CS3245

Information Retrieval

Lecture 0: Course Organization

0

Why should you care about

INFORMATION RETRIEVAL?

A keystone of today's tech

A necessity in the era of big data

The 3V's of Big Data

Volume
- Amount of data generated
- Online & offline transactions
- In kilobytes or terabytes
- Saved in records, tables, files

Velocity
- Speed of generating data
- Generated in real-time
- Online and offline data
- In streams, batch or bits

Variety
- Stuctured & unstructured
- Online images & videos
- Human generated - texts
- Machine generated - readings

https://blog.hurree.co/blog/the-pros-and-cons-of-big-data

Part of an ongoing evolution

V 1.0

Directories

- Organized by a human
- Broad subject categories organized by hierarchy
- **Select and read**

V 2.0

Search Engines

- Information retrieval (poor man's NLP) over billions of websites and pages
- Algorithms personalize the searching experience and rank results based on the search criteria
- **Free text search**

V 3.0

Conversational AI

- Advanced natural language understanding using machine learning and computational linguistics
- Knows user history and conversation state based on previous interactions
- Knowledge summarization and task completion
- **Natural language conversations**

https://medium.com/@sidjreddy/conversational-artificial-intelligence-in-the-context-of-information-revolution-a3257867d50b

More than just a piece of technology

https://trends.withgoogle.com/year-in-search/2024/

Focus of the course

Boolean
Retrieval

Language
Models

Dictionaries and
Tolerant Retrieval

Index
Compression

Postings Lists and
Choosing Terms

Index
Construction

Scoring and
Vector Space
Model

Evaluation

Probabilistic IR

A Complete
Search System

Relevance Feedback
and Query Expansion,
and XML IR

Crawling and
Link Analysis
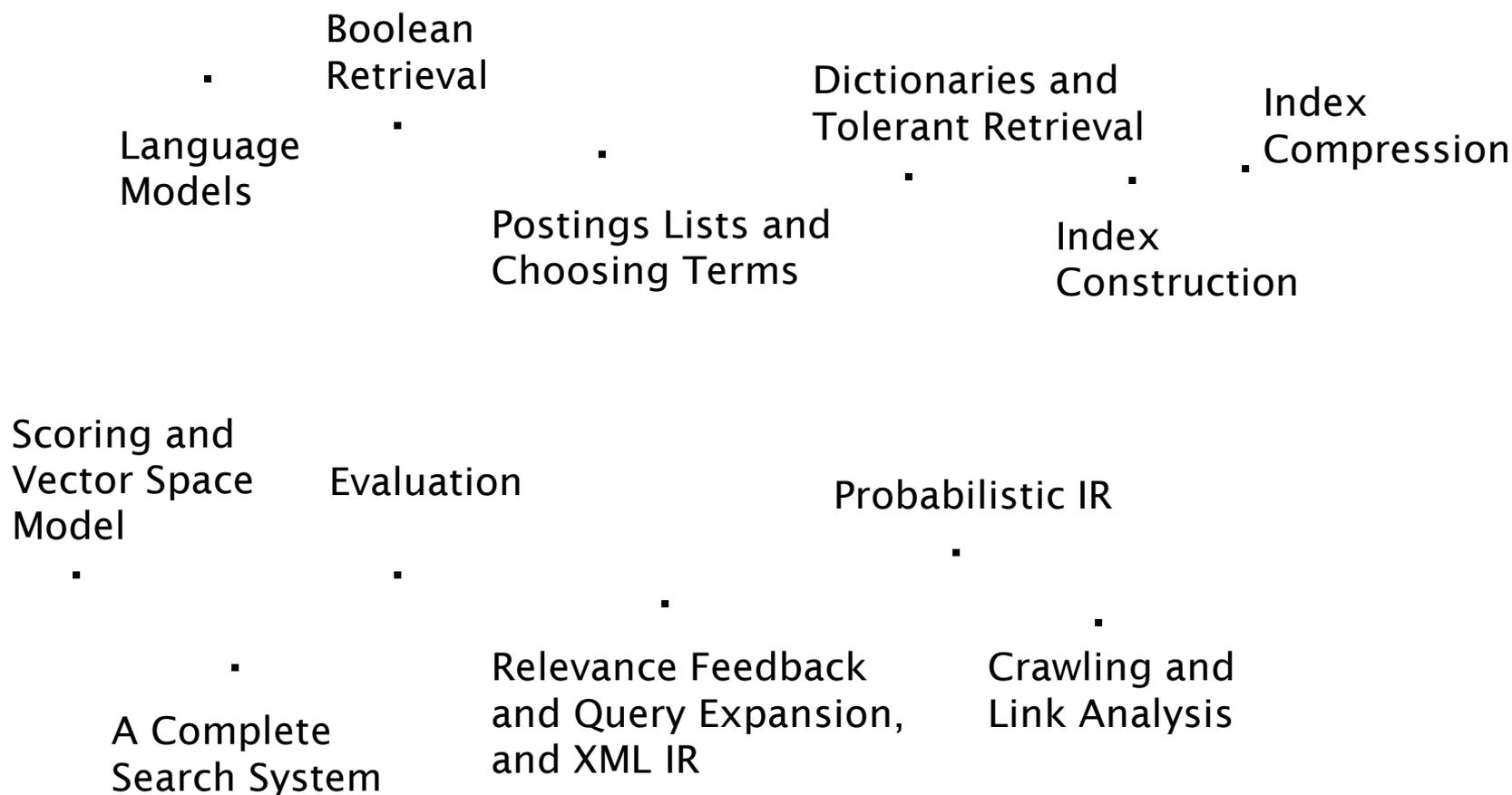
And now for the

