CSCE 670 - Information Storage and Retrieval

Lecture 1: Welcome and Course Overview

Yu Zhang

yuzhang@tamu.edu

August 26, 2025

Course Website: https://yuzhang-teaching.github.io/CSCE670-F25.html

Adapted from the slides by Prof. James Caverlee

Course Website

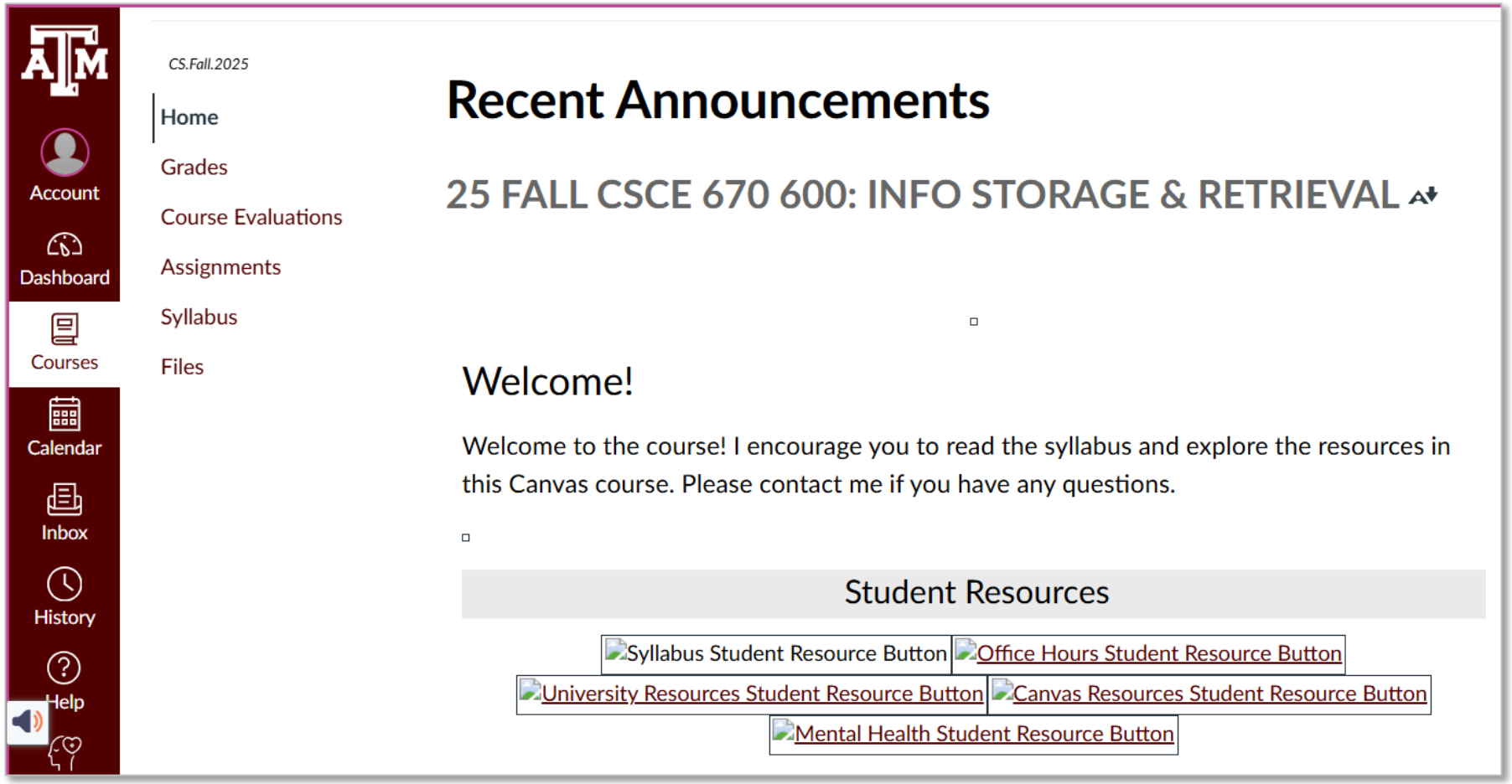
https://yuzhang-teaching.github.io/CSCE670-F25.html

Schedule (Subject to changes)

Week	Date	Topic	Slides	Optional Readings
W1	8/26	Overview and Course Logistics	PDF	-
	8/28	Boolean Retrieval	PDF	[MRS Chapter 1], [MRS Chapter 2]
W2	9/2	TF-IDF, Vector Space Model	PDF	[MRS Chapter 6], [MRS Chapter 7]
	9/4	BM25, Probabilistic Model	PDF	[MRS Chapter 11]
	9/6	**Homework 0 Due (Saturday)**		
W3	9/9	Link Analysis: PageRank		[MRS Chapter 21], [LRU Chapter 5.1/5.2]
	9/11	Link Analysis: Topic-Sensitive PageRank, HITS		[LRU Chapter 5.3/5.5]
W4	9/16	Evaluation		[MRS Chapter 8]
	9/18	Evaluation (Cont'd) **and Quiz 1**		[Sakai, SIGIR'16]

Canvas

https://canvas.tamu.edu/courses/403458



Course Logistics

- Course Website: Syllabus, Slides, Schedule, Optional Readings
- Canvas: Syllabus, Slides, Announcements, Homework, Discussions

- You may email me/TA directly (please put [CSCE670] in the subject).
- We prefer email over Canvas messages.

