COMP6714 17s2

Course Info

- LiC: Prof. Wei Wang, K17-507, x51762
  - http://www.cse.unsw.edu.au/~weiw
  - Knowledge Graph: http://kg.cse.unsw.edu.au
  - Similarity query processing / high-dimensional data
  - DB + ML
- Homepage: http://www.cse.unsw.edu.au/~cs6714
- Email:
  - **piazza**: https://piazza.com/configure-classes/summer2017/comp6714 for Q&As
- Lecture time: Fri 1200 – 1500
- Consultations:
  - will mainly offer consultations on piazza
  - otherwise, by appointment only

Assessment

- Proj1: 20%
- Proj2: 20%
- Final exam: 60%

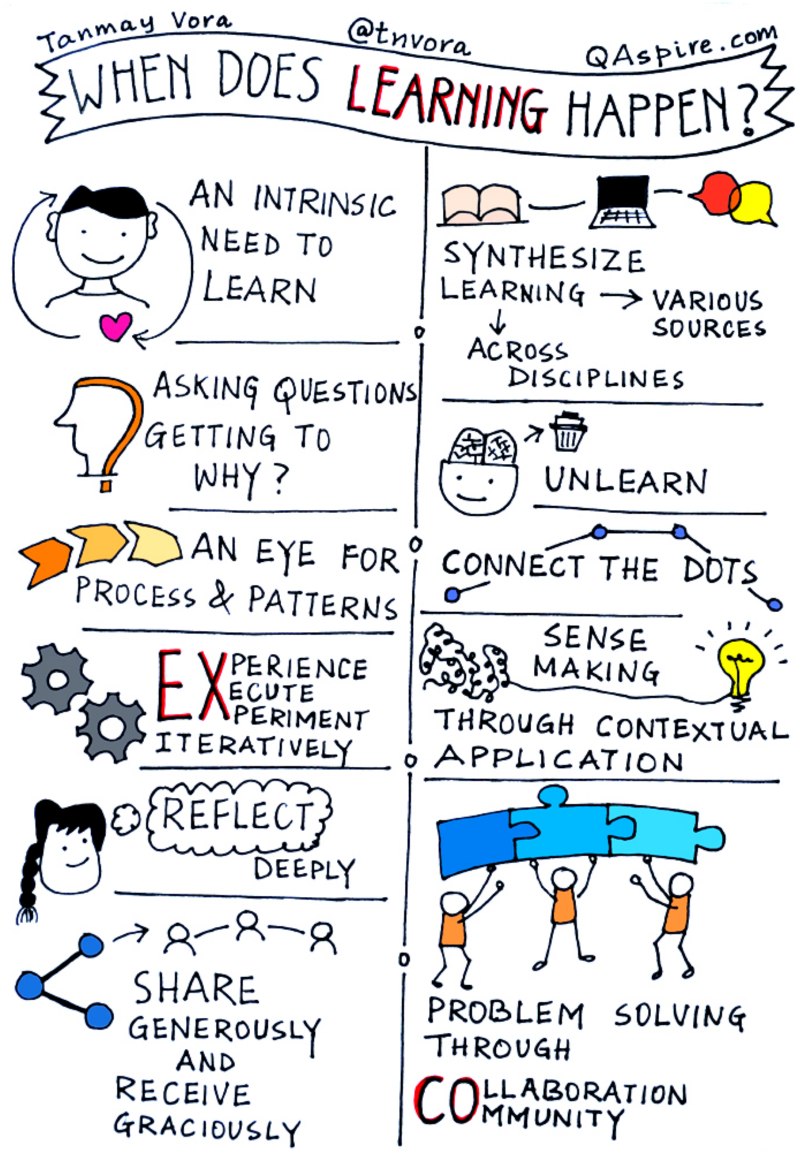
39FL if final exam mark < 40 (out of 100)

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
  - **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

- Lecture
  - Read book chapters and slides before hand
  - My lecture typically will give you a different perspective to the material
- Python notebook
  - Do the exercise by yourself
  - Very helpful in understanding concepts/algorithms

What’s New?

- New contents:
  - incorporate an additional module on vector representation of words
- 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

Knowledge Assumed (non-exhaustive)/3

- CS Architecture
  - Memory hierarchy: name the levels?
  - Bit representation: binary string for any x? How to obtain the 3^{rd}-5^{th} 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 **u** and **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

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

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

IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections
- Web search engines are best-known examples, but many others
  - *Open source* search engines are important for research and development
    - e.g., Lucene, Lemur/Indri, *Galago*
- Big issues include main IR issues but also some others

IR and Search Engines

Information Retrieval

Relevance

-Effective ranking

Evaluation

-Testing and measuring

Information needs

-User interaction

Search Engines

Performance

-Efficient search and indexing

Incorporating new data

-Coverage and freshness

Scalability

-Growing with data and users

Adaptability

-Tuning for applications

Specific problems

-e.g. Spam

Search Engine Issues

- Performance
  - Measuring and improving the efficiency of search
    - e.g., reducing *response time*, increasing *query throughput,* increasing *indexing speed*
  - *Indexes* are data structures designed to improve search efficiency
    - designing and implementing them are major issues for search engines

Search Engine Issues

- Dynamic data
  – The “collection” for most real applications is constantly changing in terms of updates, additions, deletions
    - e.g., web pages
  – Acquiring or “crawling” the documents is a major task
    - Typical measures are *coverage* (how much has been indexed) and *freshness* (how recently was it indexed)
  – Updating the indexes while processing queries is also a design issue

Search Engine Issues

- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithm, indexing strategy, interface for different applications

Spam

- For Web search, spam in all its forms is one of the major issues
- Affects the efficiency of search engines and, more seriously, the effectiveness of the results
- Many types of spam
  - e.g. spamdexing or term spam, link spam, “optimization”
- New subfield called *adversarial IR*, since spammers are “adversaries” with different goals

Natural Language Processing (NLP)

- NLP:
  - Sentence <---> Meaning
- Main challenges: Ambiguity
  - "Time flies like an arrow"
    - at least 5 possible ways of syntactically parsing, hence interpretations
  - "Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo"
    - Buffalo buffalo [Buffalo buffalo buffalo/VB] buffalo/VB Buffalo buffalo
  - "I saw a girl with a telescope": Who has the telescope?

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

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

Specialised Courses

- Other specialised courses in the Database or Data Science stream:
  - COMP9319: Advanced algorithms on compression, text/XML databases, etc.
  - COMP9313: Big data systems (hadoop, spark, etc)
  - COMP9318: Data mining / machine learning.