COMP6714 Review

Wei Wang

weiw AT cse.unsw.edu.au

School of Computer Science and Engineering
Universities of New South Wales

October 20, 2018

Course Logisitics

- **THE** formula:

$$mark = \begin{cases} 0.20 \cdot (ass1 + proj1) + 0.60 \cdot exam & \text{, if } exam \geq 40 \\ 39FL & \text{, otherwise.} \end{cases}$$

- Exam date:
- Pre-exam consultations: TBD
- 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.

About the Final Exam

- **Time**: DATE (TBD), 10 minutes reading time + 2 hr closed-book exam.
- **Accessories**: UNSW Approved Calculator. Note: watches are prohibited.
- Designed to test your *understanding* and familiarity of the core contents of the course.
- 100 (8 questions)
  - Q1: short answer questions
  - Q2–Q8:
    - choose any 5 to answer.
    - others will require some "calculation" or more steps.

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,
- NLP preliminaries: Part of Speech (POS) tagging
- **Not required**: details of the (Porter's) stemming algorithm.

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 algorithm, 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.

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*.

Deep Learning Basics

- Classification/regression: problem setting, and other related concepts.
- Feed-forward neural network:
  - Architecture, activiation and loss functions.
  - Gradient descent
  - Backpropagation
- Recurrent NN:
  - Elman's RNN
  - LSTM

Note:

- No need to memorize complex formula (e.g., 5 equations of LSTM); they will be given, and make sure you understand them.
- Simple, important, or very frequently used equations like σ, cross-entropy loss, etc. won't be given.

Language Models

- Definition, usage, and evaluation
- *n*-gram LM
  - Parameter learning, including various smoothing
- Neural LMs
  - Bengio's
  - word2vec
- Pros and cons of NLM