بسم الله الرحمن الرحيم

# دورة "استرجاع المعلومات" باللغة العربية - صيف ٢٠٢١

# Information Retrieval – Summer 2021


# 1. Introduction to IR!

*IIR: Ch. 1*

**Tamer Elsayed**

**Qatar University**

## الهدف

"اللهم علّمنا ما ينفعنا
وانفعنا بما علمتنا وزدنا علما"

النية

"من سلك طريقاً يلتمس فيه علماً
سهل الله له به طريقاً إلى الجنة"

# Main Instructor


- **Tamer Elsayed**
  - *Associate Prof of Computer Science*
  - *Qatar University*
  - telsayed@qu.edu.qa


- **Research interests:**
  - ***Information Retrieval***
  - Big Data Analytics

1992-1997: B.Sc.
1998 – 2001: M.Sc.

2002-2009: Ph.D.
2009-2010: PostDoc


2007: Soft. Eng. Intern

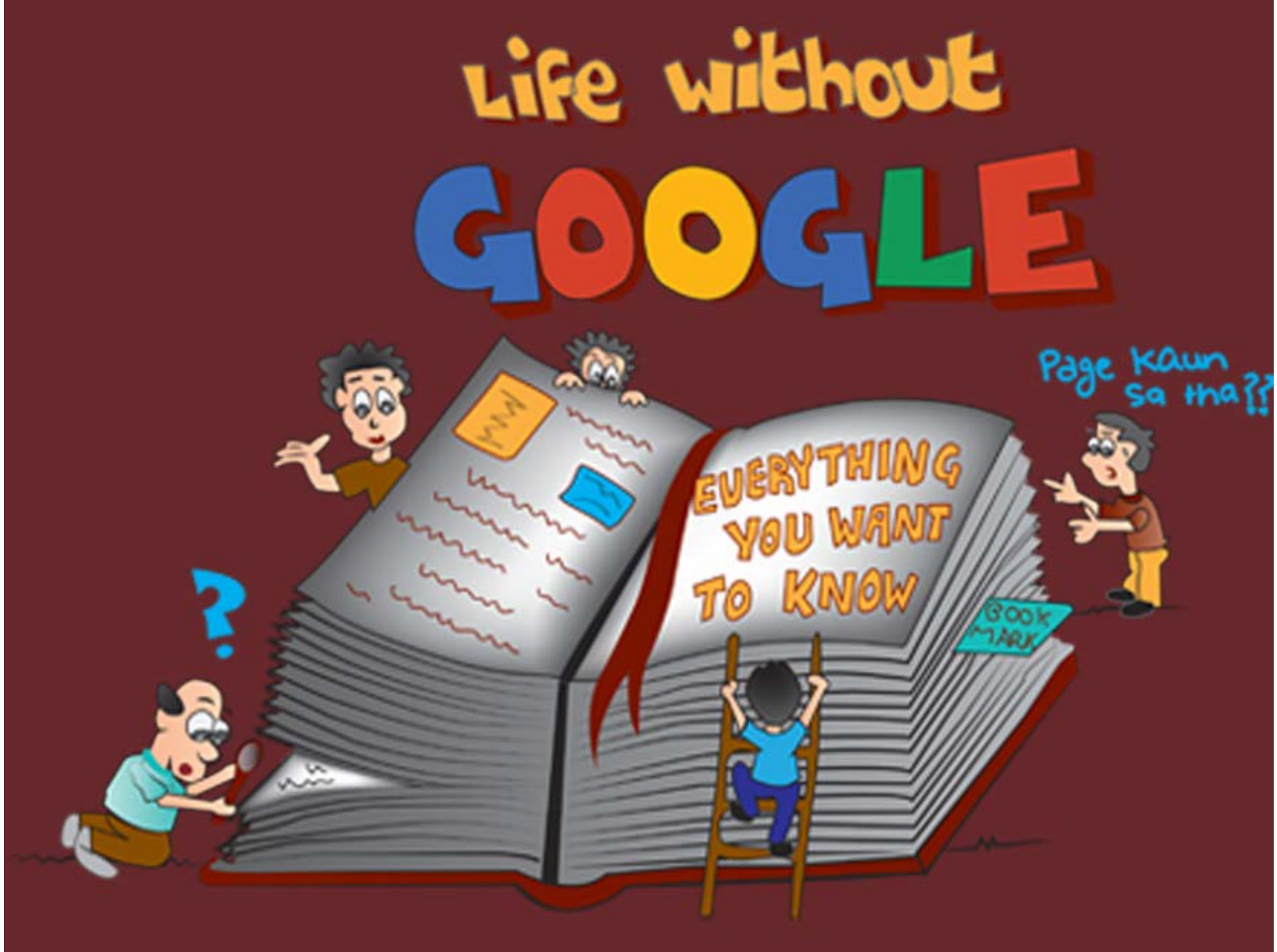


2012 – 2018: Assist. Prof.
2018 – current: Assoc. Prof.


