3

Cour

- LiC: Prof. Wei Wang, K17-507, x51762
  - http://www.cse.unsw.edu.au/~weiw
- Research interests
  - Database + AI
  - Knowledge Gra
  - Adversarial ma
- Homepage: http:/ 714
- 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.

Online lectures (pre-recorded videos)

Asses

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

details to be available later

39FL if fin out of 100)

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

Expe

- 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 subst design/analysis experience & c
  - ML/DL knowled
  - **Up-to-date** viewpoints, understan e (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/

Requir

- Lectures
  - My lecture does not follow closely the textbook, and will try to give you a different perspective whenever
- Python notebooks
  - Do the exercise by yourself
  - Very helpful in understanding concepts/algorithms

- New contents:
  - some new contents added this semester
  - please let ou’ve found
- Welcome you the course)

Knowledge Assum -exhaustive)

- Data structures & algorithms:
  - Heap/priority queue: build a heap in O(n) time?
  - Membersh orst/avg-case time compl
  - 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 Assum -exhaustive)/2

- C/C++ & Python Programming:
  - Pointer
  - sizeof(int) *p) = ? sizeof(str) = ?
  - Be able to learn to use n libraries and write & debug python pr
    - Quickly learn a python-based framework in this course

Knowledge Assum -exhaustive)/3

- CS Architecture
  - Memory hierarchy: name the levels?
  - Bit representation: binary string for any x? How to obtain the 3rd th
- Maths
  - Calculus: How to find the inimal 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

Search and Info n Retrieval

- Search on the Web is a daily activity for many people throughout the world

- Search and most popular uses of the

- Applications involving s everywhere
- The field of computer science that is most involved with R&D for search is *information retrieval (IR)*

Informati ieval

- *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, a mation."* (Salton, 196
- General definition that plied to many types of information and search applications
- Primary focus of IR since the 50s has been on *text* and *documents*

What is a ent?

- Examples:
  - web pages, email, books, news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum post ons, 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 fiel

  - e.g., bank r numbers, balances, names, addres security numbers, dates of birth,

- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is more difficult

Document cords

- Example bank database query
  - *Find records with balance > $50,000 in branches located in Amherst, MA.*
  - Matches e json with field values of records
- Example search engine
  - *bank scandals in western mass*
  - This text must be compared to the text of entire news stories

Compa xt

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval
- Exact match enough
  - Many diffe same thing in a "natural language" like
  - 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

Dimens IR

- IR is more than just text, and more than just web search
  - although t
- People doin rent media, different types of searc ions, and different tasks

Other

- New applications increasingly involve new media
  - e.g., video,
- Like text, co describe and compare
  - text may be used to represent them (e.g. tags)
- IR approaches to search and evaluation are appropriate

Dimens IR

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video		ssification
Scanned docs		nswering
Audio	Forum searc	
Music	P2P search	
	Literature search	

IR

- Ad-hoc search
  - Find relevant documents for an arbitrary text query
- Filtering (ak mination)
  - Identify re a new document
- Classification
  - Identify relevant labels for documents
- Question answering
  - Give a specific answer to a question

Big Iss R

- Relevance
  - What is it?
  - Simple (and simplistic) definition: A relevant document ion that a person was looking ed 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 Iss R

- Relevance
  - **Retrieval models** define a view of relevance
  - **Ranking algorithms** used in search engines are based on r
  - Most mod 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 Iss R

- Evaluation
  - Experimental procedures and measures for comparing system output with user expectations
    - Originate in the 60s
  - IR evaluation methods n many fields
  - Typically use *test collecti* ments, queries, and relevance judgments
    - Most commonly used are TREC collections
  - *Recall* and *precision* are two examples of effectiveness measures

Big Iss R

- Users and Information Needs
  - Search evaluation is user-centered
  - Keyword qu criptions of actual info
  - Interaction and context ant for understanding user inte
  - Query refinement techniques such as **query expansion**, **query suggestion**, **relevance feedback** improve ranking

Web

- New Challenges

- How to obt

- Additional features for

- Formal representation of semantics
  - "set" in wordnet,
  - Mamihlapinatapai
- Common sen
  - kittens are
  - SJC often experiences dela
- 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 Thing… Learn to Do

- Language modelling
  - … as soon as [???] …
- Semantic relation
  - Synonyms:
  - Compound nouns:
    - “apple cake” vs “birthday
    - What about “parsley cake”?
- Sequence modelling
  - ”time flies like an arrow”

All Based on Deep Learning for NLP

Cours

- To help you to understand search engines, evaluate and compare them, and modify them for specific a
- Provide bro important issues in information ra arch engines, and natural language processing
  - includes underlying models and current research directions