SITARE University

Chapter 1

Search Engines and Information Retrieval

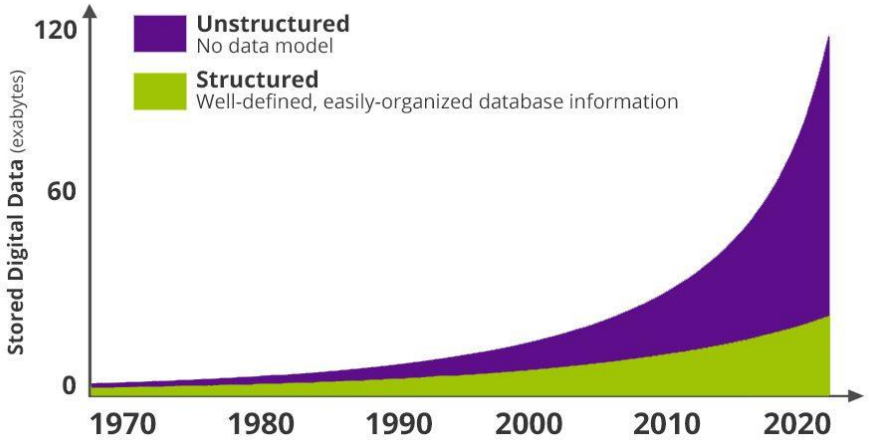
Full Credit: Croft et al. - http://www.search-engines-book.com/

In Memory: Gerard Salton (1927-1995)

Cornell Upson Hall (circa 1994)

Gerard Slaton
Amit Singhal
Chris Buckley
Cindy Robinson
Mandar Mitra

Why This Course?



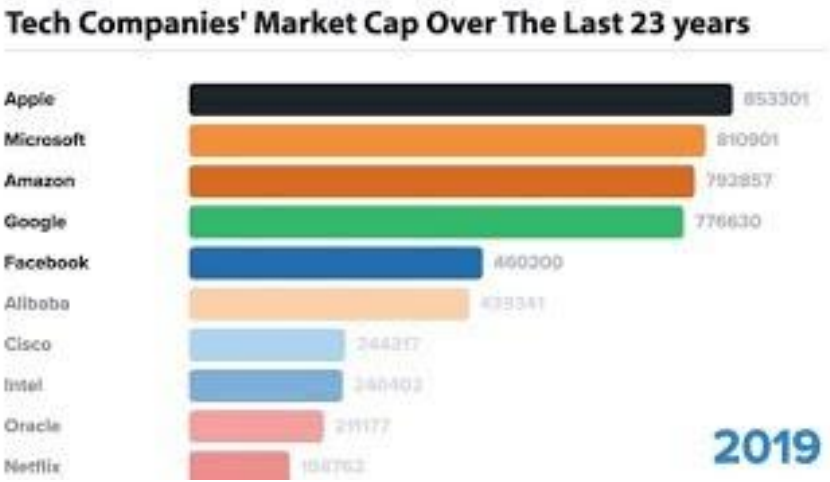


Image Source: Komprise

Search and Information Retrieval

- Search on the Web[1] is a daily activity for many people throughout the world
- Search and communication are most popular uses of the computer
- Applications involving search are everywhere
- The field of computer science that is most involved with R&D for search is *information retrieval (IR)*

[1] or is it web?

Amit's Sidebar

- ICE Metaphor
  - **Information**: news sites, search engines, X (previously twitter)
  - **Connection/Communication**: WhatsApp, Email, SMS, FB
  - **Entertainment**: YouTube, Instagram
- However, parts of ICE show up in various services
  - FB, IG, and X have DM
  - YouTube has a lot of informational content

Information Retrieval

- *“Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.”* (Salton, 1968)
- General definition that can be applied to many types of information and search applications
- Primary focus of IR since the 50s has been on *text* and *documents*

What is a Document?

