Introduction
Welcome to CMPT 456

- A course about information, search technology, big data, and data analytics
  - Many hot (probably the hottest) topics nowadays!
- Instructor: Jian Pei
- TA: Weiyuan Wu and Nguyen Cao
- Office hours: starting in Week 2
  - Instructor: 10-11 am, Fridays, TASC I 9429
  - TA: TBA
- The instructor has to go several business trips, and some classes may be canceled, or provided using video or by invited speakers
  - November 4, 6, 8, 22, December 2
About the Course

• A comprehensive introduction to information retrieval techniques with the focus on search engines and (web) search
• Major theme: key components in search engines and the related information retrieval techniques
• Focuses: information retrieval from big data
  – Fundamental principles and techniques
  – User experience
  – (if time permits) Monetarization and value-added services and business
Lab Time

• Please search “school lunch” using Google
WeChat

- A phenomenal social networking app
- Daily active users: 750 million
- Average daily active time per user: 30 minutes
- Number of chats per day: 45 billion in 2019, 38 billion in 2018
- Active users: 1.08 billion in 2018
Content (Subject to Change)

• Search engine architecture
• Crawling
• Indexing
• Retrieval and ranking
• Evaluation
• Link analysis
• Social networks and social media
Emphases

- Practical algorithms instead of NLP techniques
- Solid knowledge about information retrieval
- Problem solving skills: developing feasible techniques to solve practical information retrieval and web search problems
- Implementation skills: implement some important information retrieval and Web search techniques
- Connecting techniques and business: understanding how techniques can be used in business
Textbook and Lecture Notes

  - The textbook contains a lot of details you need to learn
  - E-book available

- Attending the classes and taking notes
  - The textbook does not cover all materials in this course
  - The lecture notes on the course web page are incomplete – taking notes in classroom will make them complete
  - Exams will be based on the lecture notes and the textbook
  - Don’t take this course if you cannot attend classes regularly

- Some figures in my slides are borrowed from the textbook and references
Schedule and Grading Scheme

• 5 writing/programming assignments (45% in total)
• 2 in-class quizzes (5% each), no advance notification
  – Cover anything taught 2+ weeks ago
• Final exam (3 hours, 45%): December 9 (Monday), 8:30-11:30 am
  – Covering the whole course

• Bring to classroom a laptop computer / iPad / smartphone that has access to the web
Prerequisites

• Comprehensive understanding and skills in data structures, such as linked data structures, B-trees, and hash functions
• Java/Python programming
• Analysis of algorithms and time complexity
• Essentials in operating systems, main memory and disk management, file systems
• Elementary probability theory and statistics, such as random variables, distributions, probability mass functions, and sampling
Expectations

• Strong incentives to learn new knowledge, not just passing another course
• Come up with new ideas creatively to solve real world problems
• Work hard to gain solid understanding and skills
Promises

• Strong reference letters for A+ students
• Discussion on entrepreneur ideas
• Research internship opportunities in my lab for creative students
Questions and Suggestions?
Data and Information

• Data: values of qualitative or quantitative variables, belonging to a set of items
  – An abstract concept
  – Can be viewed as the lowest level of abstraction from which information and then knowledge are derived

• Information: a sequence of symbols that can be interpreted as a message
Discussion

• Where are data and information come from?

• How can people find data and information?
Search – Examples

• Search using search engines
  – Billions of web pages on millions of computers
• Search for the final exam policy at the university
  – Enterprise, institutional and domain-specific search
• Instant Search at Windows, Spotlight at Mac OS and in-app search at smart phones
  – Personal information retrieval
• In many situations, “search” and “information retrieval” can be used interchangeably
Information Retrieval

“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching and retrieval of information”

— Gerard Salton, 1968
Retrieve from What?

• Usually documents (the focus of this course), can be extended to pictures, video, music, audio or speech

• In general, can be anything, such as sensors (internet of things), and unstructured in nature
Structured or Unstructured?

• From the computer point of view
  – Not in a relational database
  – Not easy for a computer to understand
• Still may be structured for human being
• Text, image, video, audio, and combinations
• What collection? Usually stored on computers
  – Information retrieval started from libraries
IR Searches Versus DB Queries

- DB queries are on attributes well defined in format
  - Example: find students whose last name is “Jackson” and whose GPA is over 3.5
- IR searches are on large text bodies
  - Example: find news stories about “data science”
  - A news story about data science may not contain term “data science” exactly
IR/Search Applications

- **Usual search**: typing in a query, receiving answers in the form of a list of documents in ranked order
- **Vertical search**: the domain of the search is restricted to a particular topic
- **Enterprise search**: finding the required information in the huge variety of computer files scattered across a corporate intranet
- **Desktop search**: the personal version of enterprise search
  - The information sources are the files stored on an individual computer, including email messages and Web pages that have recently been browsed
- **Peer-to-peer search**: finding information in networks of nodes or computers without any centralized control
IR/Search Tasks