2011-2012: Researcher


2010-2011: PostDoc

**Search Engines**
***changed the world !***


Life without
GOOGLE
Page kaun sa tha??
EVERYTHING YOU WANT TO KNOW
?
BOOK MARK

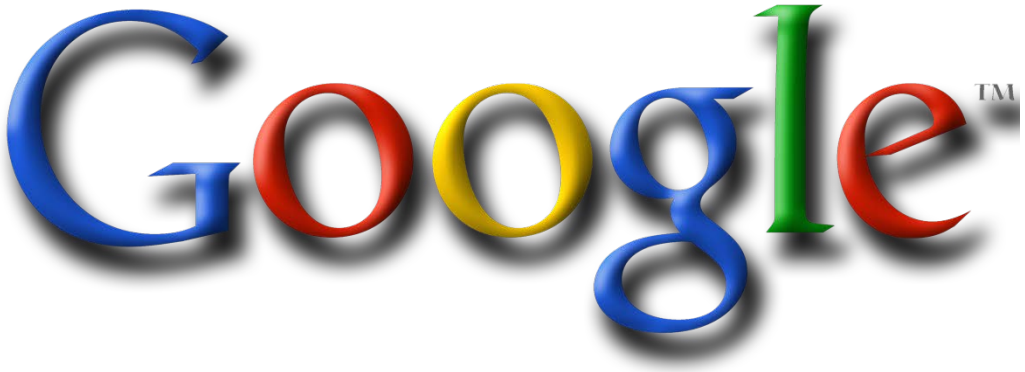
**Search Engines**
***changed the world !***

**IR is the field that *changed the world* !**


Linked in
facebook
You Tube
lost.fm
digg
facebook
orkut
SOCIAL MEDIA

# WHAT IS IR?

# IR is NOT just



# IR is NOT just the search box!

# Microblog Search


# Text Classification

# Expert Search

EXPERTSEARCH.CO.UK

There are **9** records matching your search request :

**Area of Expertise = Information Retrieval**

Your search took **0.250 seconds** to perform.

**Name:** Mr John Allcock web site

**Town/County:** London
**Organisation:** Bristows, Solicitors
**Occupation:** Solicitor and European Patent/Trade Mark Attorney

**Name:** Mr Matthew J Atha web site

**Town/County:** Wigan, Lancs
**Organisation:** Independent Drug Monitoring Unit (IDMU)
**Occupation:** Drug Abuse Research & Information Consultant

**Name:** Miss Annette Clarey web site

**Town/County:** Slough, Berks
**Organisation:** BioMark Forensics Ltd
**Occupation:** Forensic Biologist

BIOMARK FORENSICS

**Name:** Mr Andrew Fox web site

**Town/County:** Plymouth, Devon
**Organisation:** Audax Digital Forensic
**Occupation:** Computer Forensic Consultant

Audax

SAMUEL GINN COLLEGE OF ENGINEERING

Prospective Students | Business and Government | Current Students | Alumni | Faculty

College of Engineering >> Academic Programs >> Graduate School >>

**Faculty Expertise -- Search**

This faculty search tool can be used to identify faculty members within the College of Engineering who have expertise in specific areas of interest. This can be useful for identifying

**Auburn Engineering Faculty Search Engine**

Computer Science and Software Engineering

information retrieval Search

**JUAN E. GILBERT**
TSYS Distinguished Associate Professor
Computer Science and Software Engineering
3101 Shelby Center
Phone: (334) 844-6316, Fax: (334) 844-6329
E-mail: gilbert@auburn.edu
Website: http://www.juangilbert.com

- Human-Centered Computing, Human-Computer Interaction, Spoken Language Systems, Databases, Information Management, Advanced Learning Technologies, Ethnocomputing

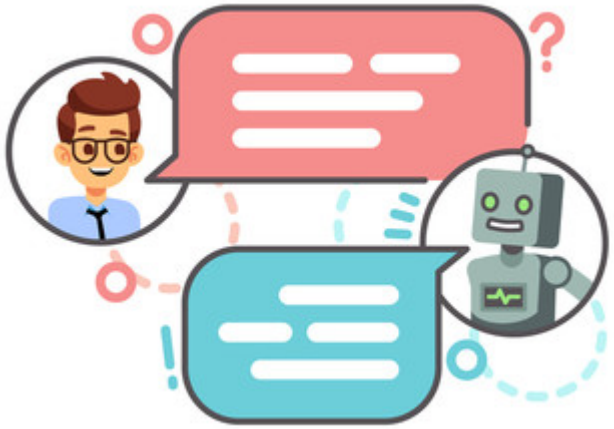
**W. H. CARLISLE**
Associate Professor
Computer Science and Software Engineering
110 Dunstan Hall
Phone: (334) 844-6308, Fax: (334) 844-6329
E-mail: carliwh@auburn.edu

- Languages and algorithms for cooperative autonomous systems, distributed processing, and distributed information sharing and system management.