- Examples:
  - web pages, email, books (or a chapter, or a page?), news stories, scholarly papers, text messages, Word™, Powerpoint™, PDF, forum postings, patents, IM sessions, etc.
- Common properties
  - Significant text content
  - Some structure (e.g., title, author, date for papers; subject, sender, destination for email)

Documents vs. Database Records

- Database records (or *tuples* in relational databases) are typically made up of well-defined fields (or *attributes*)
  - e.g., bank records with account numbers, balances, names, addresses, social security numbers, dates of birth, etc.
- Easy to compare fields with well-defined semantics to queries in order to find matches
- Text is more difficult

Documents vs. Records

- Example bank database query
  - *Find records with balance > $50,000 in branches located in Paris.*
  - Matches easily found by comparison with field values of records
- Example search engine query
  - [bank scams]
  - This text must be compared to the text of entire news stories

Comparing Text

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval
- Exact matching of words is not enough
  - Many different ways to write the same thing in a “natural language” like English
  - e.g., does a news story containing the text *“bank manager in Italy steals funds”* match the query [bank scams]?
  - Some stories will be better matches than others

Dimensions of IR

- IR is more than just text, and more than just web search
  - although these are central
- People doing IR work with different media, different types of search applications, and different tasks

Other Media

- New applications increasingly involve new media
  - e.g., video, photos, music, speech
- Like text, content is difficult to describe and compare
  - text may be used to represent them (e.g. tags, comments, speech recognition)
- IR approaches to search and evaluation are appropriate

