Introduction to Information Retrieval

CS 7263 Information Retrieval
Lecture 00

Instructor: Jiho Noh

Department of Computer Science
Kennesaw State University

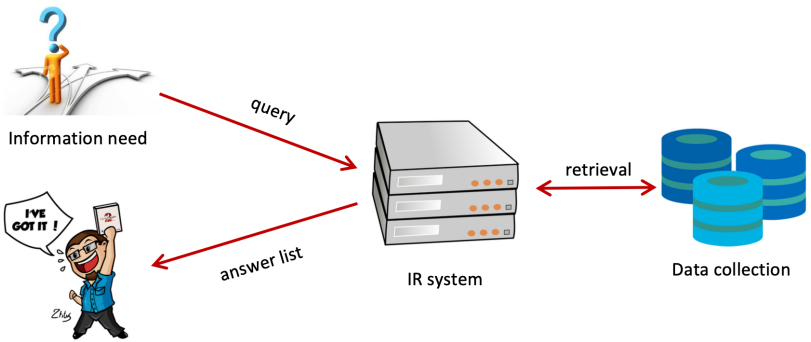
Fall 2025

About this course

Course Details: Find the syllabus on the D2L course page

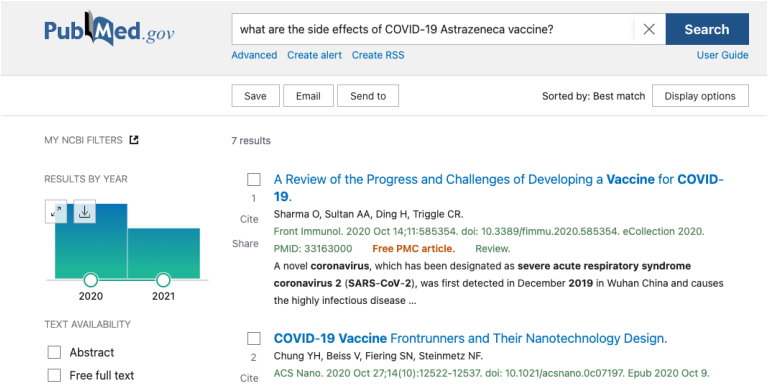
Information Retrieval (IR) Systems

- The goal of IR is
  - to find a small set of items
  - relevant to the user's information need
  - from a large collection of data.



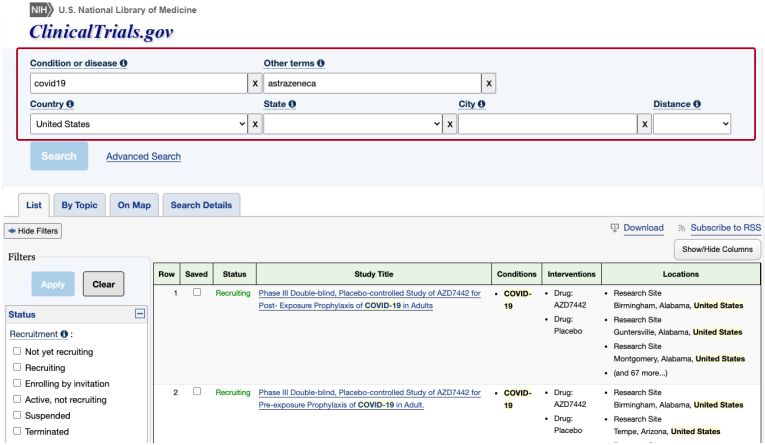
Document Search Engine — PubMed

PubMed is a search engine of references and abstracts of life science and biomedical topics.



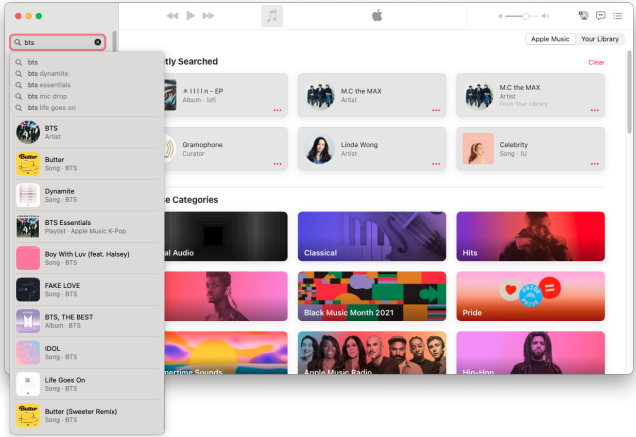
Document Search Engine — Clinical Trials