# Speech Retrieval

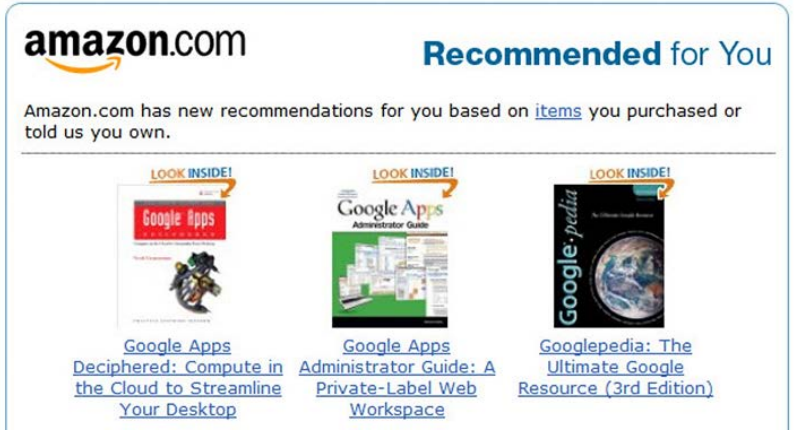


# Conversational Search



# Recommendation






Customers who viewed this item also viewed

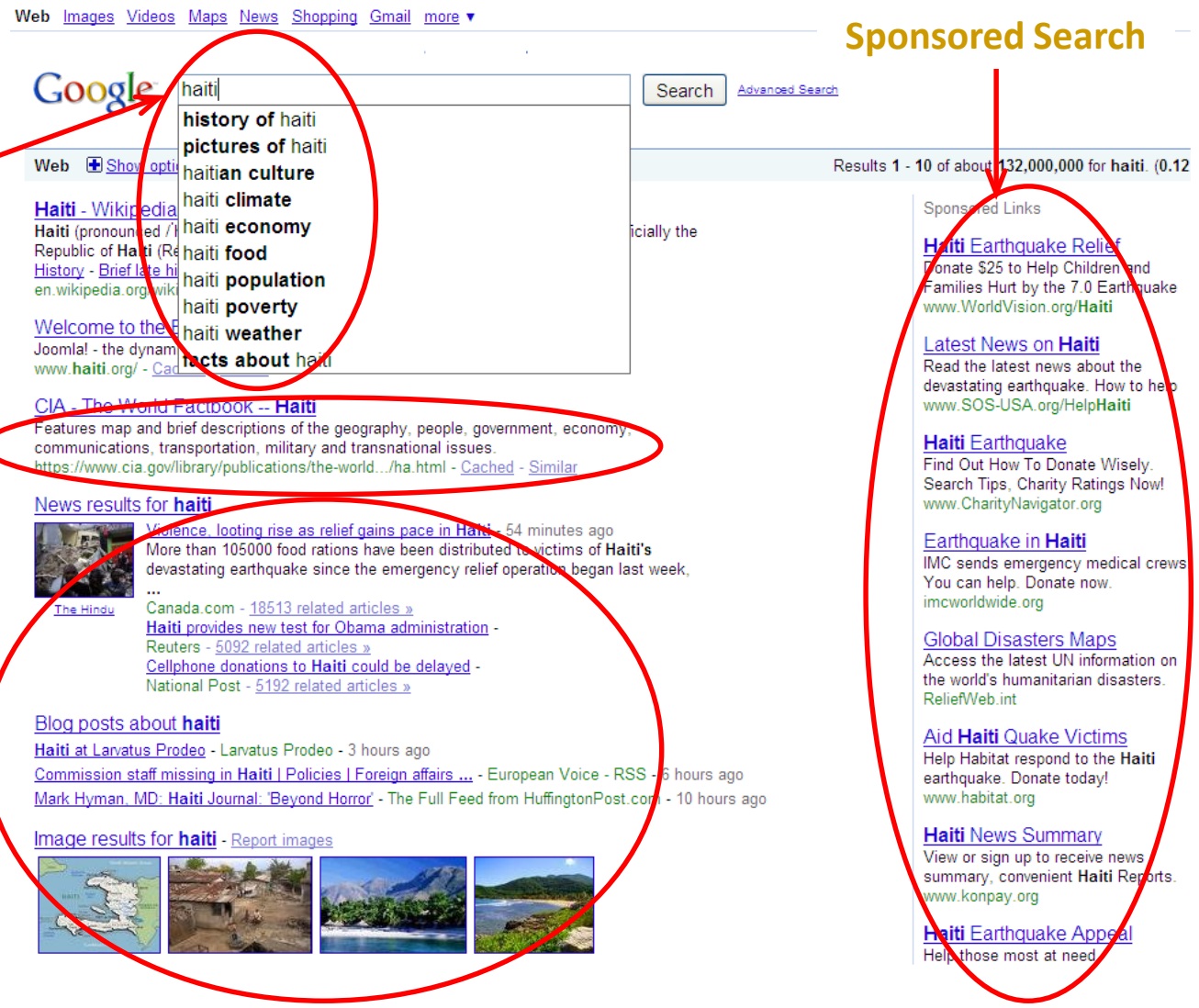
# Stuff on Search Results Page!