Grading (See Syllabus and Course Website for Details)

- Homework: 30%
  - Homework 0: 2% [due Sep 6]
  - Homework 1: 7% [due Sep 20]
  - Homework 2: 7% [due Oct 11]
  - Homework 3: 7% [due Nov 1]
  - Homework 4: 7% [due Nov 22]
- Quizzes: 20%
  - Quiz 1: 5% [in the Sep 18 class]
  - Quiz 2: 5% [in the Oct 7 class]
  - Quiz 3: 5% [in the Oct 30 class]
  - Quiz 4: 5% [in the Nov 20 class]
- Group Project: 20%
  - Proposal: 2% [due Oct 18]
  - Presentation: 9% [in the Dec 2 and Dec 4 classes]
  - Report: 9% [due Dec 9]
- Final: 30% [3:30pm – 5:30pm, Dec 16, HRBB 113]

Homework (30%)

- We will have 1 "getting-started" homework and then 4 real homework assignments
- Fun opportunity to put concepts into action
- All in Python!

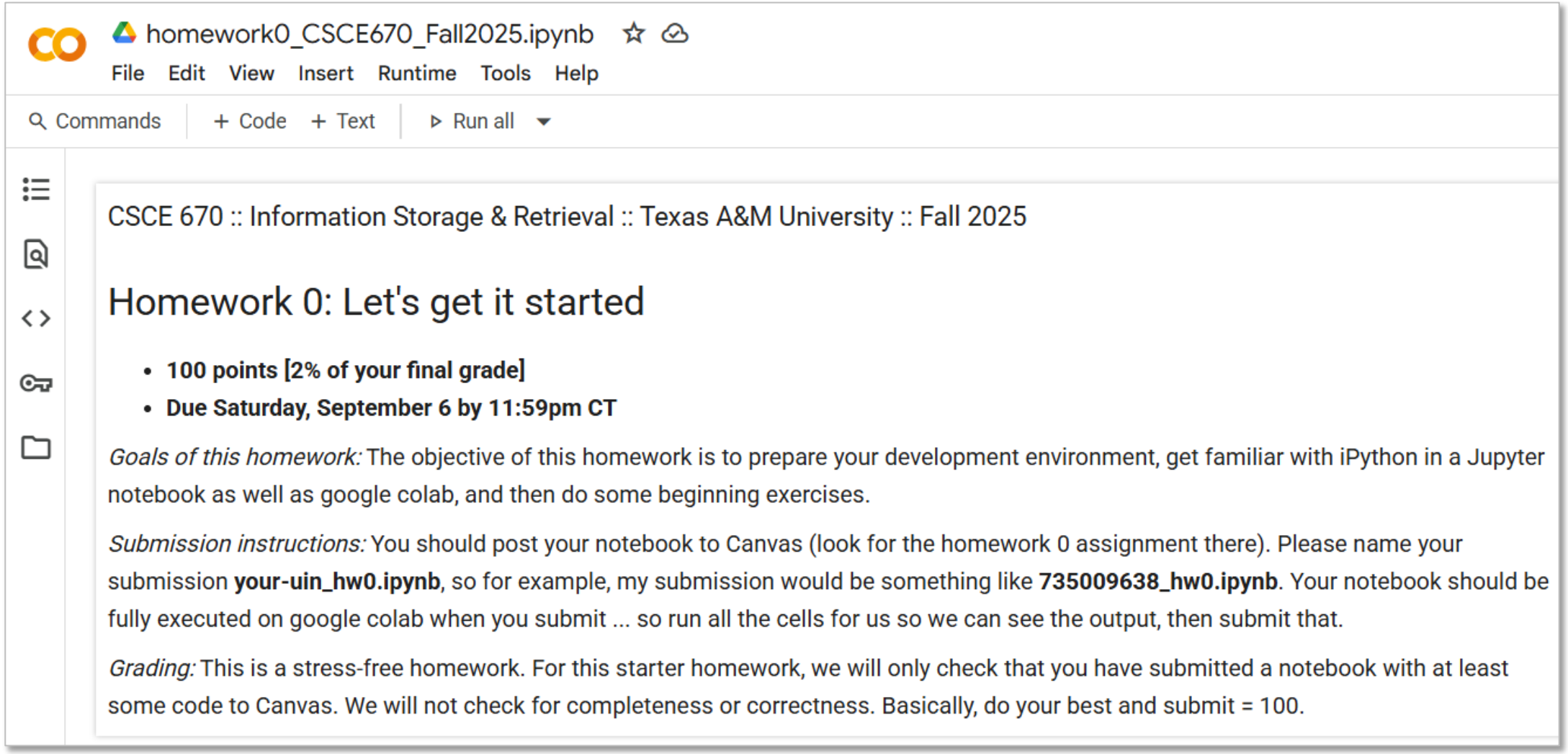
- Individual work, but you may discuss generally with others
  - You should write your own code, by yourself
  - BUT you may talk amongst yourselves about approaches/methods
    - E.g., sit in a group with no laptops, just talking = totally fine
    - E.g., sit next to each other while you code = BAD NEWS
  - You must acknowledge ALL help in your homework
    - Using code comments
    - I will show you an example in 5 minutes

AI Policy

- *"In principle you may submit AI-generated code, or code that is based on or derived from AI-generated code, as long as this use is properly documented in the comments: you need to include the prompt and the significant parts of the response. AI tools may help you avoid syntax errors, but there is no guarantee that the generated code is correct. It is your responsibility to identify errors in program logic through comprehensive, documented testing. Moreover, generated code, even if syntactically correct, may have significant scope for improvement, in particular regarding separation of concerns and avoiding repetitions. The submission itself must meet our standards of attribution and validation."*

- (from Boris Steipe (2023) "Syllabus Resources". The Sentient Syllabus Project: http://sentientsyllabus.org)

Homework 0 (due Sep 6)



Homework Late Days

- Due by 11:59pm on the due date
- You get 5 late days total
- Late day = indivisible 24-hour unit
  - E.g., if due date is 11:59pm on Saturday, and you submit at 12:01am on Sunday = one late day
  - No penalty for using a late day; no need to alert me/TA that you are using a late day
- Once you are out of late days, you get 0

Regrade Policy

- Once you receive your graded assignment (e.g., homework and quizzes), you have 7 days to request a regrade in writing (give to me)
- After 7 days = no regrades
- You must give us a written explanation of what the issue is
- We will re-grade the entire assignment

Questions?