Dimensions of IR

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

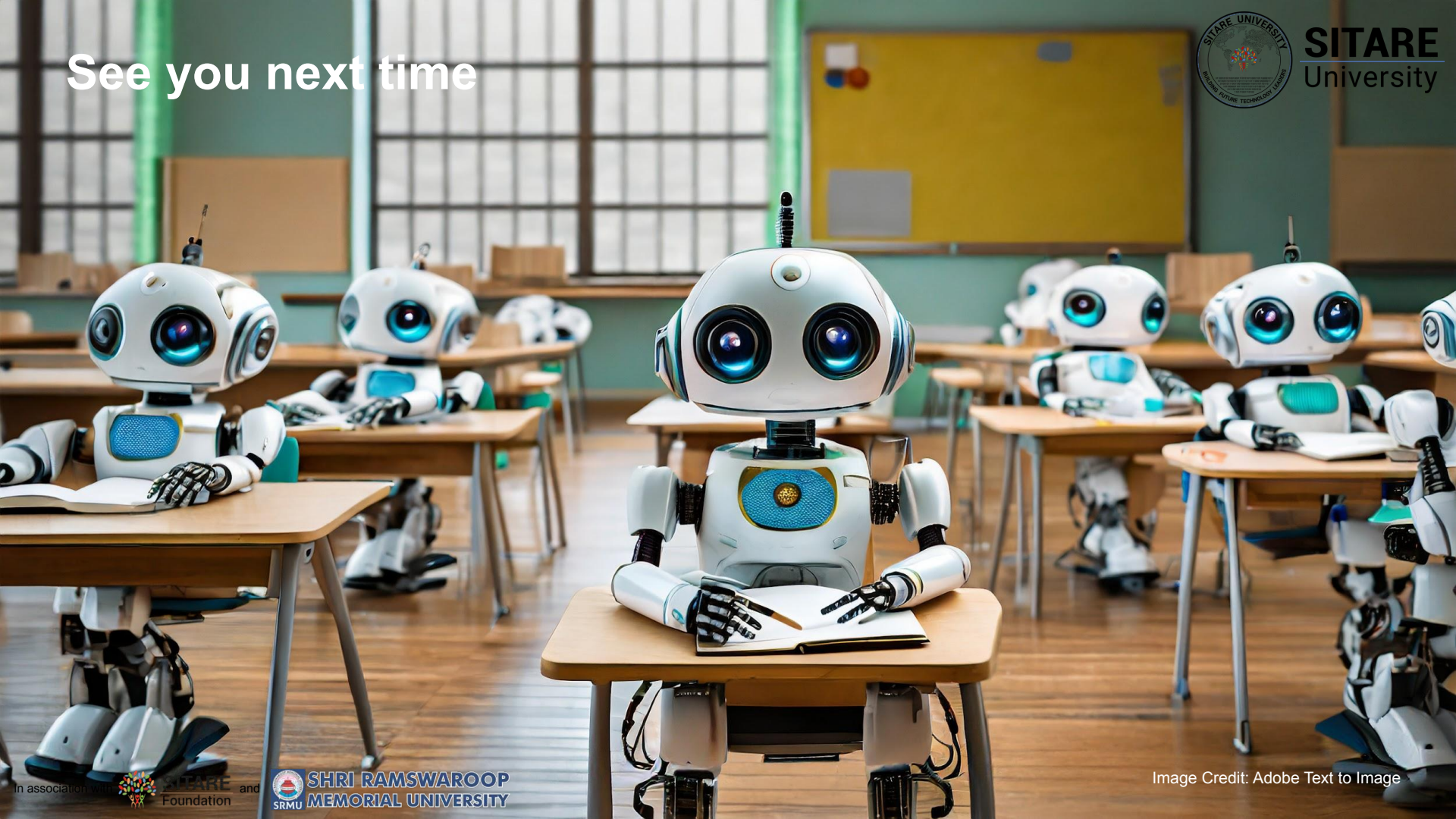
IR Tasks

- Ad-hoc search
  - Find relevant documents for an arbitrary text query
- Filtering
  - Identify relevant user profiles for a new document
- Classification
  - Identify relevant labels for documents
- Question answering
  - Give a specific answer to a question

Big Issues in IR

- Relevance
  - What is it?
  - Simple (and simplistic) definition: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
  - Many factors influence a person's decision about what is relevant: e.g., task, context, novelty, style
  - *Topical relevance* (same topic) vs. *user relevance* (everything else)
  - Topical relevance ([query words]) AND user relevance (factors not in the query words)

See you next time



SITARE University

In association with SITARE Foundation and SHRI RAMSWAROOP MEMORIAL UNIVERSITY

Image Credit: Adobe Text to Image

Big Issues in IR

- Relevance
  - *Retrieval models* define a view of relevance
  - *Ranking algorithms* used in search engines are based on retrieval models
  - Most models describe statistical properties of text rather than linguistic
    - i.e. counting simple text features such as words instead of parsing and analyzing the sentences
    - Statistical approach to text processing started with Luhn in the 50s
    - Linguistic features can be part of a statistical model

Big Issues in IR

- Evaluation
  - Experimental procedures and measures for comparing system output with user expectations
    - Originated in Cranfield experiments in the 60s
  - IR evaluation methods now used in many fields
  - Typically use *test collection* of documents, queries, and relevance judgments
    - Most commonly used are TREC collections
  - *Recall* and *precision* are two examples of <u>effectiveness</u> measures

Amit's Sidebar

- Which is better?

Google

computer science universities in india

U.S. News & World Report
https://www.usnews.com › ... › India

Best Global Universities for Computer Science in India

Here are the best global **universities** for **computer science** in **India** · Thapar **Institute** of Engineering & Technology · Vellore **Institute** of Technology · **Indian** ...

People also ask

- Which institute is best for Computer Science in India?
- Is India a good place to study Computer Science?
- What is the average fee for CS in India?
- Which IIT is best for Computer Science?

Feedback

SCImago Institutions Rankings
https://www.scimagoir.com › rankings

University Overall Rankings - Computer Science - India 2023

1 (287), VIT **University** ; 2 (522), Amrita **University** ; 3 (537), **Indian Institute** of **Technology**, Kharagpur ; 4 (541), **Indian Institute** of **Science** ; 5 (543), Thapar ...

Research.com
https://research.com › university-rankings › compute...

Best Computer Science University Ranking in India 2023

Best **Computer Science Universities in India** 2023 · Indian Statistical Institute · Indian Institute of Technology Bombay · Indian Institute of Technology ...