Query suggestion / correction

Snippet selection / summarisation

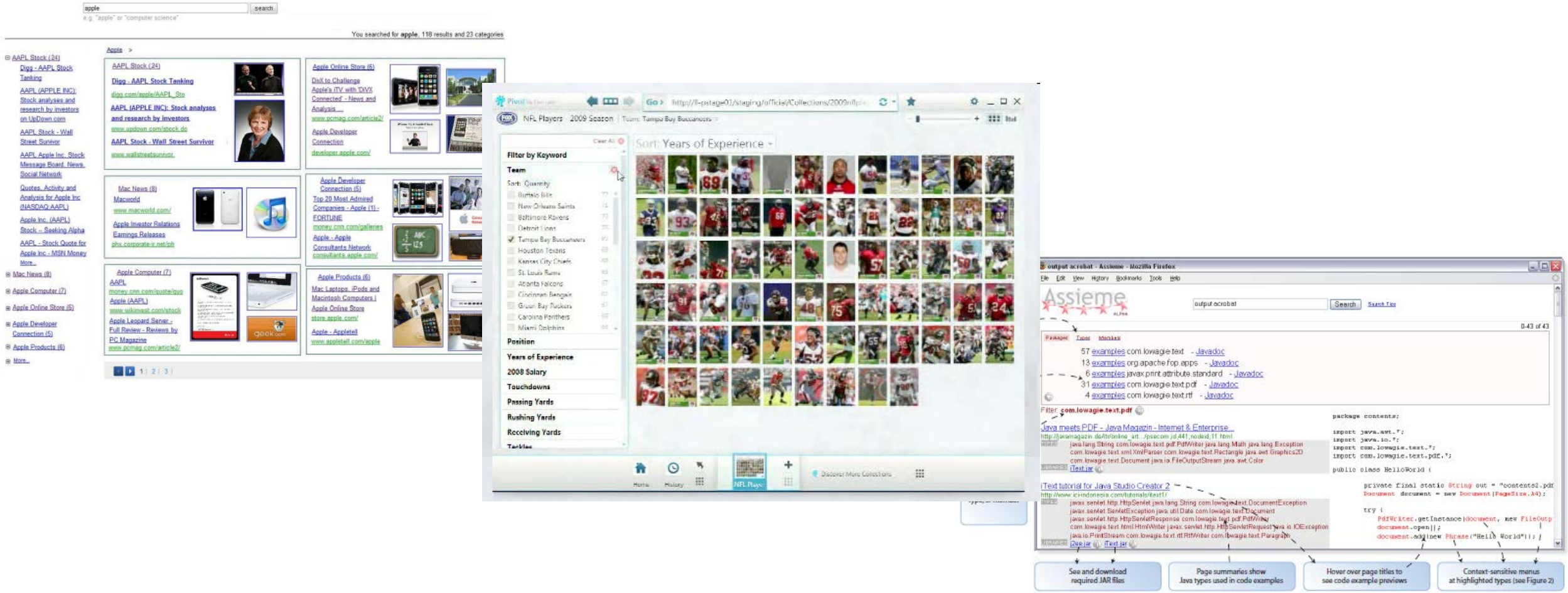
Categorization (search verticals)

Sponsored Search



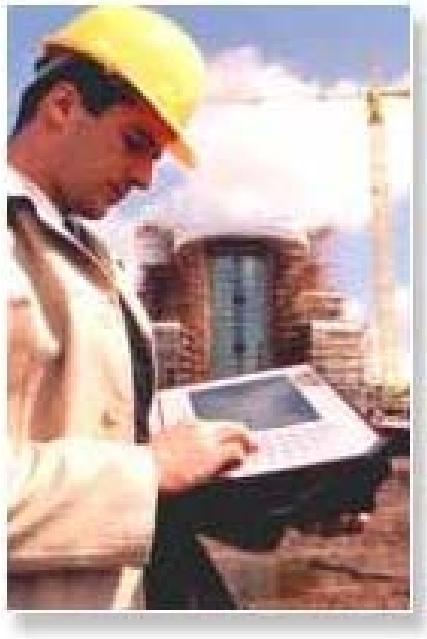

*Source: Matt Lease (IR Course at U Texas)

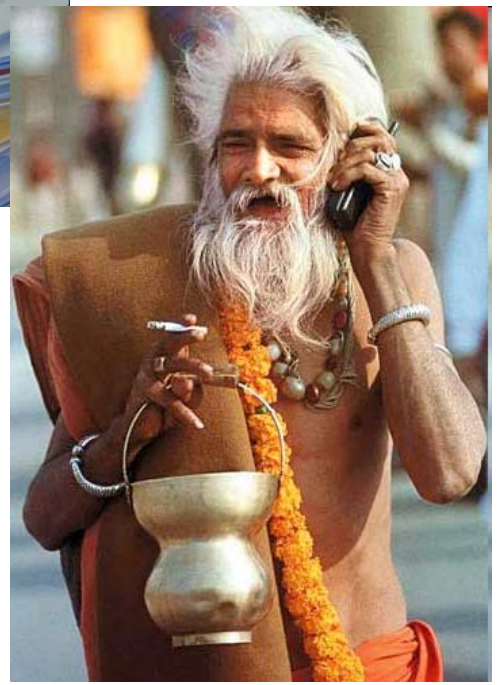
# Information Visualization



***IR is about technology***
***to connect people to information***


EOD






1
POLL


1

## Searching in videos is IR?

- Yes
- No
- Only if videos have text in them

## Spam filtering is IR?

- Yes
- No


# COURSE ORGANIZATION

# Duration & Daily Schedule

- 2 “intensive” weeks!
- Daily From today (Sunday May 30th) to Thursday June 10th
  - except for Friday Jun 4th and Saturday June 5th.

| Time | Activity | Who Attends? |
|---|---|---|
| **4:00pm-5:00pm** | **Lecture Part 1** | **All registered** |
| *5:00pm-5:15pm* | *Break* | |
| **5:15pm-6:15pm** | **Lecture Part 2** | **All registered** |
| *6:15pm-6:45pm* | *Break* | |
| **6:45pm-8:00pm** | **Lab** | ***Full-Registration only*** |

