



# CSCE 670 - Information Storage and Retrieval

## Lecture 1: Welcome and Course Overview

Yu Zhang

[yuzhang@tamu.edu](mailto:yuzhang@tamu.edu)

August 26, 2025



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

# 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	<a href="#">PDF</a>	-
	8/28	Boolean Retrieval	<a href="#">PDF</a>	<a href="#">[MRS Chapter 1]</a> , <a href="#">[MRS Chapter 2]</a>
W2	9/2	TF-IDF, Vector Space Model	<a href="#">PDF</a>	<a href="#">[MRS Chapter 6]</a> , <a href="#">[MRS Chapter 7]</a>
	9/4	BM25, Probabilistic Model	<a href="#">PDF</a>	<a href="#">[MRS Chapter 11]</a>
	9/6	<b>Homework 0 Due (Saturday)</b>		
W3	9/9	Link Analysis: PageRank		<a href="#">[MRS Chapter 21]</a> , <a href="#">[LRU Chapter 5.1/5.2]</a>
	9/11	Link Analysis: Topic-Sensitive PageRank, HITS		<a href="#">[LRU Chapter 5.3/5.5]</a>
W4	9/16	Evaluation		<a href="#">[MRS Chapter 8]</a>
	9/18	Evaluation (Cont'd) <b>and Quiz 1</b>		<a href="#">[Sakai, SIGIR'16]</a>

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

The screenshot shows the Canvas LMS interface for a course. On the left is a dark red sidebar with the TAMU logo at the top, followed by icons and labels for Account, Dashboard, Courses, Calendar, Inbox, History, and Help. The main content area has a top navigation bar with 'CS, Fall, 2025' and a list of links: Home, Grades, Course Evaluations, Assignments, Syllabus, and Files. The 'Home' link is active. The main heading is 'Recent Announcements'. Below it is the announcement title '25 FALL CSCE 670 600: INFO STORAGE & RETRIEVAL' with an up/down arrow icon. The announcement body starts with 'Welcome!' followed by a paragraph: 'Welcome to the course! I encourage you to read the syllabus and explore the resources in this Canvas course. Please contact me if you have any questions.' Below the paragraph is a small square icon. At the bottom of the announcement is a grey bar titled 'Student Resources'. Under this bar are five buttons, each with a small icon and a text label: 'Syllabus Student Resource Button', 'Office Hours Student Resource Button', 'University Resources Student Resource Button', 'Canvas Resources Student Resource Button', and 'Mental Health Student Resource Button'.

CS, Fall, 2025

Home

Grades

Course Evaluations

Assignments

Syllabus

Files

## Recent Announcements

### 25 FALL CSCE 670 600: INFO STORAGE & RETRIEVAL

#### Welcome!

Welcome to the course! I encourage you to read the syllabus and explore the resources in this Canvas course. Please contact me if you have any questions.

#### Student Resources

Syllabus Student Resource Button Office Hours Student Resource Button

University Resources Student Resource Button Canvas Resources Student Resource Button

