimization for age, finding content blocks, etc.
Link Analysis

- The pagerank algorithm: theory and practice
- **Not required**: the topic-specific/personalized pagerank
Thanks and Good Luck!