- **Links sent over email. Same links for all days.**
- **Lab material is available for all.**

# Lab Instructors

**Fatima Haouari**

PhD Student

Qatar University

**Watheq Mansour**

MSc Student

Qatar University

**Maram Hasanain**

PhD Student

Qatar University

# What is in Arabic?

- Explanation and discussion in Arabic
- Arabic dataset in labs


- Slides in English
- Terminology in English
- References/textbooks in English

# Goals

- Learn the ***main concepts*** of "information retrieval"
- Gain ***practical experience*** through training on some IR tools

- Preparing ***lecturers*** to spread IR in their societies
- Preparing ***researchers*** in this field

- Motivate attendees to ***gain more knowledge*** in this important technical area.

# What you will learn …

- The ***basics*** of search engines
- How to ***build a (simple) search engine*** (programmatically)
- How to ***evaluate*** search engine performance
- Some ***open issues*** for scientific research in IR

# Topics

| Date | Day # | Lecture (4:00pm-6:15pm) & Lab (6:45pm-8:00pm) |
|---|---|---|
| Sun May 30 | 1 | Introduction & Boolean Retrieval |
| Mon May 31 | 2 | Indexing & Preprocessing |
| Tue Jun 1 | 3 | Evaluating Search Effectiveness |
| Wed Jun 2 | 4 | Ranked Retrieval I: Vector Space Model and BM25 |
| Thu Jun 3 | 5 | Ranked Retrieval II: Language Models |
| Sun Jun 6 | 6 | Query Expansion & Relevance Feedback |
| Mon Jun 7 | 7 | Term Representation & Embeddings (and PageRank) |
| Tue Jun 8 | 8 | Transformers & BERT |
| Wed Jun 9 | 9 | Ranked Retrieval III: Neural Models |
| Thu Jun 10 | 10 | Some IR Research Problems and Resources |

# PLEASE …

# Interactions

- I will ask quick questions!
  - Use reactions to answer
- Several polls every lecture

- Please give reactions
  - at any time ☺

- Ask questions on chat
  - Start with “Q: ” or “ س:”

# Resources

- GitHub
  - https://github.com/telsayed/IR-in-Arabic/tree/master/Summer2021

- Lecture slides
- Readings
- Lab notebooks

- Updated daily

# Prerequisites

- No prior knowledge of IR is required.
- At least, undergrad-level courses
  - Programming (Python)
  - Data structures or Algorithms
  - Basic Probability theory
  - Basic Linear algebra

# References


- Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze, 2008.
- Search Engines: Information Retrieval in Practice, by W. Bruce Croft, D. Metzler, and T. Strohman, 2010.

- An introduction to Neural Information Retrieval, by Bhaskar Mitra and Nick Craswell, 2018.
- Pretrained Transformers for Text Ranking: BERT and Beyond, by Jimmy Lin, Rodrigo Nogueira, and Andrew Yates, 2020.

# BIG Disclaimer!

**This is the "*FIRST*" version!**

# BIG Disclaimer!

- Extensive -- daily lectures!
- Lab material is new -- just prepared for you!
- Explained in Arabic!
- Big class!

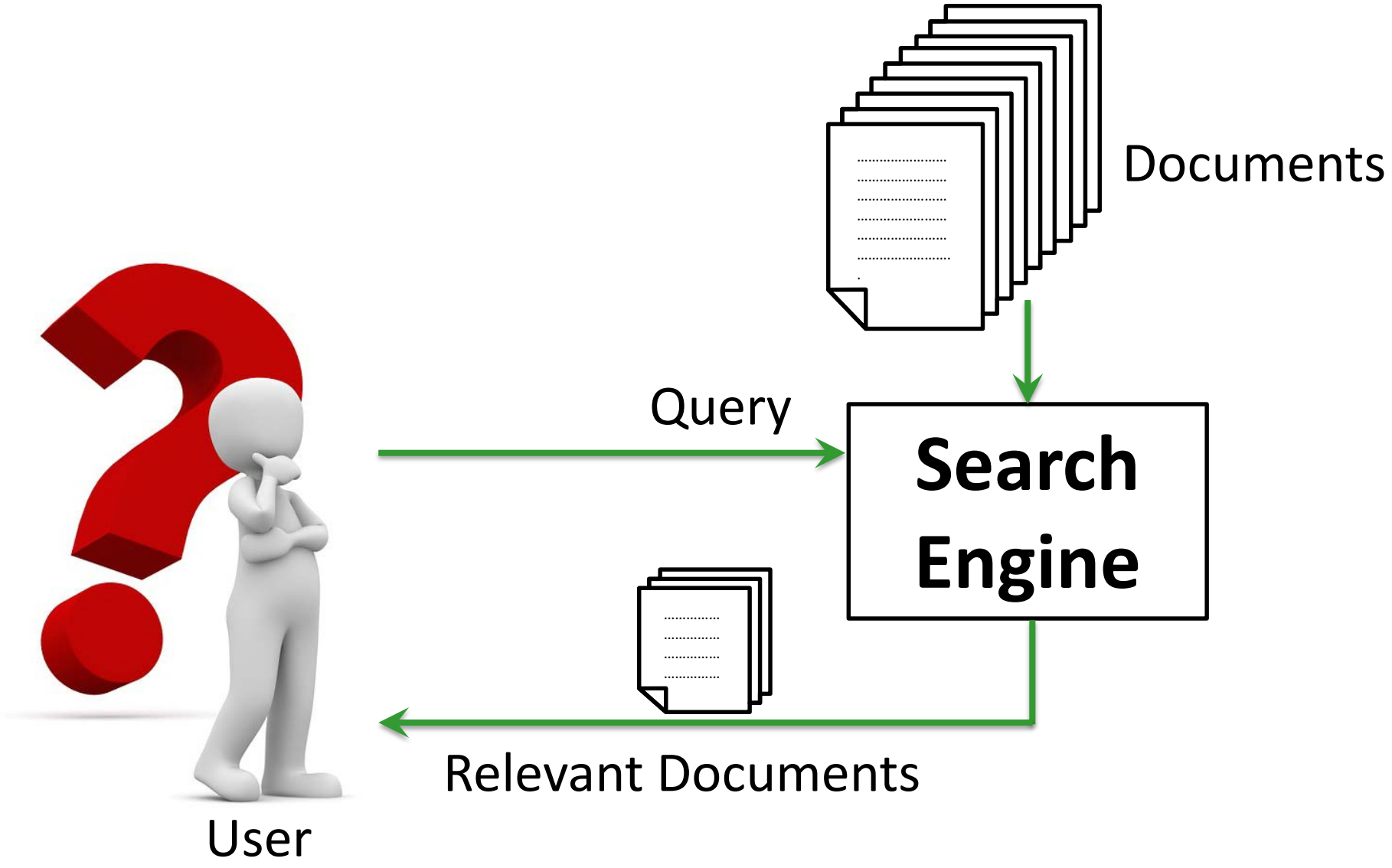
***Be patient …***
***and tell us how to make it better!***