Microsoft Bing

computer science universities in india

EduRank.org
https://edurank.org/cs/in

100+ Best Computer Science universities in India …

Web Jul 18, 2023 · 100 Best universities for Computer Science in India. 1. **Indian Institute of Technology** Kanpur. Uttar Pradesh | Kanpur. For Computer Science. 2. **Indian Institute of Science.** 3. **Indian Institute of Technology** Kharagpur. 4. **Indian Institute of** …

Cyber Security 69

Below is the list of 100 best universities for Cyber Security in I…

Software Engineering 18

Below is a list of best universities in India ranked based on th…

See results only from edurank.org

US News
https://www.usnews.com/.../india/computer-science

Best Global Universities for Computer Science in India

Web Here are the best global **universities** for **computer science** in **India**. Thapar Institute of Engineering & Technology; Vellore Institute of Technology; Indian Institute of Technology …

Collegedunia
https://collegedunia.com/engineering/computer-science-colleges

Top Computer Science(Engineering)Colleges in India

Web 4 rows · **Jharkhand University** of Technology - [JUT], Ranchi- Andhra **University** - [AU], Visakhapatnam- ...

CD RANK	COLLEGES	COURSE FEES	PLACEMENT	USER REVIEWS
#1	IIT Madras - Ind…	Courses	Cutoff 2023 (Rank)	Cutoff 2022 (R…
#2	IIT Delhi - India…	Courses	Cutoff 2023 (Rank)	Cutoff 2022 (R…
-	Featured Parul …	₹ 1,49,000 BE/…	₹ 29,50,000 Highest P…	8.1 / 10 Based o…
#3	IIT Bombay - In…	Courses	Cutoff 2023 (Rank)	Cutoff 2022 (R…

See all 4 rows on collegedunia.com

Amit's Sidebar

- Exercise
  - Read the top three results on the previous slide, and write down which ranking do you prefer and why?
  - Deadline: 48 hours
  - 2 marks

Big Issues in IR

- Users and Information Needs
  - Search evaluation is user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as *query expansion*, *query suggestion*, *relevance feedback* improve ranking

IR and Search Engines

- A search engine is the practical application of information retrieval techniques to large scale text collections
- Web search engines are best-known examples, but many others
  - *Open source* search engines are important for research and development
    - e.g., Lucene, Lemur/Indri, *Galago*
- Big issues include main IR issues but also some others

IR and Search Engines

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*

Search Engine Issues

- Performance
  - Measuring and improving the efficiency of search
    - e.g., reducing *response time*, increasing *query throughput*, increasing *indexing speed*
  - *Indexes* are data structures designed to improve search efficiency
    - designing and implementing them are major issues for search engines

Search Engine Issues

- Dynamic data
  - The “collection” for most real applications is constantly changing in terms of updates, additions, deletions
    - e.g., web pages
  - Acquiring or “crawling” the documents is a major task
    - Typical measures are *coverage* (how much has been indexed) and *freshness* (how recently was it indexed)
  - Updating the indexes while processing queries is also a design issue

Search Engine Issues

- Scalability
  - Making everything work with millions of users every day, and many terabytes of documents
  - Distributed processing is essential
- Adaptability
  - Changing and tuning search engine components such as ranking algorithm, indexing strategy, interface for different applications

Spam

- For Web search, spam in all its forms is one of the major issues
- Affects the efficiency of search engines and, more seriously, the effectiveness of the results
- Many types of spam
  - e.g. spamdexing or term spam, link spam, “optimization”
  - White on white text, only meant for search engines
- New subfield called *adversarial IR*, since spammers are “adversaries” with different goals

Spam vs SEO

- Comes in shades

best vacuum cleaners in india | vacuum cleaner price | vacuum cleaners | best vaccum cleaners for home | best vacuum cleaners

QUICK LINKS

Dyson Air Purifiers	Best Geysers
Best Protein Powder	Best Car Mats
Apple Ipad	Sofa Sets Online
Best Hair Dryers	

Amit's Sidebar

- [query] will denote a query as it kinda looks like a search box :-)
- Important Concepts: Vocabulary mismatch
  - [pain in the bottom of my foot] vs

Mayo Clinic
https://www.mayoclinic.org › syc-20354846
Plantar fasciitis - Symptoms and causes

Amit's Sidebar

- What is the meaning of a word?
  - Cricket

Crickets are orthopteran insects which are related to bush crickets, and, more distantly, to grasshoppers. In older literature, such as Imms,[3] "crickets" were placed at the family level (*i.e.* Gryllidae), but contemporary authorities including Otte now place them in the superfamily **Grylloidea**.[1] The word has been used in combination to describe more distantly related taxa[3] in the suborder Ensifera, such as king crickets and mole crickets.