Mental Health Student Resource Button

# 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]

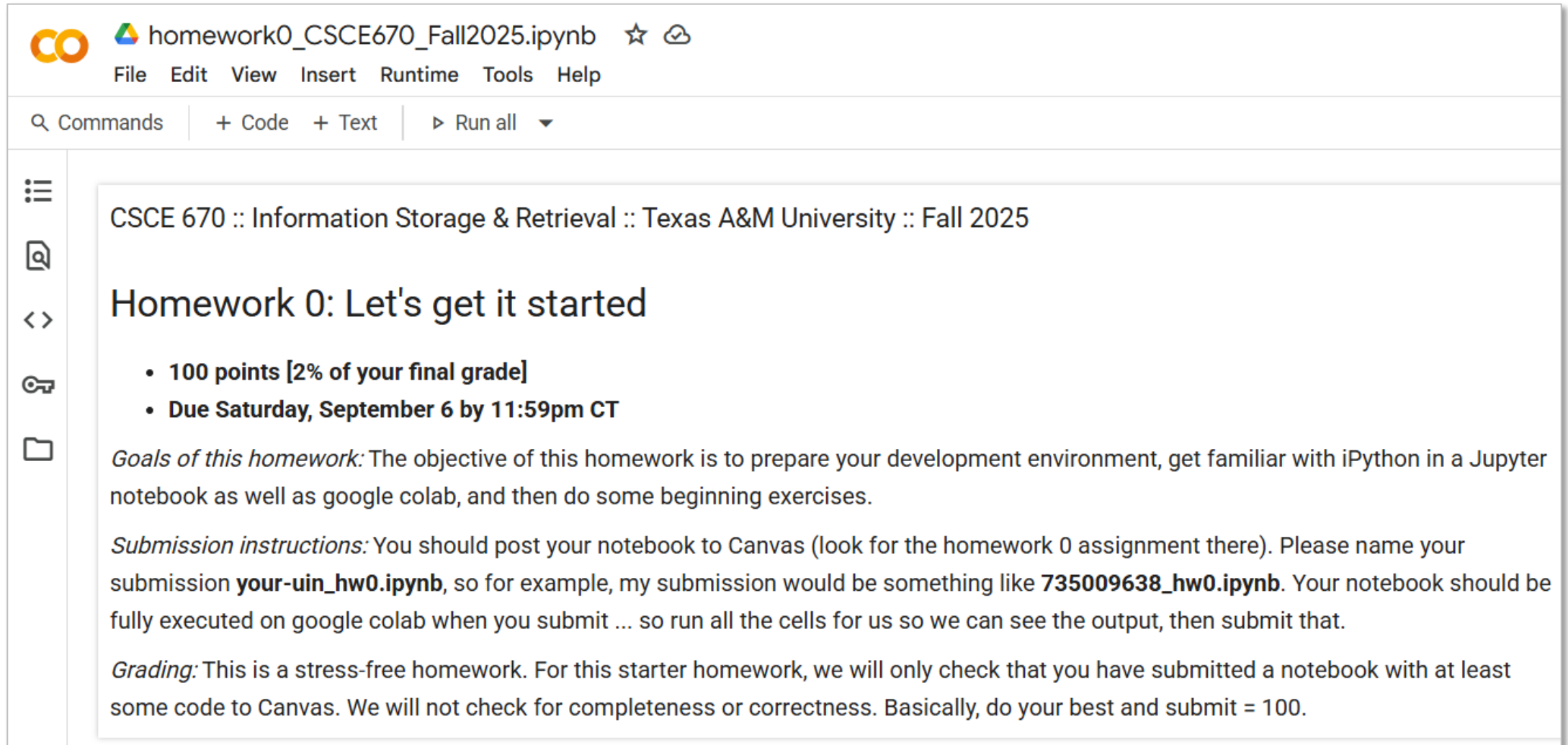
# Homework (30%)

- 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)



The screenshot shows a Jupyter Notebook titled "homework0\_CSCE670\_Fall2025.ipynb". The interface includes a top menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". Below the menu is a toolbar with "Commands", "+ Code", "+ Text", and "Run all". On the left is a sidebar with icons for a menu, search, code editor, key, and folder. The main content area displays the following text:

CSCE 670 :: Information Storage & Retrieval :: Texas A&M University :: Fall 2025

## Homework 0: Let's get it started

- **100 points [2% of your final grade]**
- **Due Saturday, September 6 by 11:59pm CT**

*Goals of this homework:* The objective of this homework is to prepare your development environment, get familiar with iPython in a Jupyter notebook as well as google colab, and then do some beginning exercises.

*Submission instructions:* You should post your notebook to Canvas (look for the homework 0 assignment there). Please name your submission **your-uid\_hw0.ipynb**, so for example, my submission would be something like **735009638\_hw0.ipynb**. Your notebook should be fully executed on google colab when you submit ... so run all the cells for us so we can see the output, then submit that.

*Grading:* This is a stress-free homework. For this starter homework, we will only check that you have submitted a notebook with at least some code to Canvas. We will not check for completeness or correctness. Basically, do your best and submit = 100.



# 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 ( $5\% \times 4 = 20\%$ )

- 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.,  $30\% + 5\% \times \text{number of excused quiz absences}$ ).

# 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)

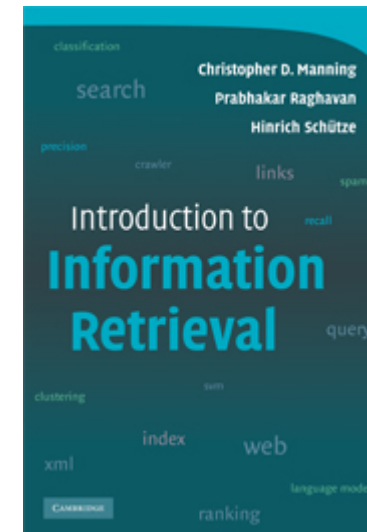
-  To give you/myself a longer Thanksgiving break
-  There will be 14-18 groups presenting in these 2 lectures. Zoom allows us to quickly switch between shared screens, reducing transition time between groups and giving each group more time to present.