# Today's Roadmap

- Introduction to IR
- How IR "sees" documents?
- Boolean retrieval


# INTRODUCTION TO IR …

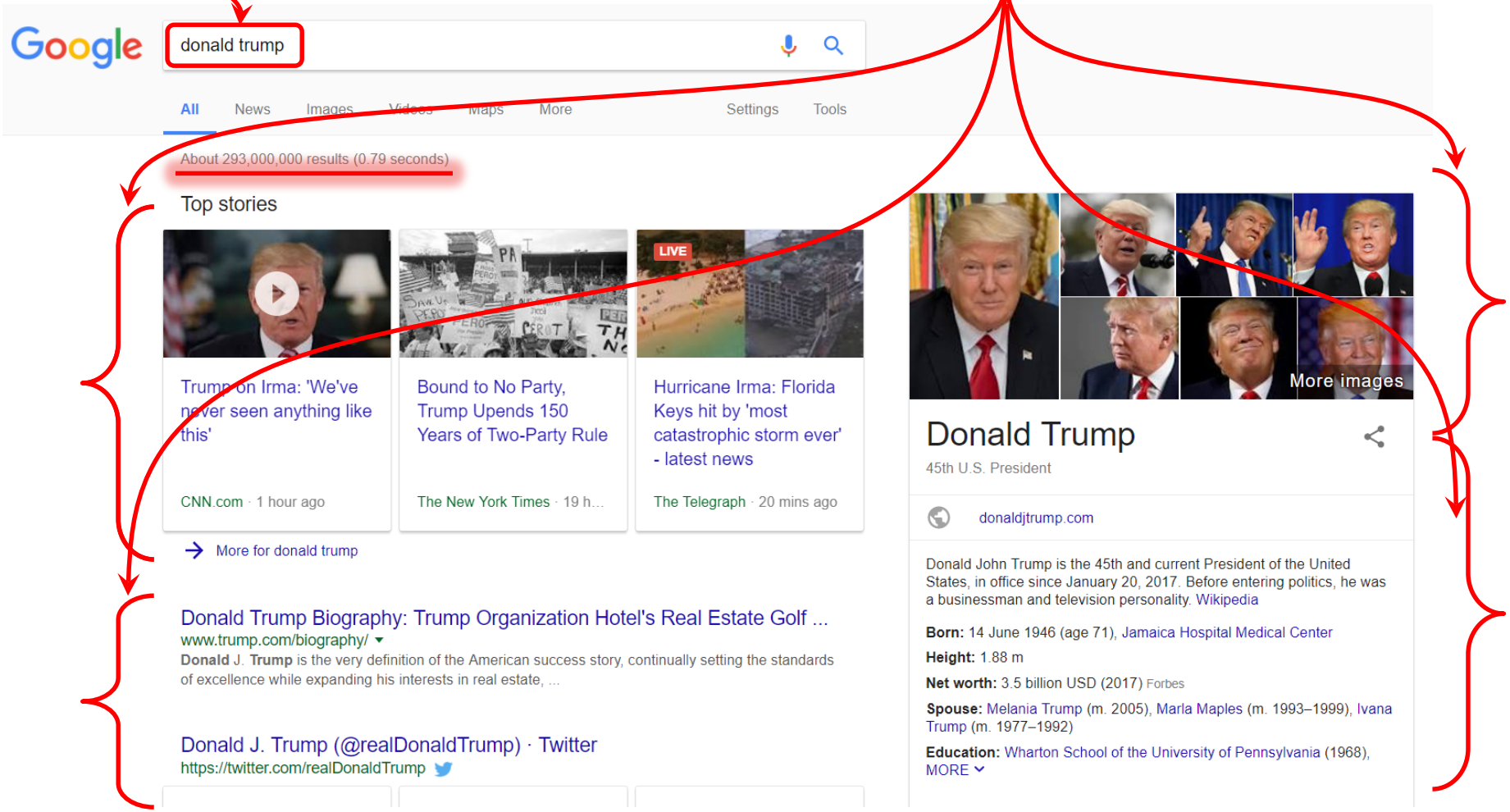
# IR in a nutshell



# IR, basic form

- Given **Query q**, find **relevant documents** ← **?**

**search results D**

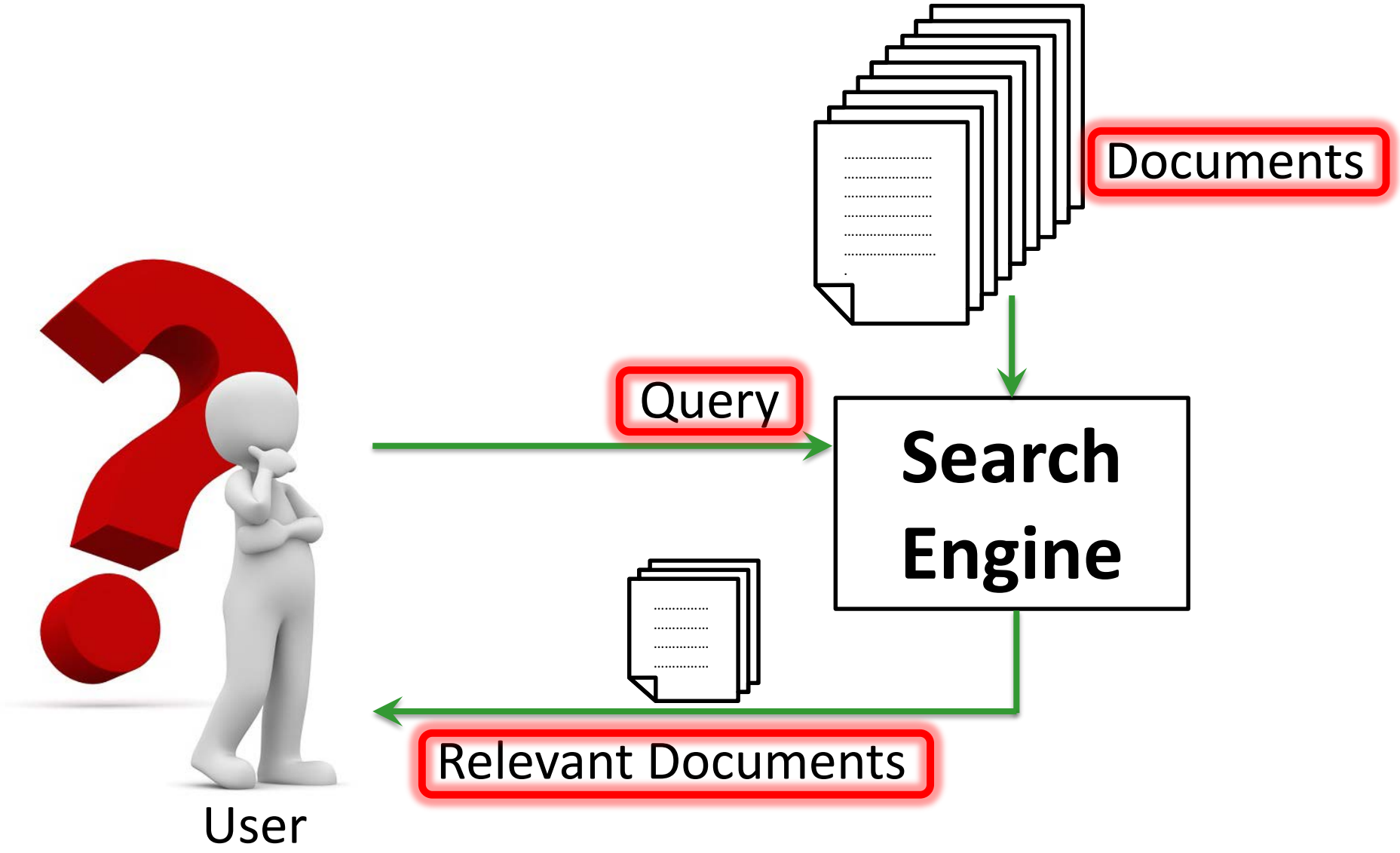


# Two main issues in IR


- **Effectiveness**
  - need to find **relevant** documents
  - needle in a haystack
  - very different from relational DBs (SQL)

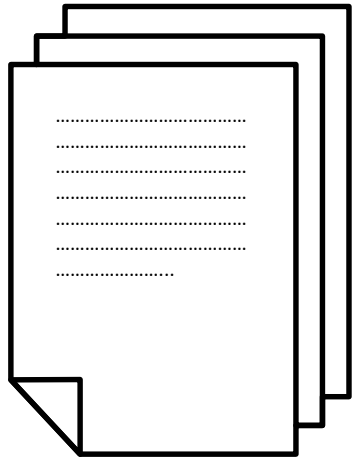
- **Efficiency**
  - need to find them quickly
  - vast quantities of data (10's billions pages)
  - thousands queries per second (Google, ~40,000)
  - data constantly changes, need to keep up

