

# Introduction to Information Retrieval

## CS4422/7263 Information Retrieval Lecture 00

Instructor: Jiho Noh

Department of Computer Science  
Kennesaw State University

CS4422/7263 Summer 2025



# 1 About this course

## 2 What is Information Retrieval (IR)

## 3 Challenges in IR

## 4 Course Objectives

# About this course

Course Details: Find the syllabus on the D2L course page

1 About this course

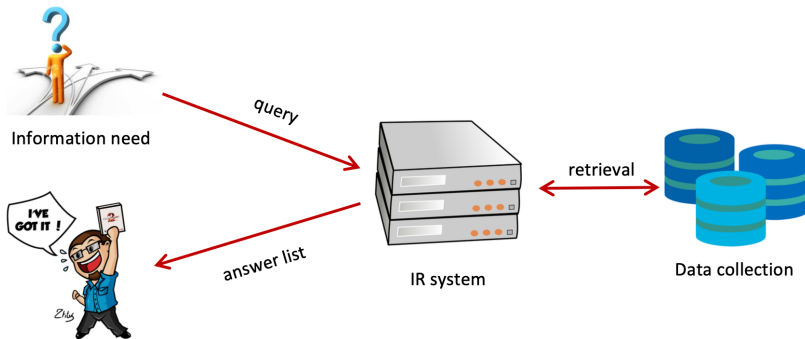
2 **What is Information Retrieval (IR)**

3 Challenges in IR

4 Course Objectives


# 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.

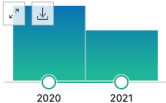
×Search

[Advanced](#) [Create alert](#) [Create RSS](#) [User Guide](#)

Save Email Send to Sorted by: Best match Display options

MY NCBI FILTERS 

RESULTS BY YEAR



2020 2021

TEXT AVAILABILITY

☐ Abstract

☐ Free full text

7 results

☐ [A Review of the Progress and Challenges of Developing a \*\*Vaccine\*\* for \*\*COVID-19\*\*.](#)

1

Cite Sharma O, Sultan AA, Ding H, Triggie CR.

Share Front Immunol. 2020 Oct 14;11:585354. doi: 10.3389/fimmu.2020.585354. eCollection 2020.

PMID: 33163000 [Free PMC article.](#) [Review.](#)

A novel **coronavirus**, which has been designated as **severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)**, was first detected in December **2019** in Wuhan China and causes the highly infectious disease ...

☐ [COVID-19 Vaccine Frontrunners and Their Nanotechnology Design.](#)

2

Cite Chung YH, Beiss V, Fiering SN, Steinmetz NF.

ACS Nano. 2020 Oct 27;14(10):12522-12537. doi: 10.1021/acsnano.0c07197. Epub 2020 Oct 9.

# Document Search Engine — Clinical Trials

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

NIH U.S. National Library of Medicine  
**ClinicalTrials.gov**

Condition or disease ⓘ

covid19

X

Other terms ⓘ

astrazeneca

X

Country ⓘ

United States

X

State ⓘ

X

City ⓘ

X

Distance ⓘ

Search Advanced Search

List By Topic On Map Search Details

Hide Filters

Download Subscribe to RSS

Show/Hide Columns

Filters

Apply Clear

Status

Recruitment ⓘ :

☐ Not yet recruiting

☐ Recruiting

☐ Enrolling by invitation

☐ Active, not recruiting

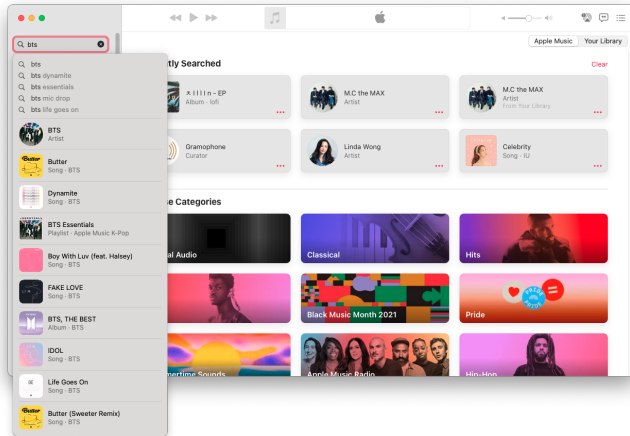
☐ Suspended

☐ Terminated

Row	Saved	Status	Study Title	Conditions	Interventions	Locations
1	<input type="checkbox"/>	Recruiting	<a href="#">Phase III Double-blind, Placebo-controlled Study of AZD7442 for Post- Exposure Prophylaxis of COVID-19 in Adults</a>	• COVID-19	• Drug: AZD7442 • Drug: Placebo	• Research Site Birmingham, Alabama, <b>United States</b> • Research Site Guntersville, Alabama, <b>United States</b> • Research Site Montgomery, Alabama, <b>United States</b> • (and 67 more...)
2	<input type="checkbox"/>	Recruiting	<a href="#">Phase III Double-blind, Placebo-controlled Study of AZD7442 for Pre-exposure Prophylaxis of COVID-19 in Adult.</a>	• COVID-19	• Drug: AZD7442 • Drug: Placebo	• Research Site Birmingham, Alabama, <b>United States</b> • Research Site Tempe, Arizona, <b>United States</b>