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.”

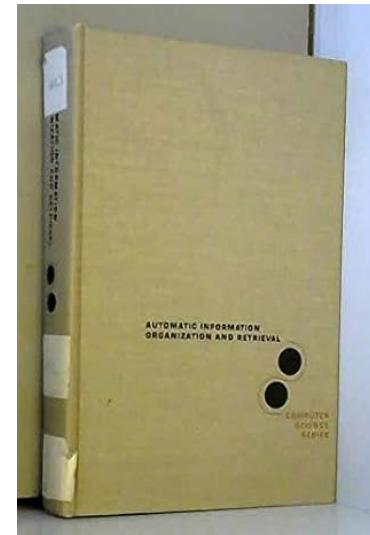


WIKIPEDIA  
The Free Encyclopedia

(according to Wikipedia)

# Information Retrieval is ...

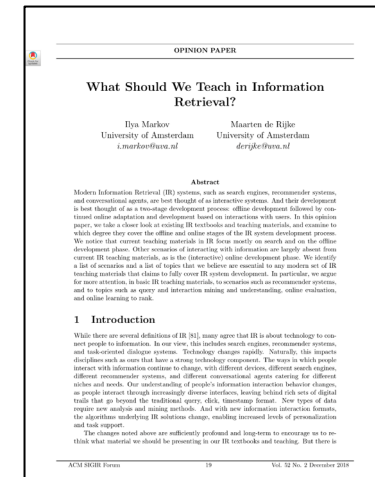
- “... a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”



(according to Gerard Salton “Father of IR” 1968)

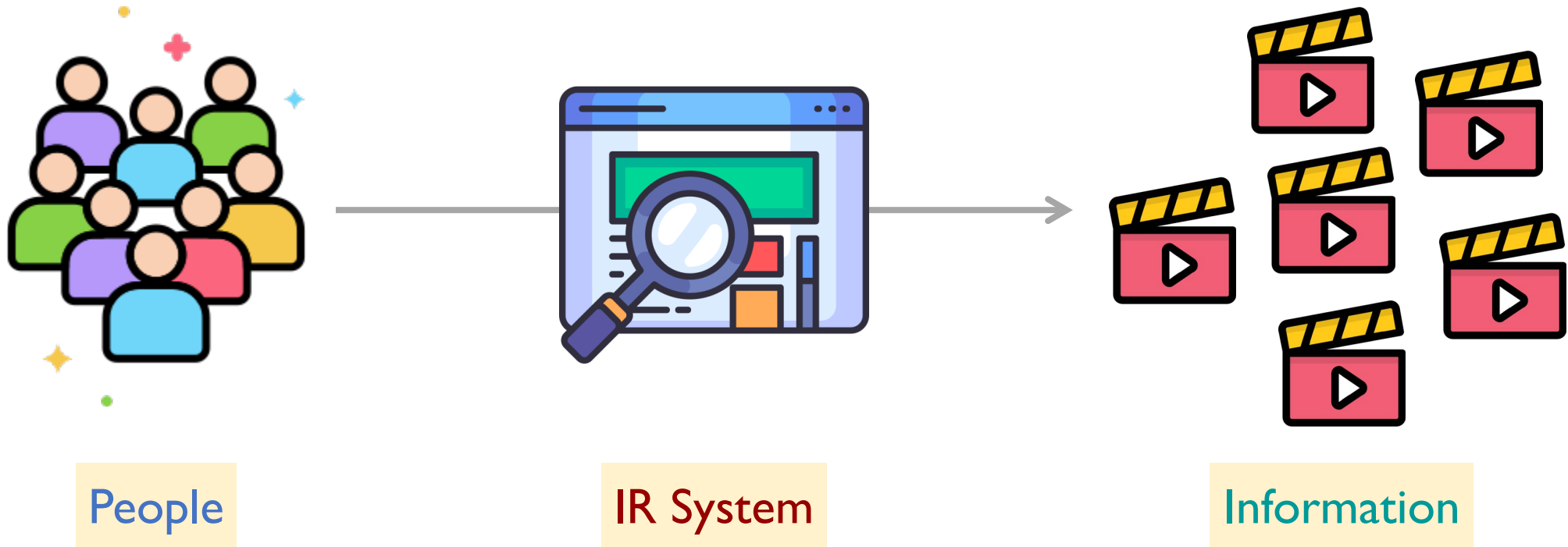
# Information Retrieval is ...

- “... *about technology to connect people to information.*”



(according to Markov and de Rijke 2018)

# IR connects people to information



- Examples?