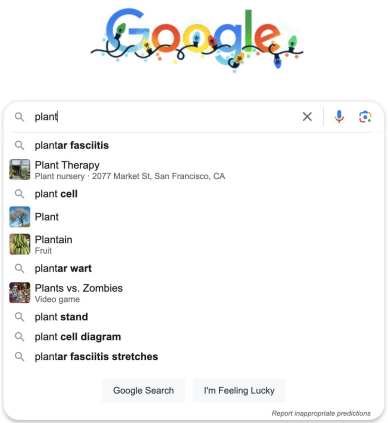
Try Cricket Wireless on Us

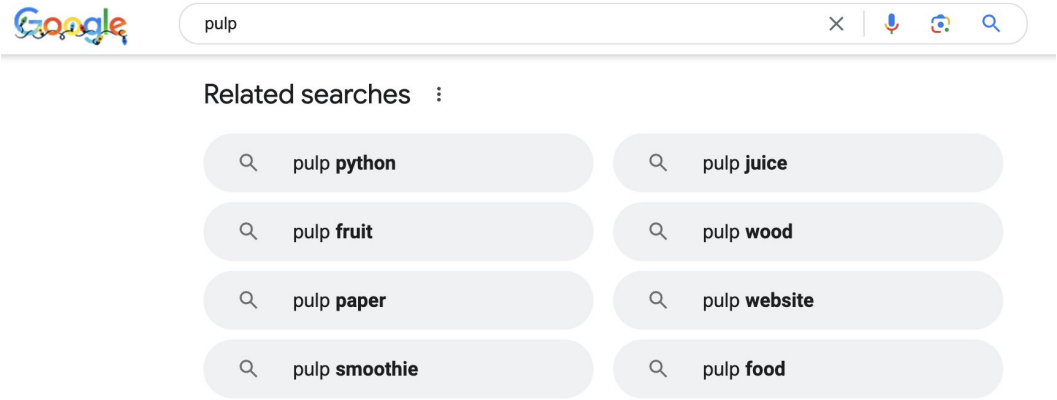
People who come to Cricket stay with Cricket. Now you can see why. Our 14-day free trial lets you test-drive Cricket on your phone without disrupting your existing service with your current carrier.

Get TryCricket App

Amit's Sidebar

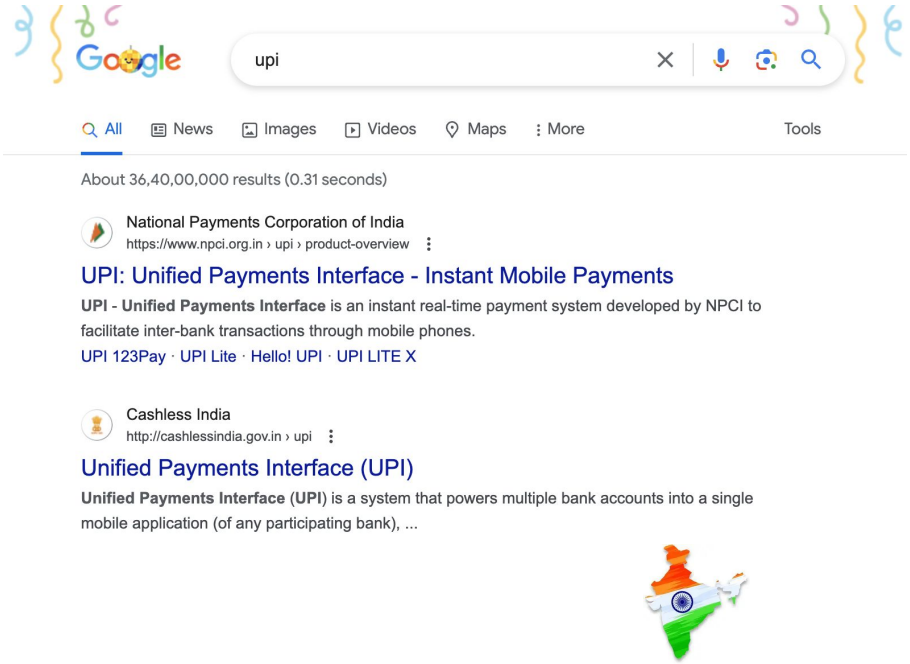
- Important Concepts: Query Formulation
  - Users need help with query formulation: suggest, refinements, etc.