COURSE ORGANIZATION

Lecturer

Zhao Jin zhaojin@comp.nus.edu.sg

(Just call me "**Jin**" for short)

- Undergraduate, PhD, Instructor, (Senior) Lecturer and Assistant Dean

- Programming Methodology, Orbital, Information Retrieval

Teaching Assistants

- ## Sim Shi Qian
  - ### e1154547@u.nus.edu
- ## Ou Jin Bin
  - ### e0902046@u.nus.edu
- ## Bhardwaj Tannish
  - ### e1070595@u.nus.edu

Course activities

- 12 (physical) **lectures** (Weeks 1-12)
  - 12-2pm on the Fridays
  - Recorded on a best effort basis
  - Video Conference Room (COM1-VCRM at level 2)

  - Makeup lecture in Week 10 due to Well-Being day, No lecture in Week 13 due to Good Friday

  - Each with an **online quiz** via Canvas (12 **quizzes** in total)

Course activities

- 5 (physical, optional) **tutorials** (odd weeks in general)
  - 4 groups: 11am-12nn / 12nn-1pm (Thu), 10-11am/11am-12nn (Fri)
  - Not recorded (solutions to be released afterwards)
  - For discussions, clarifications and participation marks
  - SR 9 (COM1-02-09)

  - Makeup tutorials in Week 4 due to CNY, via an online video in Week 13 due to Good Friday
  - Tutorials in Weeks 4/5/7/9/11

Course activities

- 4 **homework assignments**
  - maximum group size: 1/2/2/4

- 1 (physical, open-book) **final exam**

- 2 (online, optional) **help sessions** (via Zoom)
  - 3-5pm on the **Friday of Week 2** for technical issues
  - During the **Exam Weeks** for exam matters + Past year paper
  - Recorded on a best effort basis.

Grading

Component	Percentage	Remarks
Participation	5%	Forum + Tutorial (optional)
4 Homework assignments	40% (5%/10%/10%/15%)	Due in **Weeks 4/7/10 and Reading Week**
12 Quizzes	5% (0.5% each)	Due **weekly** before the subsequent lecture
Final Exam	50%	**6 May (Tue)**, 9-11am

Course web sites

https://www.comp.nus.edu.sg/~cs3245

- Official website
- General information
- Course materials
- Homework assignments

https://canvas.nus.edu.sg/

- Canvas course
- Announcements
- Recordings
- Internal resources (e.g., past year papers)
- Homework submissions

When using a search engine to find our course materials, make sure you find our site for 24/25 Sem II.

Course web sites

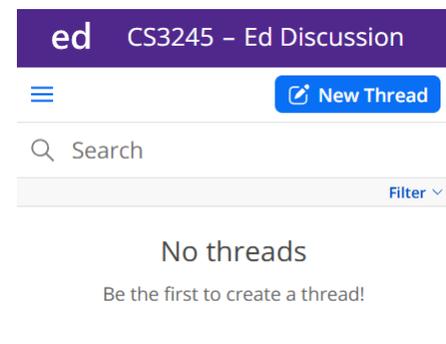https://edstem.org/us/courses/70898/discussion

- Ed forum
- Discussion (+ Teammate recruitment)
- Participation counts!