# Example: Course Explorer



Students



Boolean  
Retrieval  
(Week 1)



Courses  
about  
“learning”

Title ▾	CRN Syllab... ▾	S. ▾	Crse ▾	Sect ▾	Hrs ▾	Instructor(s) ▾
<input type="text" value="learning"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<a href="#">MACHINE LEARNING</a>	57951 <a href="#">Syllabus</a>	CSCE	633	700	3	Bobak Mortazavi (P)
<a href="#">DEEP LEARNING</a>	36429 <a href="#">Syllabus</a>	CSCE	636	600	3	Anxiao Jiang (P)
<a href="#">DEEP LEARNING</a>	62232 <a href="#">Syllabus</a>	CSCE	636	700	3	Anxiao Jiang (P)
<a href="#">DEEP REINFORCEMENT LEARNING</a>	55177 <a href="#">Syllabus</a>	CSCE	642	600	3	Guni Sharon (P)
<a href="#">DEEP REINFORCEMENT LEARNING</a>	60328 <a href="#">Syllabus</a>	CSCE	642	700	3	Guni Sharon (P)
<a href="#">SPTP: DEEP LEARNING AND LLMS</a>	62415 <a href="#">Syllabus</a>	CSCE	689	600	3	Tomer Joseph Galanti (P)



# Example: PubMed, Semantic Scholar, Google Scholar



Researchers



TF-IDF,  
BM25  
(Week 2)



Papers about  
“respiratory  
disease”

NIH National Library of Medicine  
National Center for Biotechnology Information

Log in

PubMed®

respiratory disease

Search

Advanced Create alert Create RSS User Guide

Save Email Send to Sort by: Best match Display options

MY CUSTOM FILTERS

2,024,858 results

RESULTS BY YEAR

1846 2026

Global Impact of **Respiratory Disease**: What Can We Do, Together, to Make a Difference?

1

Cite Levine SM, Marciniuk DD.  
Chest. 2022 May;161(5):1153-1154. doi: 10.1016/j.chest.2022.01.014. Epub 2022 Jan 17.  
PMID: 35051424 **Free PMC article.** No abstract available.

Lessons from Dairy Farmers for Occupational Allergy and **Respiratory Disease**.

# Example: PubMed, Semantic Scholar, Google Scholar



Researchers



word2vec,  
BERT  
(Weeks 8-  
10)



Papers about  
“Byzantine  
privacy”

SEMANTIC SCHOLAR

Byzantine privacy

About 405,000 results for “Byzantine privacy”

Fields of Study ▾ Date Range ▾ Has PDF Author ▾ Journals & Conferences ▾

**Privacy-Preserving and Byzantine-Robust Federated Learning**

Caiqin Dong Jian Weng +4 authors Shui Yu Computer Science ·

IEEE Transactions on Dependable and Secure... · 1 March 2024

**TLDR** An efficient maliciously secure shuffling protocol that maintains robustness in the presence of poisoning attacks and preserves the **privacy** of local models simultaneously and leverages three-party computation (3 PC) to securely achieve a **Byzantine-robust** aggregation method.[Expand](#)

39 IEEE Save Cite

**Efficient Byzantine-Robust and Privacy-Preserving Federated Learning on Compressive Domain**

# Example: Google



Aggies




PageRank  
(Week 3)



Webpages  
about “tamu  
fall break  
time”

Google

tamu fall break time

 Texas A&M  
<https://registrar.tamu.edu/academic-calendar>

**Academic Calendar - Registrar - Texas A&M University**

Fall Break - No classes · 342547 · 12 p.m., Fall 2025 Mid-Semester Grades Due · 342548. Tuesday, October 14th, 2025. Fall Break - No classes · 342547. Wednesday ... [See more](#)

Fall 2024   Fall 2025   Spring 2025   Spring 2026

# 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

The screenshot shows the Amazon website interface. At the top, the Amazon logo is on the left, and the delivery location 'Deliver to Yu Bryan 77802' is in the center. A search bar on the right contains the text 'information retrieval'. Below the search bar, navigation links include 'Top', 'About this item', 'Similar', 'From the Author', 'Questions', and 'Reviews'. A section titled 'Similar books' features a row of four book covers:

- Information Retrieval: Implementing and Evaluating Search Engines** by Stefan Bittcher, Charles L.A. Clarke, and Gordon V. Cormack. It has a 4.5-star rating (20 reviews).
- Relevant Search: With applications for Solr and Elasticsearch** by Doug Ford and John Deveney. It has a 4.5-star rating (37 reviews).
- Foundations of Statistical Natural Language Processing** by Christopher D. Manning and Hinrich Schütze. It has a 4.5-star rating (71 reviews).
- AI-Powered Search** by Terry Golubov, Doug Ford, and Peter Hume. It has a 4.5-star rating (11 reviews), is a paperback, and priced at \$57.12.