# Search for Other Contents — Songs

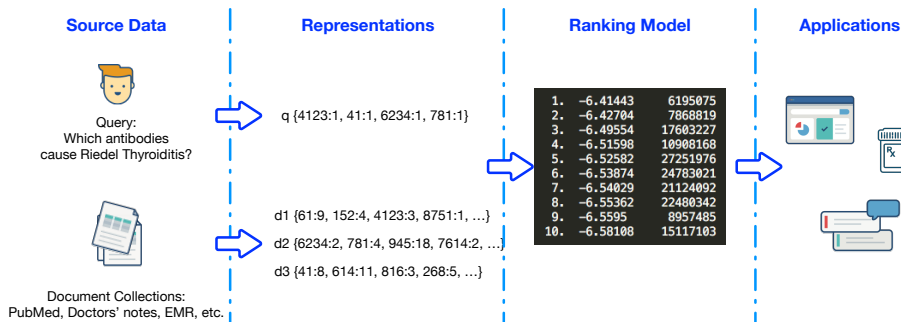
Examples of Content — text, images, videos, music, any other items for recommendation systems





# Document Retrieval Pipeline

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



# Bag-of-words Model

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

A bag of words with their frequencies

{side: 1, effects: 1, COVID-19: 1, AstraZeneca: 1, vaccines: 1}

Vocabulary

```
0: <UNK>
1: the
2: a
...
267: effects
...
347: side
...
```

word mapped to its index  
in the pre-defined vocabulary

{347: 1, 267: 1, 1657: 1, 2110: 1, 943: 1}

# 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**.

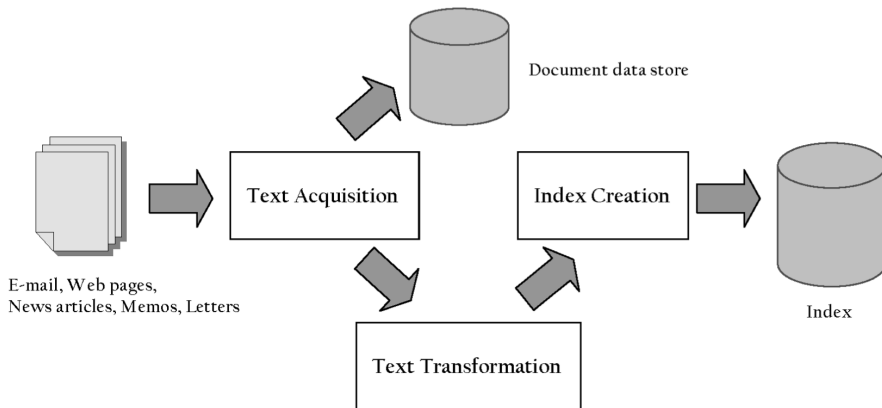
# Relevance

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.



1 About this course

2 What is Information Retrieval (IR)

3 Challenges in IR

4 Course Objectives

# 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.

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

*semantic gap*

{347:1, 267:1, 1657:1, 2110:1, 943:1}

*relevant?*

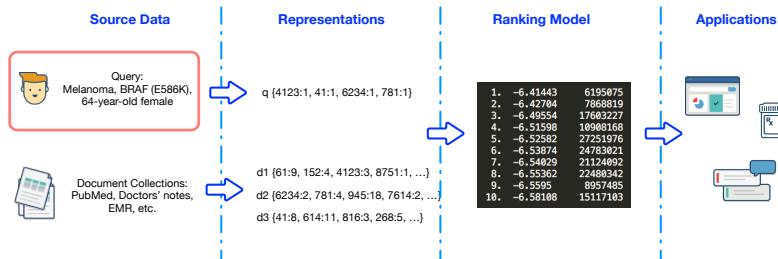
{0:163, 1:227, ..., 347:1, 267:1, 1657:1, 943:1, ...}

*semantic gap*

A Document: The economic effects of COVID-19 containment measures...  
On the supply side ..., until when vaccines are widely available ...