4 Quizzes (5% × 4 = 20%)

- In-class
- 40 minutes, but designed to only take 25-30 minutes
- 7 multiple-choice questions
- Answering 5 questions correctly will earn you full credit (5%)

# correct answers	0	1	2	3	4	5	6	7
credit	0%	1%	2%	3%	4%	5%	5%	5%

- Closed book
  - Laptops, books, and notes are NOT allowed.
- Calculators are NOT required, and the questions will NOT involve calculations (such as square roots or logarithms) that cannot be done easily by hand.

Absence Policy

- Please refer to Student Rule 7 (https://student-rules.tamu.edu/rule07/) about excused absences, including definitions, and related documentation and timelines.
  - For students who miss a quiz due to an excused absence, your quiz score will be counted as part of the final exam.
  - Specifically, your final exam weight will increase by 5% for each quiz missed with an excused absence (i.e., 30% + 5% × number of excused quiz absences).

Final (30%)

- In our regular classroom
- 3:30pm – 5:30pm on Dec 16, 2025 (Tuesday)
- 120 minutes; Comprehensive
- You can bring one cheatsheet
  - Cheatsheet = 8.5” x 11” standard sheet of paper with anything on it, front and back

Group Project (20%)

- Teams of 3 or 4 (any deviation from this size requires prior approval from the instructor)
  - Option 1: A prototype (search engine or recommender system)
  - Option 2: A research project (e.g., reasoning-search interleaved LLMs)
  - Option 3: A survey (e.g., recent studies on search-enhanced LLMs)
- Topic-wise: your choice, as long as it is related to information retrieval!
- Project presentations during our last two classes
- Super-fun opportunity for you to explore some compelling aspect of IR

Group Project (20%)

- More details to be discussed in the Sep 25 class

W5	9/23	Learning to Rank		[MRS Chapter 14], [Nallapati, SIGIR'04], [Joachims, KDD'02]
	9/25	Learning to Rank (Cont'd) and Course Project Info		[Burges et al., ICML'05]

- Project proposal due on Oct 18 (so you still have plenty of time)

W8	10/14	No Class (Fall Break)		
	10/16	Word Embedding, word2vec, GloVe		[Mikolov et al., NIPS'13], [Pennington et al., EMNLP'14]
	10/18	Project Proposal Due (Saturday)		

- Project presentations during our last two classes

W15	12/2	Project Presentations (Zoom)
	12/4	Project Presentations (Zoom)

Zoom

- The following 3 classes will be held on Zoom.

W15	12/2	Project Presentations (Zoom)
	12/4	Project Presentations (Zoom)

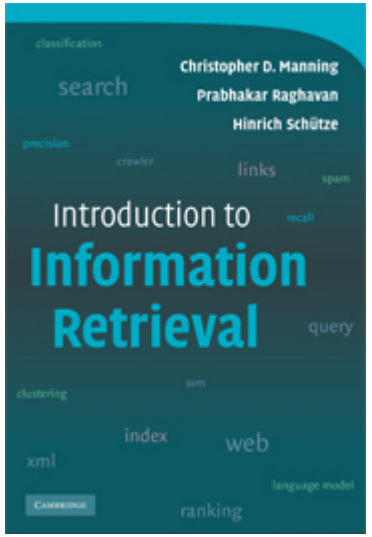
- ❌ To give you/myself a longer Thanksgiving break
- ✅ There will be 14-18 groups presenting in these 2 lectures. Zoom allows us to quickly switch between shared screens, reducing transition time between groups and giving each group more time to present.

W12	11/11	Large Language Models with Search Engines (Guest Lecture by Bowen Jin, **Zoom**)		[Jin et al., arXiv'25]

Questions?

Information Retrieval is …

- "… *finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).*"



(according to Manning, Raghavan, Schutze 2008)

Information Retrieval is …

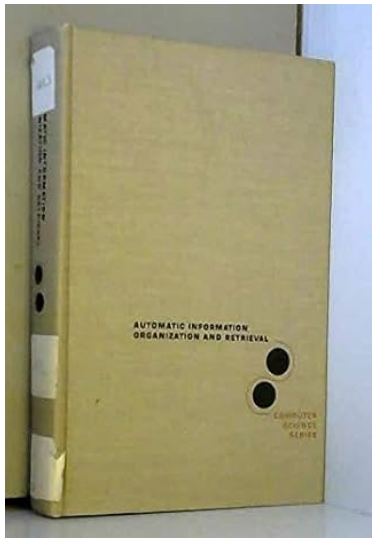
- "… *the process of obtaining information system resources that are relevant to an information need from a collection of those resources.*"

WIKIPEDIA
The Free Encyclopedia

(according to Wikipedia)

Information Retrieval is …

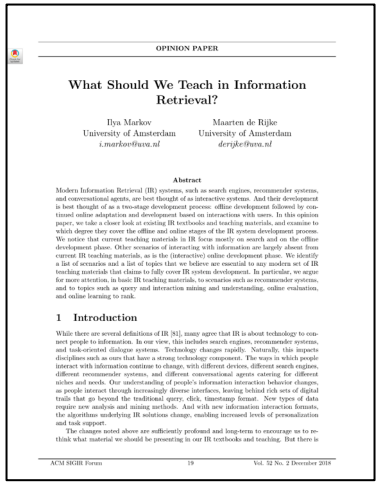
- "… *a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.*"



(according to Gerard Salton "Father of IR" 1968)

Information Retrieval is …

- "… *about technology to connect people to information.*"



OPINION PAPER

What Should We Teach in Information Retrieval?