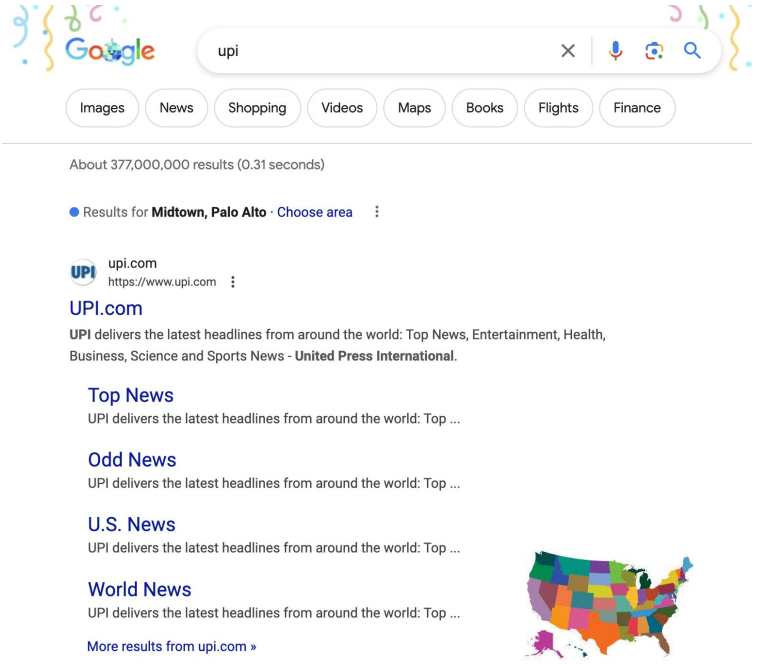




Amit's Sidebar

- Localization



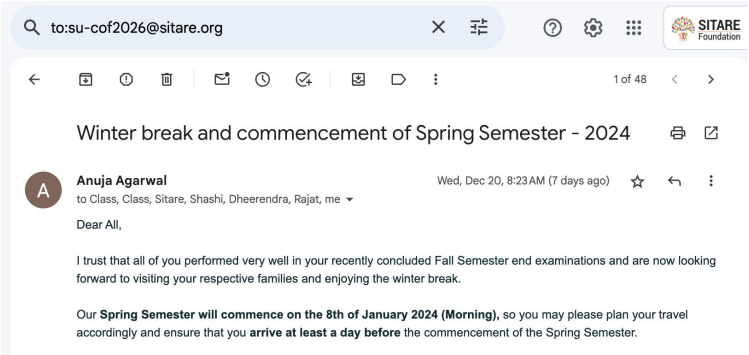


Amit's Sidebar

- Important Concepts: RECALL
  - Recall: ability to find (almost) every relevant document
  - If I return the entire web my search engine has 100% recall, we have found all the relevant webpages, but at the cost of having to read billions of useless documents.
- Important Concepts: PRECISION
  - Precision: ability to find *only* relevant document, and no non-relevant document
  - If I return only one (possibly) relevant webpage, I have high precision, but at what cost?

Amit's Sidebar

- Important Concepts: RECALL
  - Recall: critical in information poor environments (your email)
  - [start of semester] vs



- Important Concepts: PRECISION
  - Precision: critical in information rich environments like the web

Course Goals

- To help you to understand search engines, evaluate and compare them, and modify them for specific applications
- Provide broad coverage of the important issues in information retrieval and search engines
    - includes underlying models and current research directions

Chapter Exercise

- Do exercise 1.1 but only compare the top THREE results between Google and Bing for TEN queries that you have done recently. (6 marks)
- Do exercise 1.4 (2 marks)
- Due Date: Four days

See you next time

SITARE University

In association with SITARE Foundation and SHRI RAMSWAROOP MEMORIAL UNIVERSITY

Image Credit: Adobe Text to Image