Clinical Trials is a database of clinical studies conducted around the world



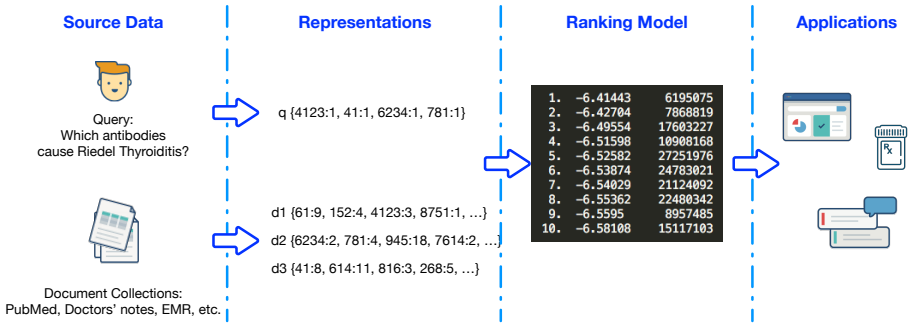
Search for Other Contents — Songs

Examples of Content — text, images, videos, music, any other items for recommendation systems

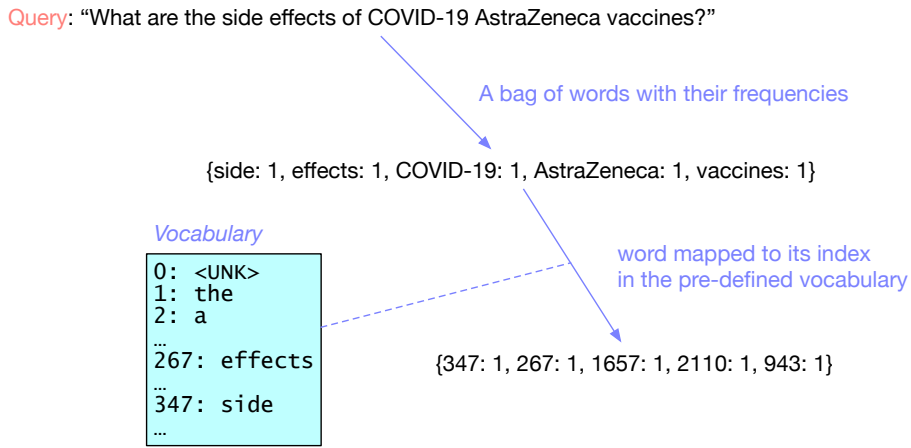


Document Retrieval Pipeline

- Document Retrieval is the process of semantic matching between a stated user query and free-text documents



Bag-of-words Model



Ranking Documents

Query (q):

- What are the side effects of COVID-19 AstraZeneca vaccines?

A Document (d):

- The most common side effects with COVID-19 Vaccine AstraZeneca in the trials were usually mild or moderate pain and tenderness at the injection site, headache, tiredness, muscle pain. ...

```
q: {347:1, 267:1, 1657:1, 2110:1, 943:1}
d: {0:14, 1:163,
        ..., 347:3, 267:3, 1657:17, 2110:5, 943:7, .. }
```

A ranking model measures the relevance of d to q, such as BM25(q, d) which are typically based on exact-term matching.

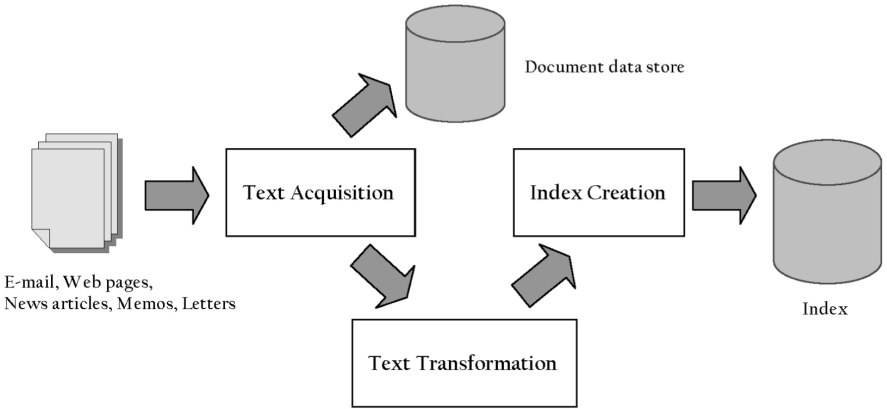
Relevance

Relevance is a subjective judgment and may include:

- Being on the proper subject
- Being timely (recent information)
- Being authoritative (from a trusted source)
- Satisfying the goals of the user and his/her intended use of the information (information need)

Indexing Process

- The **indexing process** builds the structure that enable searching: collect information from external resource, process data, create index, and store them for searching needs.