Ilya Markov
University of Amsterdam
i.markov@uva.nl

Maarten de Rijke
University of Amsterdam
derijke@uva.nl

Abstract

Modern Information Retrieval (IR) systems, such as search engines, recommender systems, and conversational agents, are best thought of as interactive systems. And their development is best thought of as a two-stage development process: offline development followed by continued online adaptation and development based on interactions with users. In this opinion paper, we take a closer look at existing IR textbooks and teaching materials, and examine to which degree they cover the offline and online stages of the IR system development process. We notice that current teaching materials in IR focus mostly on search and on the offline development phase. Other scenarios of interacting with information are largely absent from current IR teaching materials, as is the (interactive) online development phase. We identify a list of scenarios and a list of topics that we believe are essential to any modern set of IR teaching materials that claims to fully cover IR system development. In particular, we argue for more attention, in basic IR teaching materials, to scenarios such as recommender systems, and to topics such as query and interaction mining and understanding, online evaluation, and online learning to rank.

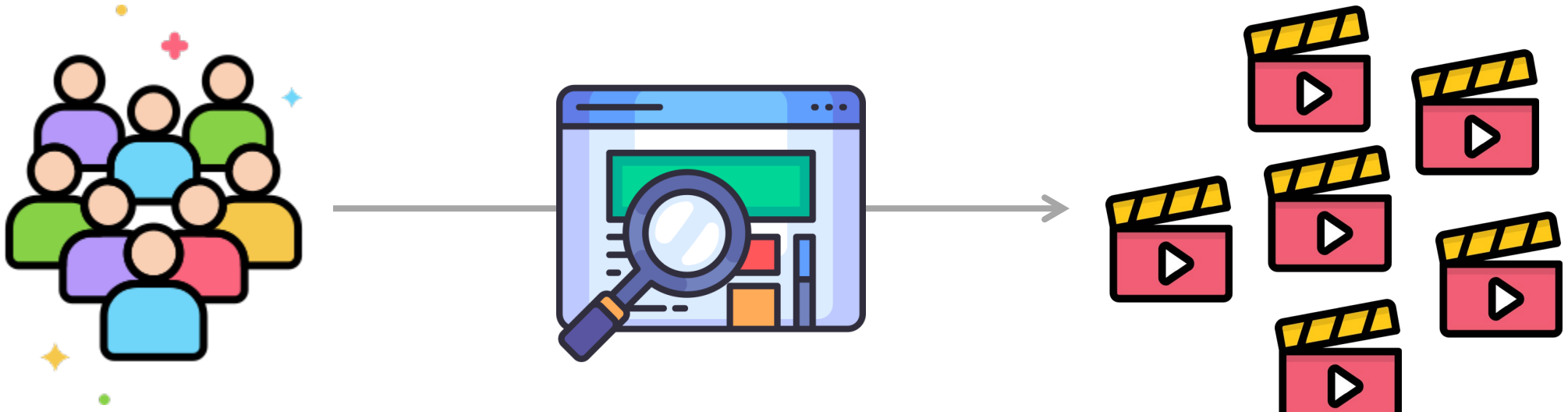
1 Introduction

While there are several definitions of IR [81], many agree that IR is about technology to connect people to information. In our view, this includes search engines, recommender systems, and task-oriented dialogue systems. Technology changes rapidly. Naturally, this impacts disciplines such as ours that have a strong technology component. The ways in which people interact with information continue to change, with different devices, different search engines, different recommender systems, and different conversational agents catering for different niches and needs. Our understanding of people's information interaction behavior changes, as people interact through increasingly diverse interfaces, leaving behind rich sets of digital trails that go beyond the traditional query, click, timestamp format. New types of data require new analysis and mining methods. And with new information interaction formats, the algorithms underlying IR solutions change, enabling increased levels of personalization and task support.

The changes noted above are sufficiently profound and long-term to encourage us to rethink what material we should be presenting in our IR textbooks and teaching. But there is

(according to Markov and de Rijke 2018)

IR connects people to information



People

IR System

Information

- Examples?

Example: Course Explorer

Students

Boolean Retrieval (Week 1)

Courses about "*learning*"

Title	CRN Syllab...	S.	Crse	Sect	Hrs	Instructor(s)
learning						
MACHINE LEARNING	57951 Syllabus	CSCE	633	700	3	Bobak Mortazavi (P)
DEEP LEARNING	36429 Syllabus	CSCE	636	600	3	Anxiao Jiang (P)
DEEP LEARNING	62232 Syllabus	CSCE	636	700	3	Anxiao Jiang (P)
DEEP REINFORCEMENT LEARNING	55177 Syllabus	CSCE	642	600	3	Guni Sharon (P)
DEEP REINFORCEMENT LEARNING	60328 Syllabus	CSCE	642	700	3	Guni Sharon (P)
SPTP: DEEP LEARNING AND LLMS	62415 Syllabus	CSCE	689	600	3	Tomer Joseph Galanti (P)

Example: PubMed, Semantic Scholar, Google Scholar

Researchers

TF-IDF, BM25 (Week 2)