# Example: Amazon



Customers  
who bought an  
IR book



LLMs for  
RecSys  
(Week 13)



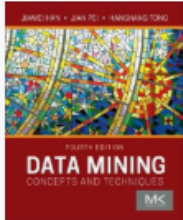
Other data  
mining books  
the customers  
may also like

Rufus ai  
beta

I just bought the book "Introduction to Information Retrieval". Could you recommend a data mining book to me?

Great choice on "Introduction to Information Retrieval"! Since you're building a strong foundation in data science, here are excellent data mining books that complement your information retrieval studies:


**Classic Data Mining Textbooks** [See more](#)

 **Data Mining: Concepts and Techniques (The Morgan Kaufman...)**  
by Jiawei Han, Jian Pei, et al.  
4.4 ★★★★★ (68)  
Paperback  
\$57<sup>95</sup> List: \$89.95  
[Prime members](#) get FREE delivery

Rufus ai  
beta

[Prime members](#) get FREE delivery  
Wed, Aug 20

The definitive comprehensive textbook covering all essential data mining concepts, from basic principles to advanced techniques  
[More details](#)

 **Data Mining: The Textbook**  
by Charu C. Aggarwal  
4.5 ★★★★★ (44)  
Hardcover  
\$57<sup>00</sup> List: \$89.99  
[Prime members](#) get FREE delivery  
Wed, Aug 20

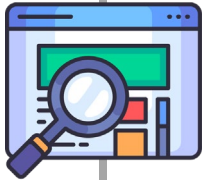
In-depth textbook with strong theoretical foundation and practical algorithms for data mining applications



# Example: TikTok



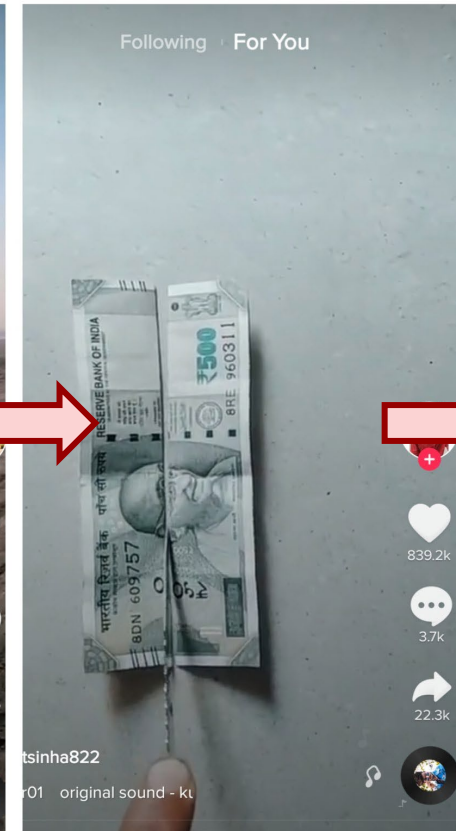
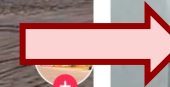
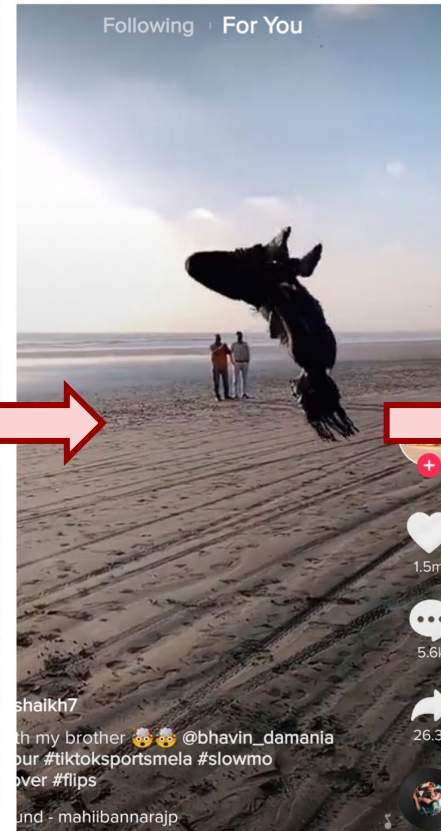
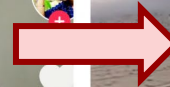
Video scrollers



Sequential  
RecSys  
(Week II)



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>