COMP6714 20t3
Course Info

• LiC: Prof. Wei Wang, K17-507, x51762
  – http://www.cse.unsw.edu.au/~weiw

• Research interests
  – Database + AI
  – Knowledge Graph and NLP
  – Adversarial machine learning

• Homepage: http://www.cse.unsw.edu.au/~cs6714

• Email:
  – piazza (for Q&As): https://piazza.com/class/keps8hkcron4ef
  – Lecture time:
    • Wed 1600 – 1800
    • Fri 1400 – 1600

• Consultations:
  – collaborative consultations on piazza
  – weekly consultation in Zoom (TBA)

I will enrol students manually by the end of W1 and again by the end of W2. No need to send emails to me.
Assessment

- Assignment 1: 25%
- Project 1: 25%
- Final exam: 50%

39FL if final exam mark < 40 (out of 100)

(details to be available later)
Reminder

• Online learning is even harder to keep up
  – Good time management and planning skill required
• Know the regulation on Academic Integrity and Plagiarism:
  https://student.unsw.edu.au/plagiarism
  – Not an exhaustive list of instances of misconducts
  – Pay attention to:
    • Collusion, including Contract cheating/ghost-writing
    • Copying, including from the Internet
Expectation

• What are expected in this course
  – Many modules covering a **broad spectrum** of IR/NLP/SE
  – Heavy workload expected: must read and digest the **textbook and slides + additional notes**
  – Requires substantial **algorithm/data structure** design/analysis experience & capability + some maths.
  – ML/DL knowledge will be a plus
  – **Up-to-date** viewpoints, understanding, knowledge (from the academia & industry)
    – ➔ Plan your time well
• I speak fast
  – ➔ focus
  – ➔ ask questions – you are helping your classmates too
• Review after the lecture
Real Learning

- After
  - You know the answer
  - You forgot
  - You made mistakes

- Life-long learning is inevitable

- Learn the right learning method
  - Rote learning is **USELESS**

Source: http://qaspire.com/2016/01/18/when-does-real-learning-happen/
Requirements

• Lectures
  – My lecture does not follow closely the textbook, and will try to give you a different perspective whenever possible

• Python notebooks
  – Do the exercise by yourself
  – Very helpful in understanding concepts/algorithms
What’s New?

• New contents:
  – some new contents added this semester
  – please let me know the glitches you’ve found

• Welcome your feedback (throughout the course)
Knowledge Assumed (non-exhaustive)

• Data structures & algorithms:
  – Heap/priority queue: build a heap in O(n) time?
  – Membership query: tradeoffs? worst/avg-case time complexities = ?
  – Recursion:
  – DFS/BFS/Best-first search

Given an array A of integers. Design an algorithm to return two elements x, y in A, such that x + y = 100 if any, and
1. the algorithm takes O(n*log(n)) time, or
2. the algorithm takes O(n) time
Knowledge Assumed (non-exhaustive)/2

- C/C++ & Python Programming:
  - Pointer
  - `sizeof(int) = ?` `sizeof(p) = ?` `sizeof(*p) = ?` `sizeof(str) = ?`
  - Be able to learn to use new Python libraries and write & debug python programs
    - Quickly learn a python-based framework in this course
Knowledge Assumed (non-exhaustive)/3

• CS Architecture
  – Memory hierarchy: name the levels?
  – Bit representation: binary string for any x? How to obtain the 3rd-5th bits of a byte?

• Maths
  – Calculus: How to find the minimum/minimal value of a function f(x)?
  – Probabilities and statistics: \( rv; \) linearity of expectation; indicator variable; number of heads by tossing a biased coin n times; Bayesian theorem
  – Linear algebra: inner/dot product of \( \mathbf{u} \) and \( \mathbf{v} = \) ? matrix multiplication
Introduction
Search and Information Retrieval

• Search on the Web is a daily activity for many people throughout the world
• Search and communication are most popular uses of the computer
• Applications involving search are everywhere
• The field of computer science that is most involved with R&D for search is information retrieval (IR)
Information Retrieval

• “Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968)

• General definition that can be applied to many types of information and search applications

• Primary focus of IR since the 50s has been on text and documents
What is a Document?

• Examples:
  – web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, etc.