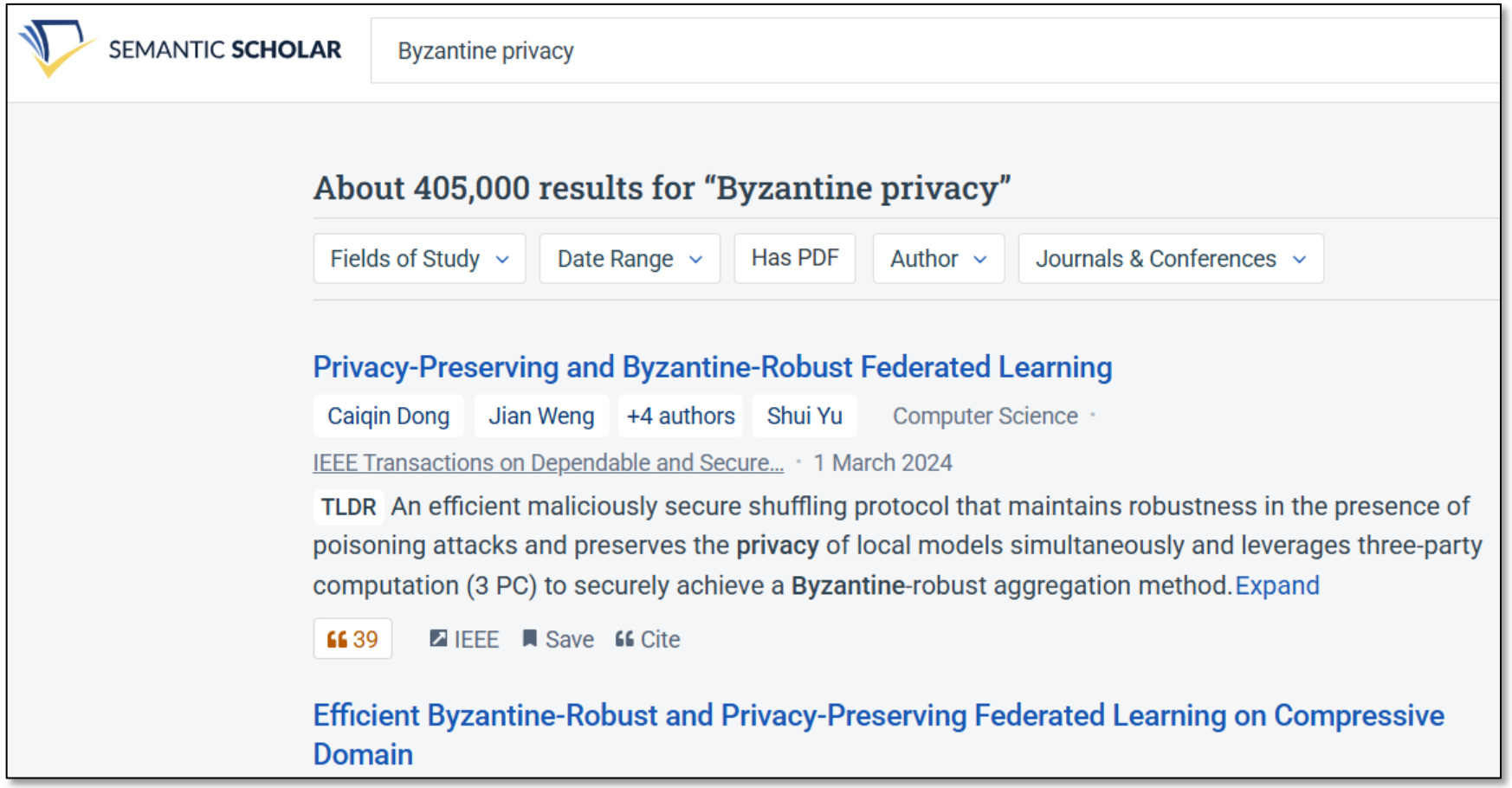
Papers about "*respiratory disease*"

Example: PubMed, Semantic Scholar, Google Scholar

Researchers

word2vec, BERT (Weeks 8-10)

Papers about "*Byzantine privacy*"