https://pollev.com/jin

- PollEverywhere
- Live Q&A during lectures

Required Textbook

- https://nlp.stanford.edu/IR-book/information-retrieval-book.html
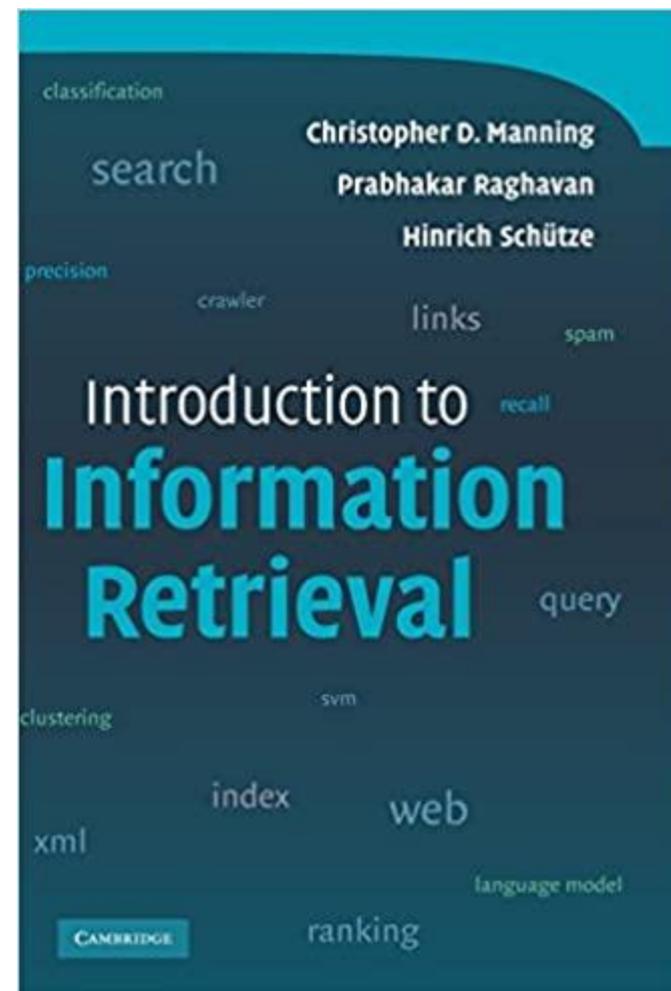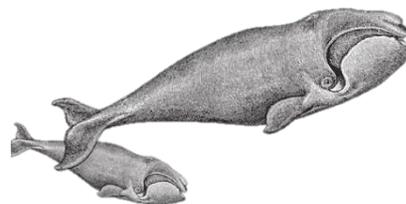
Python

- To be used for all homework assignments.

- Easy to transition to and comes with many useful libraries (e.g., for text processing and statistics).

- Some materials to be released for your reference.

- https://www.python.org/

NLTK (Natural Language Toolkit)

- To be used for all homework assignments
  - Some restrictions apply.

- Coded in Python (surprise!)
  - Yes, we know that this is a course on **information retrieval** and not **natural language processing**, but the two are forever intertwined.

- https://www.nltk.org/

SoC Compute Cluster + Slurm

- A high-performance computing cluster in SoC
  - For running (and grading) your homework submissions, which can be quite compute intensive.
  - Python Version: 3.10.12
  - Managed by Slurm Workload Manager

- Activation
  - https://mysoc.nus.edu.sg/~myacct/services.cgi

Login w/ your NUSNET account

The SoC Compute Cluster is a Unix based compute cluster on which you can run your compute intensive jobs. You can also use this to do parallel computing experiments.

This service is currently enabled

Disable

SoC Compute Cluster + Slurm

- ## Login
  - Turn on VPN if you are off-campus
  - Use ssh or any other similar tools
  - Server: xlog[0-2].comp.nus.edu.sg
  - Login w/ Your **SoC Unix** account

  Apply for one @ https://mysoc.nus.edu.sg/~newacct/ if you don't have one.

```
C:\Users\dcszjin>ssh zhaojin@xlog0.comp.nus.edu.sg
zhaojin@xlog0.comp.nus.edu.sg's password:
Welcome to Ubuntu 20.04.5 LTS (GNU/Linux 5.4.0-126-ge
```

- ## User guides
  - Compute Cluster Guide:
    https://dochub.comp.nus.edu.sg/cf/guides/compute-cluster/start
  - Slurm Quick Start Guide:
    https://dochub.comp.nus.edu.sg/cf/guides/compute-cluster/slurm-quick

To-dos

- Install Python + NLTK on your computer if you plan to work on your homework assignments locally.
  - Remember to match the Python version in the Cluster.

- Familiarize yourself with the Cluster + Slurm
  - Install NLTK on your account in the Cluster
  - Try to run Python programs in the cluster using Slurm

- Any technical issues...
  - Post in the forum
  - Attend the technical help session

Gilligan's Island Rule

- Discussion is acceptable and encouraged.

- However, NO written record (electronic or otherwise) shall be taken away from a discussion.

- Do something else for at least a half-hour before switching back to CS3245.
  - E.g., check Instagram, watch Youtube, play games

- This will assure that you are able to reconstruct what you learned from the discussion, by yourself, using your own brain.

Freedom of Information Rule

- If you collaborate with other students for your assignments but are not making a group submission with them, you must always fill in the name(s) of your collaborators on your assignment.

- You will be assessed for the parts for which you claim is your own contribution.

- Failing to do so will be considered as plagiarism.

No-Sponge Rule

- For homework assignments which group submissions are allowed, **every member in the group must actively contribute**.

- Members of the group have the responsibility to
  - NOT tolerate anyone who is putting forth no effort
  - NOT let anyone who is making a good faith effort "fall through a crack"

- Let us know about the dysfunctional group situations as early as possible.

Use of AI Tools

- OK for learning purposes.

- NOT allowed for all homework assignments
  - You must **implement the core logic on your own**.

Plagiarism

- A serious violation of NUS code of student conduct.
    - https://studentconduct.nus.edu.sg/wp-content/uploads/NUS-Code-of-Student-Conduct.pdf

- "No mercy" policy
    - Both parties, the student who copied, and the student who allowed others to copy, will be penalized equally.

- Bottom Line: Write your own code and do NOT share it with anyone!