- **Ad hoc search**
- **Filtering/tracking**: detecting stories of interest based on a user’s interests and providing an alert using email or some other mechanism
- **Classification/categorization**: using a defined set of labels or classes and automatically assigning those labels to documents
  - Yahoo Directory http://dir.yahoo.com
- **Question answering**: aiming at more specific questions and returning specific answers instead of a list of documents
  - Example: “Which Canadian university was the first to offer academic courses on a year-round trimester system, and since when?” – answer: SFU, since 1965
## Three 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 documents</td>
<td>Desktop search</td>
<td>Question answering</td>
</tr>
<tr>
<td>Audio</td>
<td>Peer-to-peer search</td>
<td>…</td>
</tr>
<tr>
<td>Music</td>
<td>Literature search</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Relevance

• Is a text document relevant in the sense of containing the information that the user was looking for?
  – Exact matching? Often poor results
  – Vocabulary mismatch problem

• Topical relevance: a document is topically relevant to a query if it is on the same topic
  – A news story about a tornado in Kansas is topically relevant to the query “severe weather events”

• User relevance: a document is user relevant if the user is interested in
  – If a user has read the story, or if the story is too old or in a language that the user does not understand, the user may not be interested
Retrieval Models

• A formal representation of the process of matching a query and a document

• Typically, retrieval models capture the statistical properties of text rather than the linguistic structure
  – More sophisticated models incorporate linguistic features, but tend to be of secondary importance
  – An important difference from NLP
Simple Models vs. Complex Models

• “Who shot Abraham Lincoln?”
  – NLP methods use sophisticated linguistic processing techniques, such as syntactic and semantic analysis

• A surprisingly simple, data-driven, redundancy-based method
  – Simply searching for the phrase “shot Abraham Lincoln” on the web
  – Look for names frequently appears on those pages

• Simple models are as good as sophisticated ones when large amount of data is used
Evaluation

- How well does a document ranking match users’ expectations?
- Precision: the proportion of retrieved documents that are relevant
- Recall: the proportion of relevant documents that are retrieved
  - Assumption: all the relevant documents for a given query are known – may not hold all the time
  - Using test collections (corpora): samples of typical queries and a list of relevant documents for each query
  - The best-known test collections: TREC (http://trec.nist.gov)
  - Clickthrough data: records of documents that were clicked on during a search session
Information Needs

• “What are the courses at SFU talking about document indexes?”
  – Issue a query “course, SFU, document indexes” to a search engine
  – Do the query results answer the question?

• Information need: the topic about which the user desires to know more
  – Unfortunately, often cannot be fed into a search engine
Queries

• Query: what the user conveys to the computer in an attempt to communicate the information need
  – Multiple queries may be formed to capture an information need
  – A query may not capture the information need sufficiently
  – One-word queries are very common in web search

• Query refinement by query suggestion, expansion and relevance feedback
Search Engines

- Practical application of IR techniques to large scale (text) databases
- Web search engines, e.g., Google, Bing, ...
  - Crawl many terabytes of data, provide sub-second response times to millions of queries everyday around the world
- Enterprise search engines, e.g., Verity, Autonomy, ...
  - Process the large variety of information sources in a company, and use company-specific knowledge (e.g., data mining) as part of search and related tasks
- Desktop search engines
  - Rapidly incorporate new documents, web pages, and email, and provide an intuitive interface for search this heterogeneous mix of information
- One real system can be configured as different search engines
Open Source Search Engines

• Many systems exist
• **Lucene**: a popular Java-based search engine
  – Has been used for a wide range of commercial applications
  – The IR techniques there are relatively simple
• **Lemur**: an open source toolkit that includes the Indri C++-based search engine
  – Has primarily been used by IR researchers to compare advanced search techniques
Big Issues in Search Engines

- Performance and scalability
  - Response time: the delay between submitting a query and receiving the result list
  - Throughput: the number of queries that can be processed in a given time
  - Indexing speed: the rate at which text documents can be transformed into indexes for searching
- Coverage: how much of the existing information has been indexed and stored in the search engine database
- Recency and freshness: the “age” of the information in the search engine database
- Customizability and adaptability: how much a search engine can be used in different applications?
Core IR and Search Engine Issues

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
IR & Search in a Bigger Picture

• Many challenges can be modeled as search/information retrieval problems
  – Example: search in data lakes
• An information infrastructure view
An Information Infrastructure View

Search/queries/information consumption

Search engine

Data
An End-to-End Interaction View

User -> Information needs -> Value -> Information -> Data

User

Information needs

Value

Information

Data
Implicit Queries

• Using the context to predict a user’s information need
  – Location
  – Time
  – Action
  – …

• Granularity of information needs
  – Queries
  – Tasks
  – …
Example: Yelp
To-Do List

• Read Chapter 1 in the textbook
• How do people compare Google and Bing? Find out from the web by search
• Assignment 1