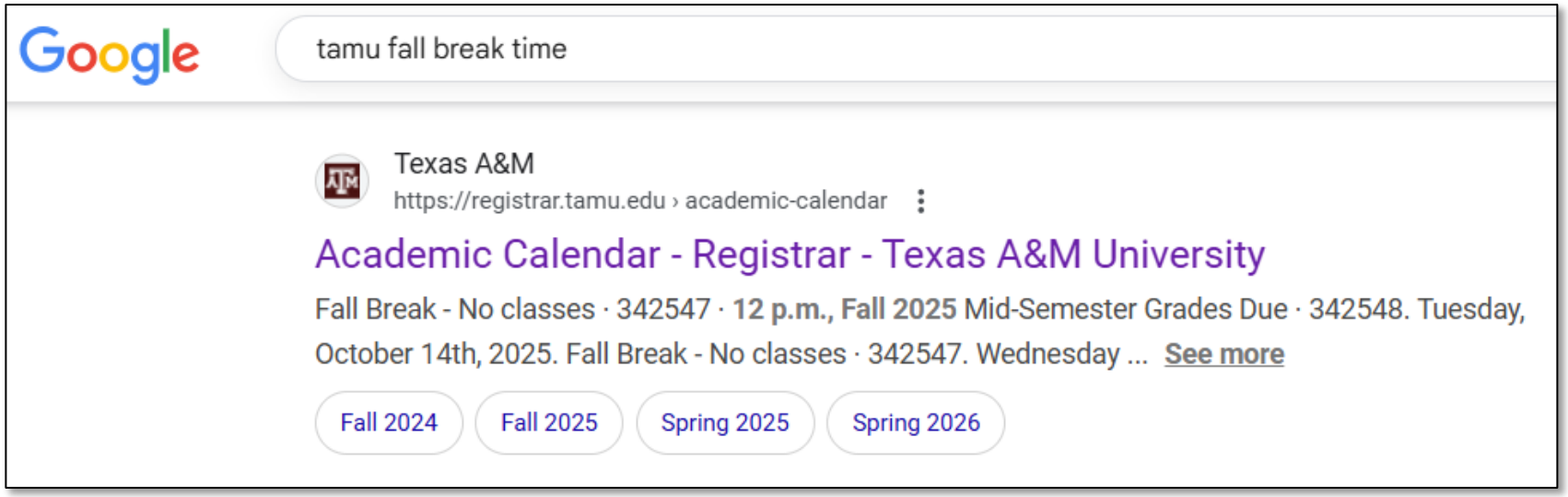


Example: Google

Aggies

PageRank (Week 3)

Webpages about "*tamu fall break time*"



Example: ChatGPT

IR Beginners

LLMs for Search (Weeks 12, 13)

Most cited paper in IR

What is the most-cited paper in IR? Please provide sources.

You're asking: **"What is the most-cited paper in Information Retrieval (IR)? Please provide sources."**

Most-Cited IR Paper

The **classic and arguably most-cited foundational paper in Information Retrieval** is:

"A Vector Space Model for Automatic Indexing"

- Authors: **Gerard Salton, A. Wong, and C. S. Yang**
- Published in *Communications of the ACM*, November 1975.
- This landmark paper introduced the **vector space model (VSM)** and the use of **TF–IDF weighting**, both of which are cornerstones of modern IR systems Wikipedia +1 .

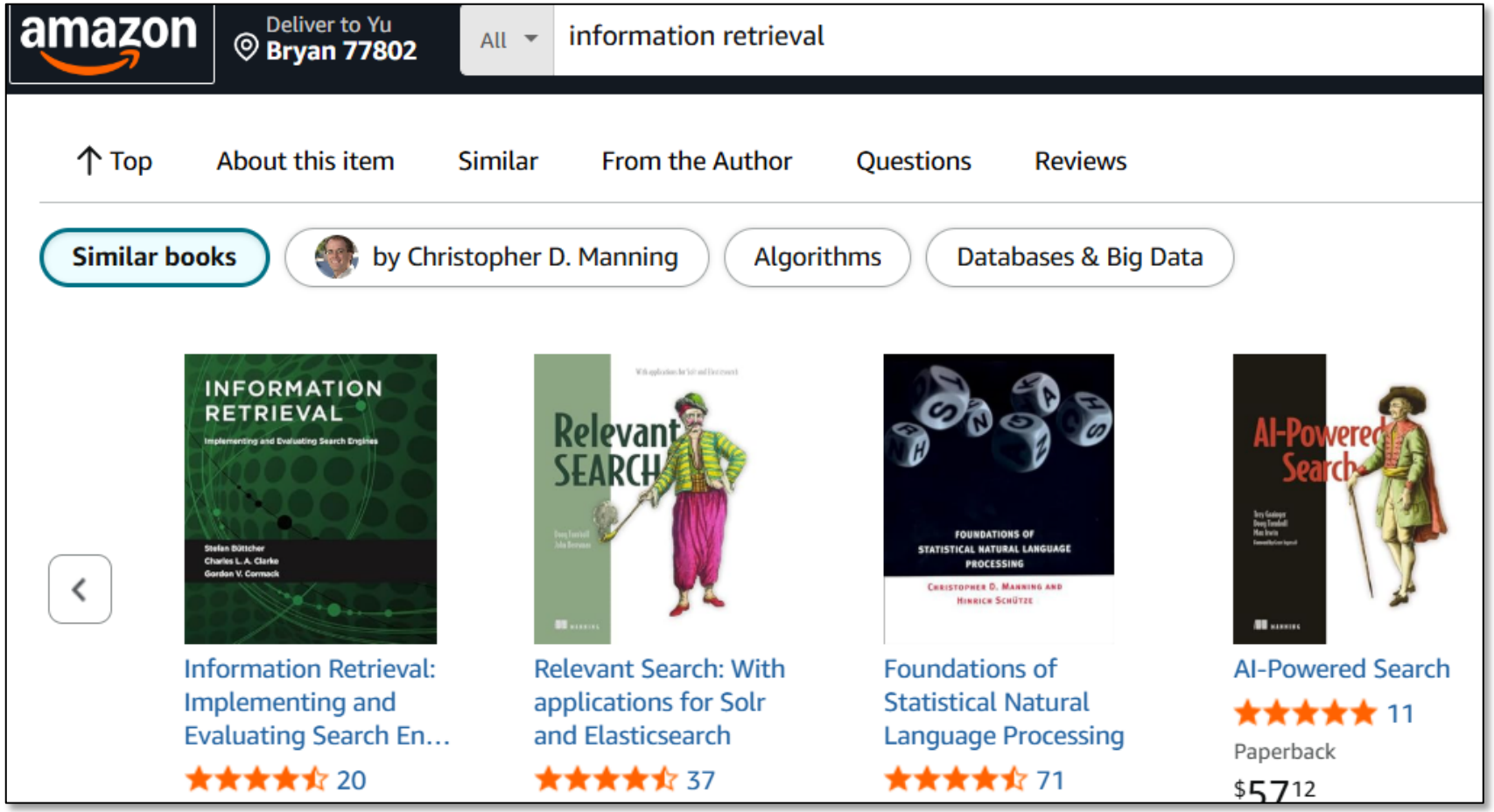
Gerard Salton, often called the "**father of Information Retrieval**," developed this model at Cornell and significantly shaped IR theory and practice Wikipedia .