- 1 About this course
- 2 What is Information Retrieval (IR)
- 3 Challenges in IR
- 4 Course Objectives**

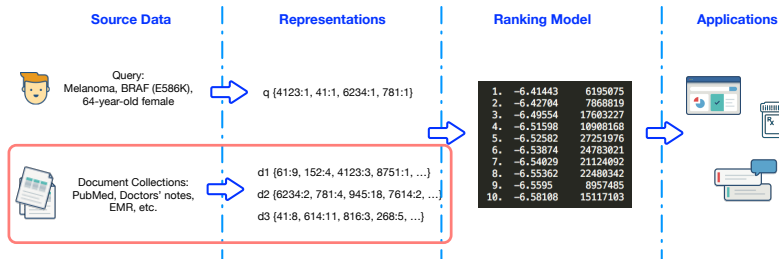
# Course Objectives



- 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, ...

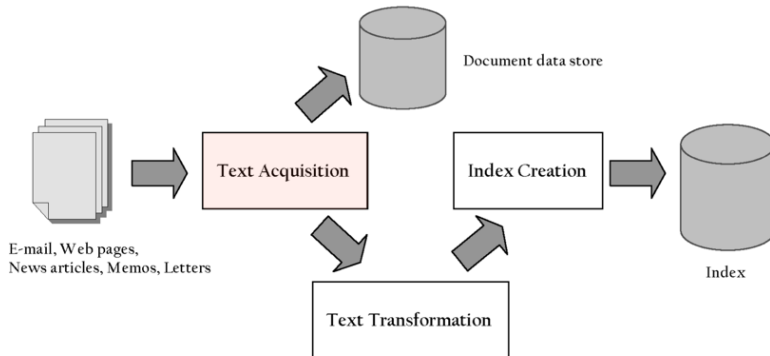
# Course Objectives (Cont.)



## ● 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, ...

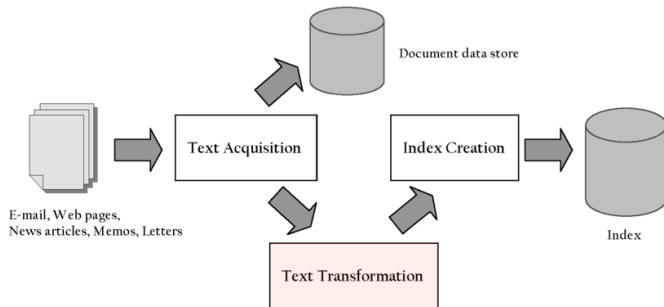
# Course Objectives (Cont.)



## • 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, ...

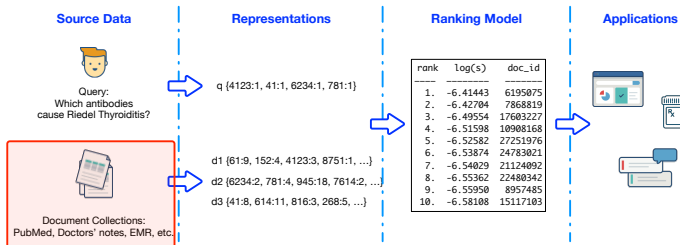
# Course Objectives (Cont.)



## ● 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.)

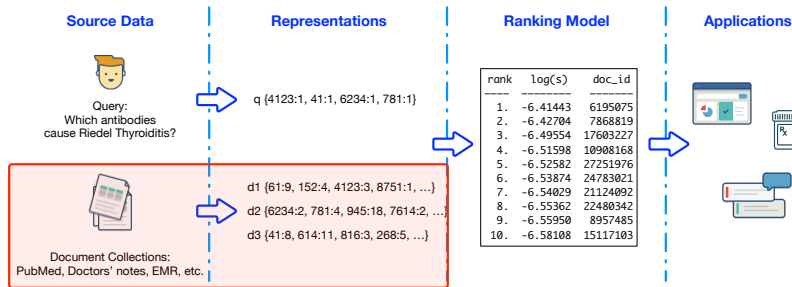


## Machine Learning Approaches for Document Understanding

- ▶ How can we leverage machine learning methods to better understand the 'real' meaning of documents?
- ▶ text classification, clustering, topic modeling, learning-to-rank, ...



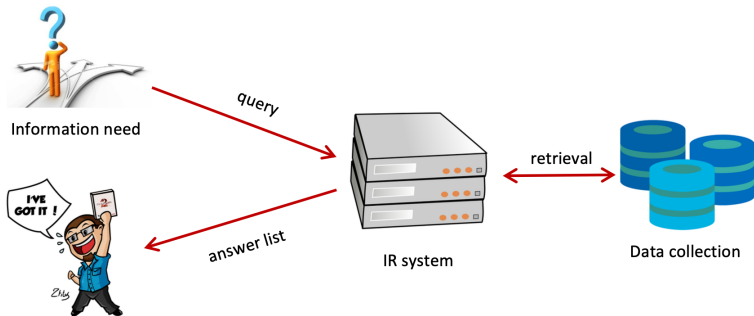
# Course Objectives (Cont.)



## ● 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?