Vocabulary Mismatch

- Vocabulary Mismatch is a common phenomenon in the usage of natural language, occurring when different people name the same thing or concept differently.
- q: What are the side effects of COVID-19 AstraZeneca vaccines?
- d+: The most common risks associated with coronavirus vaccines are mild pain at the injection site, . . .
- d-: The economic effects of COVID-19 containment measures . . . On the supply side . . ., until when vaccines are widely available . . .

Vocabulary Mismatch (Cont.)

Problems with keywords

- May not retrieve relevant documents that include synonymous terms.
  - *restaurant* vs. *café*
  - *PRC* vs. *China*
- May retrieve irrelevant documents that include ambiguous terms.
  - *bat* (baseball vs. mammal)
  - *Apple* (company vs. fruit)
  - *bit* (unit of data vs. act of eating)

Beyond Keywords

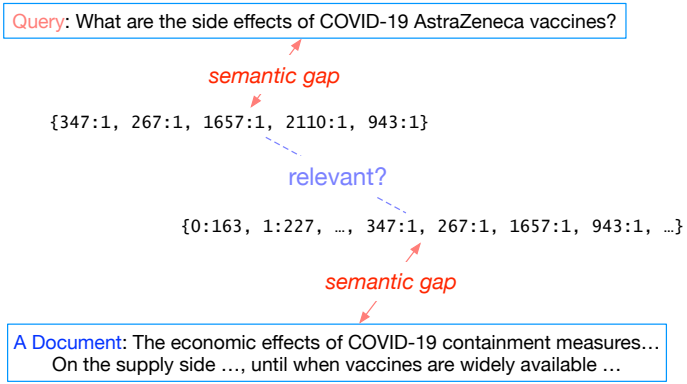
- We will cover the basics of keyword-based IR, but ...
- We will focus on extensions and recent developments that go beyond keywords.
- We will cover the basics of building an efficient IR systems,

but . . .

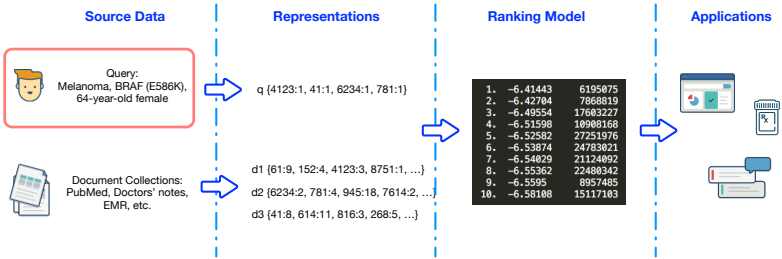
- We will focus on basic capabilities and algorithms rather than systems that allow scaling to industrial size databases.

Semantic Gap

- Semantic Gap is the difference between two descriptions of a theme by different linguistic representations.

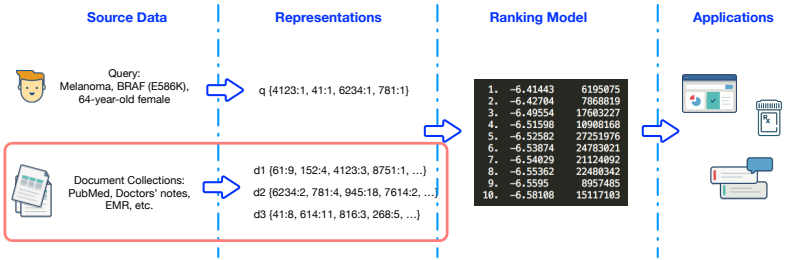


Course Objectives



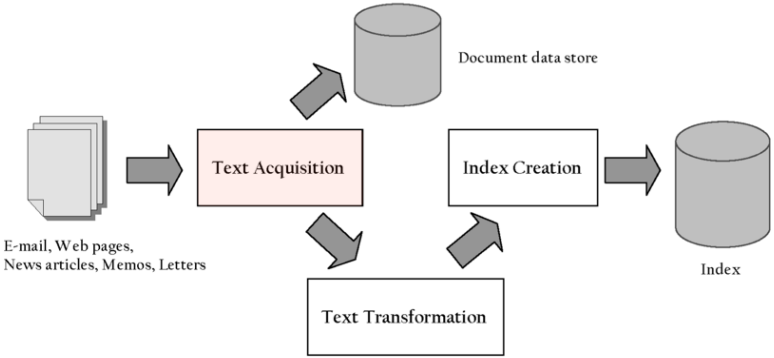
- Query Transformation and Refinement
  - How can we enhance the query representation of the user's information need?
  - spell checking, query expansion, relevance feedback, controlled vocabulary, · · ·

Course Objectives (Cont.)