Example: Amazon

Customers who bought an IR book

Recommender Systems (Weeks 6, 7, 10)

Other books the customers may also like

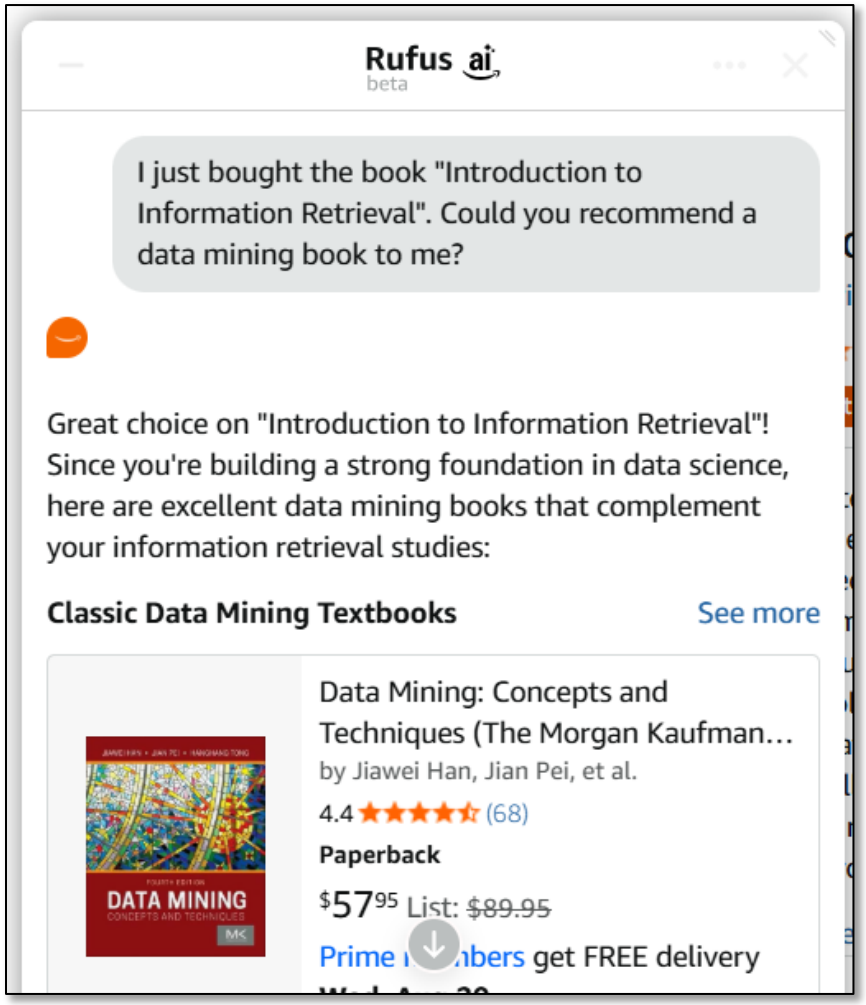


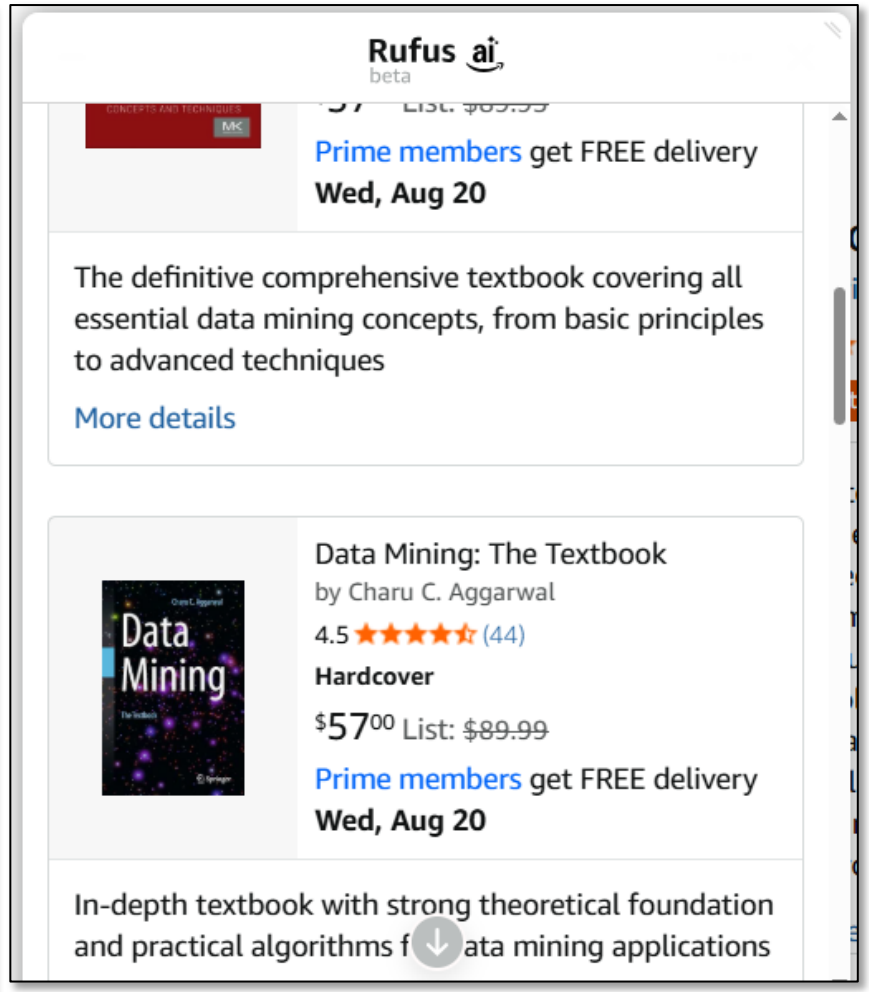
Example: Amazon

Customers who bought an IR book

LLMs for RecSys (Week 13)