• Common properties
  – Significant text content
  – Some structure (e.g., title, author, date for papers; subject, sender, destination for email)
Documents vs. Database Records

- Database records (or *tuples* in relational databases) are typically made up of well-defined fields (or *attributes*)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is more difficult
Documents vs. Records

• Example bank database query
  – *Find records with balance > $50,000 in branches located in Amherst, MA.*
  – Matches easily found by comparison with field values of records

• Example search engine query
  – *bank scandals in western mass*
  – This text must be compared to the text of entire news stories
Comparing Text

• Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval

• Exact matching of words is not enough
  – Many different ways to write the same thing in a “natural language” like English
  – e.g., does a news story containing the text “bank director in Amherst steals funds” match the query?
  – Some stories will be better matches than others
Dimensions of IR

• IR is more than just text, and more than just web search
  – although these are central
• People doing IR work with different media, different types of search applications, and different tasks
Other Media

• New applications increasingly involve new media
  – e.g., video, photos, music, speech
• Like text, content is difficult to describe and compare
  – text may be used to represent them (e.g. tags)
• IR approaches to search and evaluation are appropriate
# Dimensions of IR

<table>
<thead>
<tr>
<th>Content</th>
<th>Applications</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Web search</td>
<td>Ad hoc search</td>
</tr>
<tr>
<td>Images</td>
<td>Vertical search</td>
<td>Filtering</td>
</tr>
<tr>
<td>Video</td>
<td>Enterprise search</td>
<td>Classification</td>
</tr>
<tr>
<td>Scanned docs</td>
<td>Desktop search</td>
<td>Question answering</td>
</tr>
<tr>
<td>Audio</td>
<td>Forum search</td>
<td></td>
</tr>
<tr>
<td>Music</td>
<td>P2P search</td>
<td></td>
</tr>
<tr>
<td><strong>JSON</strong></td>
<td>Literature search</td>
<td></td>
</tr>
</tbody>
</table>

*Semistructured*
IR Tasks

• Ad-hoc search
  – Find relevant documents for an arbitrary text query

• Filtering (aka information dissemination)
  – Identify relevant user profiles for a new document

• Classification
  – Identify relevant labels for documents

• Question answering
  – Give a specific answer to a question
Big Issues in IR

• Relevance
  – What is it?
  – Simple (and simplistic) definition: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
  – Many factors influence a person’s decision about what is relevant: e.g., task, context, novelty, style
  – *Topical relevance* (same topic) vs. *user relevance* (everything else)
Big Issues in IR

• Relevance
  – Retrieval models define a view of relevance
  – Ranking algorithms used in search engines are based on retrieval models
  – Most models describe statistical properties of text rather than linguistic
    • i.e. counting simple text features such as words instead of parsing and analyzing the sentences
    • Statistical approach to text processing started with Luhn in the 50s
    • Linguistic features can be part of a statistical model
Big Issues in IR

• Evaluation
  – Experimental procedures and measures for comparing system output with user expectations
    • Originated in Cranfield experiments in the 60s
  – IR evaluation methods now used in many fields
  – Typically use test collection of documents, queries, and relevance judgments
    • Most commonly used are TREC collections
  – Recall and precision are two examples of effectiveness measures
Big Issues in IR

• Users and Information Needs
  – Search evaluation is user-centered
  – Keyword queries are often poor descriptions of actual information needs
  – Interaction and context are important for understanding user intent
  – Query refinement techniques such as query expansion, query suggestion, relevance feedback improve ranking
Web Search

• New Challenges

• How to obtain data?

• Additional features for Web data
NLP

• Formal representation of semantics
  – "set" in wordnet,
  – Mamihlapinatapai

• Common sense knowledge
  – kittens are cute
  – SJC often experiences delays.

• Inference:
  – The cat ate a mouse → No carnivores eat animals
  – Pr[A went to Primary School in B → A was born in B] = 0.613

The ball did not fit into the box because it was too big/small.
Some of the Things We’ll Learn to Do

• Language modelling
  – ... as soon as [???] ...

• Semantic relation
  – Synonyms: aubergine = ???
  – Compound nouns:
    • “apple cake” vs “birthday cake”
    • What about “parsley cake”?

• Sequence modelling
  – ”time flies like an arrow”

All Based on Deep Learning for NLP
Models
Select one of the available models

English GoogleNews Negative300

Nearest words
Given a word, this demo shows a list of other words that are similar to it, i.e. nearby in the vector space.

Show nearest
Case sensitive:  Top N: 100

aubergine

eggplant
mauve
fennel
coriander
tarragon
cauliflower
paprika
fuchsia
shallots
apricot
parmesan
parsley
time flies like an arrow
Course Goals

• To help you to understand search engines, evaluate and compare them, and modify them for specific applications

• Provide broad coverage of the important issues in information retrieval, search engines, and natural language processing
  – includes underlying models and current research directions