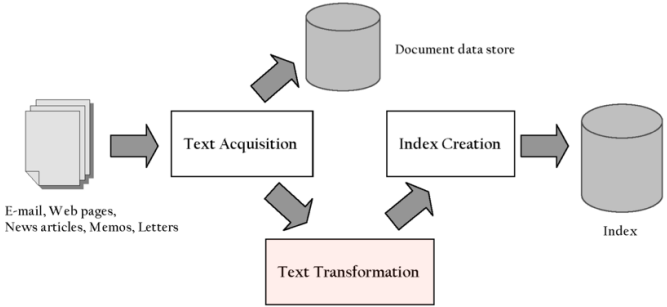
- Retrieval Models
  - What are the retrieval models and how can we improve the scoring methods for measuring the document relevance?
  - set theory/algebraic/probabilistic models, TF-IDF, Okapi BM25, ...

Course Objectives (Cont.)



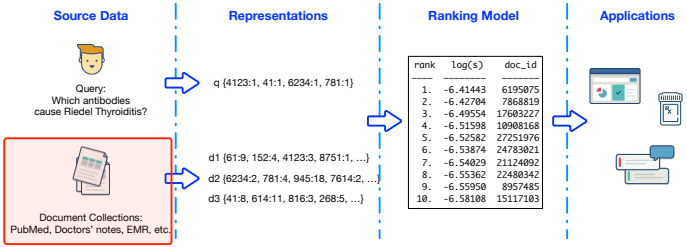
- Web Crawler
  - How do we collect and store information on the Internet, effectively and efficiently?
  - HTTP requests, URL management, graph traversal, webpage access policies, database, detecting duplicates, ...

Course Objectives (Cont.)



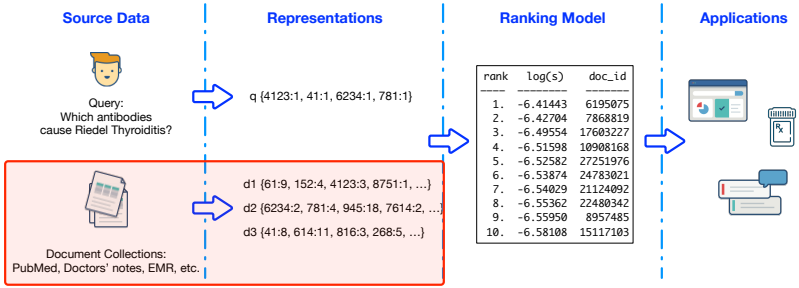
- Text Processing
  - How do analyze and process the collection of the Internet data to simplify searching?
  - text statistics, techniques for handling textual data, document parsing, lemmatization, stopping, stemming, encoding schemes, ...

Course Objectives (Cont.)



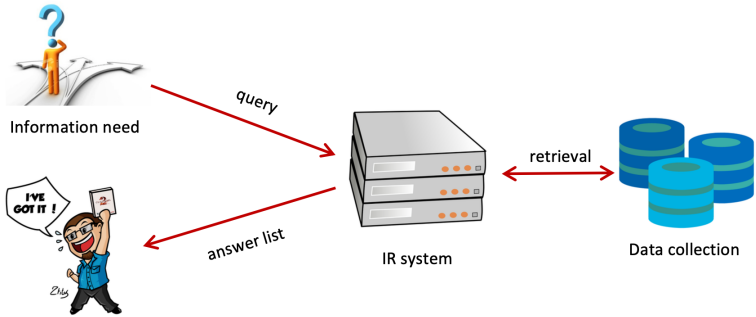
- Machine Learning Approaches for Document Understanding
  - How can we leverage machine learning methods to better understand the 'real' meaning of documents?
  - text classification, clustering, topic modeling, learning-to-rank, ...

Course Objectives (Cont.)



- Deep Learning Approaches
  - How can we bridge the semantic gaps between source data and their representations?
  - neural networks, word embeddings, language models and transformer

Course Objectives (Cont.)



- Deep Learning Approaches
  - What are the evaluation methods to measure the utility of the IR system?
  - IR evaluation metrics, ranking measures

Beyond Information Retrieval

- Semantic Understanding
- Multimodal Retrieval
- Personalization
- Cross-domain recommendation
- Large Language Models for IR
- Retrieval Augmented Machine Learning
- Question Answering
- Conversational IR models

Beyond Information Retrieval (Cont.)

- Interactive search
- FATE (Fairness, Accountability, Transparency, Ethics, and Explainability)
- Knowledge representation and reasoning
- Document representation and content analysis
  - Summarization
  - Readability
  - Opinion mining and sentiment analysis
  - ...

Summary

- Questions? Discussion?