Other data mining books the customers may also like



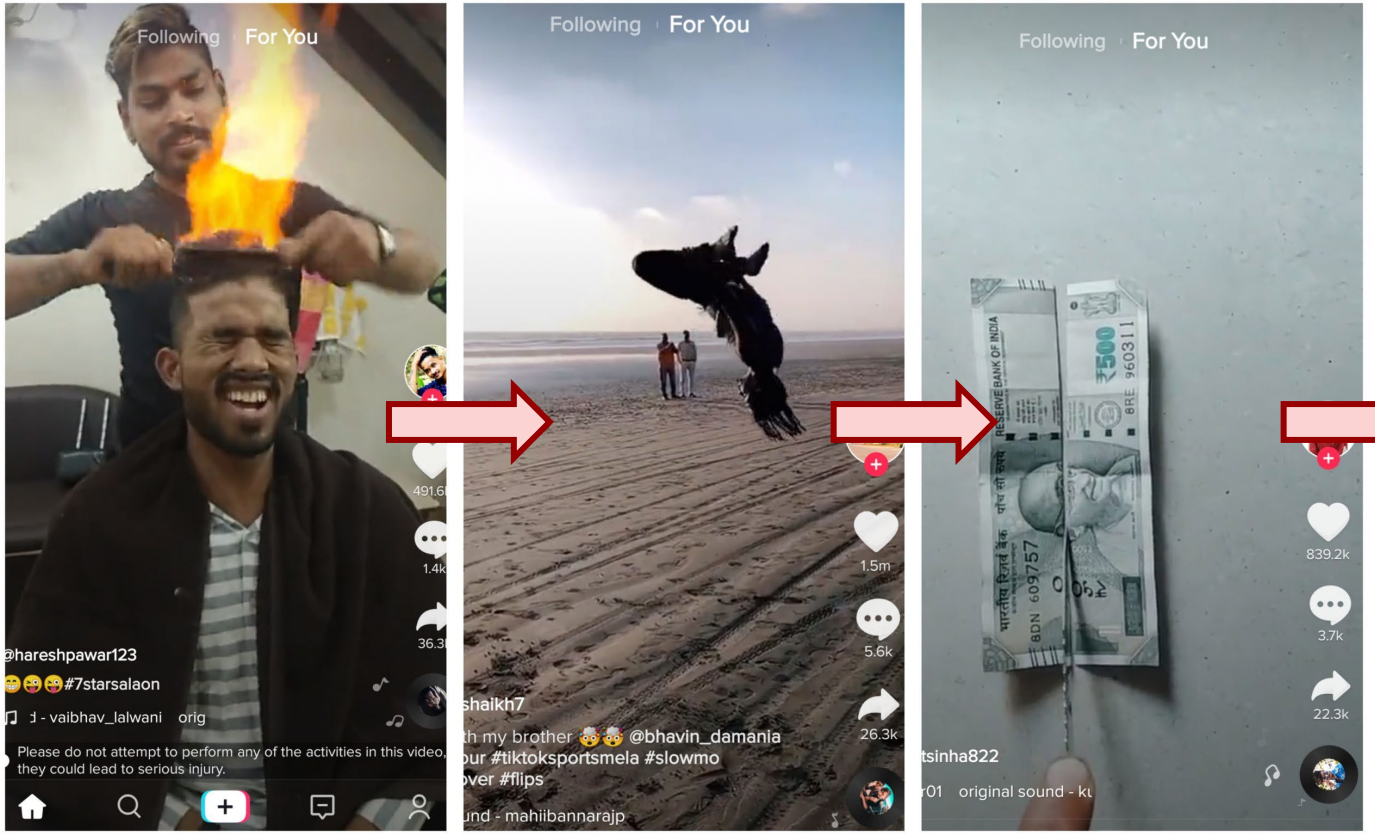


Example: TikTok

Video scrollers

Sequential RecSys (Week 11)

Next video they may be interested in



IR algorithmically mediates what items a user encounters

YouTube's Product Chief on Online Radicalization and Algorithmic Rabbit Holes

Neal Mohan discusses the streaming site's recommendation engine, which has become a growing liability amid accusations that it steers users to increasingly extreme content.

We as computer scientists need to …

- Understand these algorithms
  - How can we build a search engine or a recommender system? What algorithms can we use? What "features" are important? How to evaluate if it is doing a good job?
- Design new approaches
  - Can we keep pace with rapid developments in industry and in academia? Adopt new ML/DL approaches? Anticipate the "next" big thing?
- Be mindful of the power we wield! Important issues around fairness, bias, misinformation, and other negative outcomes.

This Course

- Phase 1: Search Engines
  - basics, Boolean and ranked retrieval, link analysis, evaluation, learning to rank (ML + ranking), …
- Phase 2: Recommender Systems
  - basics, non-personalized recommendation, collaborative filtering, matrix factorization, implicit recommendation, …
- Phase 3: From Foundations to Modern Methods
  - embedding learning, Transformer, "small" language models, … (for search and recommendation)
- Phase 4: Large Language Models (!!)

Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE670-F25.html