# IR main components



# Documents

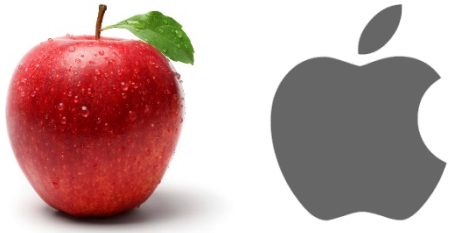
- Document = the **element** to be **retrieved**
  - Unstructured nature
  - Unique ID
  - N documents --> **Collection**



- web-pages, emails, book, page, sentence, tweets
- photos, videos, musical pieces, code
- answers to questions
- product descriptions, advertisements
- people

# Queries

- Free text to **express** user's **information need**
- **Same** information need can be described by **different** queries
  - Are chatting Apps secure?
  - Live chat protection
  - Breaches in online chat
- **Same** query can represent **different** information needs
  - Apple
  - Jaguar



# Queries – different forms

- Web search → keywords, narrative …
- Image search → keywords, sample image
- QA → question
- Music search → humming a tune
- Filtering/recommendation → user's interest/history
- Scholar search → structured (author, title ..)

- Advanced search
  - #wsyn(0.9 #field (title, #phrase (homer,simpson)) 0.7 #and (#> (pagerank,3), #ow3 (homer,simpson)) 0.4 #passage (homer, simpson, dan, castellaneta))

# Relevance

- At an abstract level, IR is about:
  - does item ***d*** match query ***q***? … or …
  - is item ***d*** **relevant** to query ***q***?
- Relevance is a tricky notion
  - will the user like it / click on it?
  - will it help the user achieve a task? (satisfy information need)
  - is it novel (not redundant)?
- Relevance ➔ similarity
  - i.e. ***d***,***q*** share similar “meaning”
  - about the same topic / subject / issue

# Information Need/Query/Relevance

## Information need

- Topic about which the ***user*** desires to know more
- In the user's mind!

## Query

- What the user conveys to the computer
- Considered **one representation** of the information need

## Relevance

- Document having a value with respect to the information need
  - i.e., a document is relevant if it satisfies the information need

# A central problem in search

Searcher

Author


Concepts

Concepts

Query Terms

Document Terms

**Do these represent the same concepts?**

# What is the challenge in relevance?

- No clear semantics!
  - "William Shakespeare"
    - Author history's? list of plays? a play by him?
- Inherent ambiguity of language!
  - polysemy: "Apple", "Jaguar"
- Relevance is highly subjective!
  - Rel: yes/no, Rel: perfect/excellent/good/fair/bad

# Information Retrieval (IR) is …

**Finding material** (usually documents)
of an **unstructured** nature (usually text)
that satisfies an **information need**
from within **large collections**


2
POLL


2

## Is effectiveness more important than efficiency in IR?

- Yes
- No, efficiency is more important
- Both are equally important

## Relevance is having a value with respect to the query.

- Yes
- No, with respect to the *search engine*
- No, with respect to the *information need*

# IR == DB?

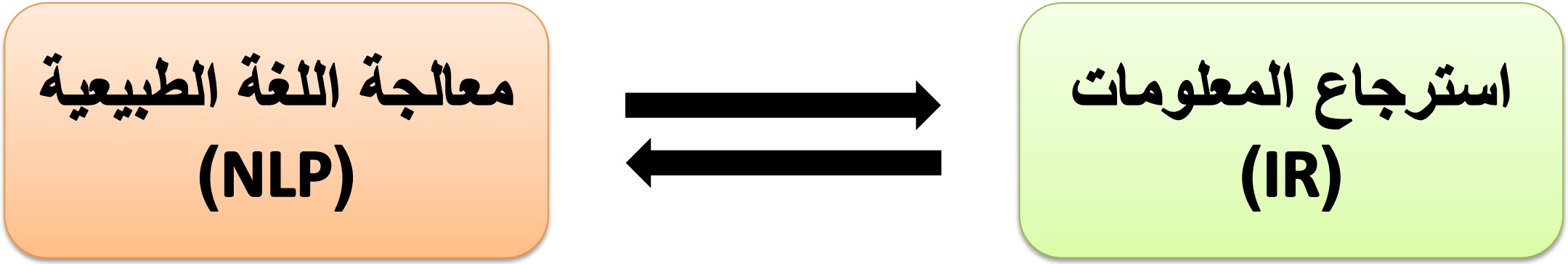
# IR is NOT "DB"

| | Databases | IR |
|---|---|---|
| What we're retrieving | Structured data. Clear semantics based on a formal model. | Mostly unstructured. Free text with some metadata. |
| Queries we're posing | Formally-defined (relational algebra, SQL). Unambiguous. | Free text ("natural language"), Boolean |
| Results we get | Exact (always "correct") | Imprecise (need to measure effectiveness) |
| Interaction with system | One-shot queries. | Interaction is important. |

# IR == NLP?

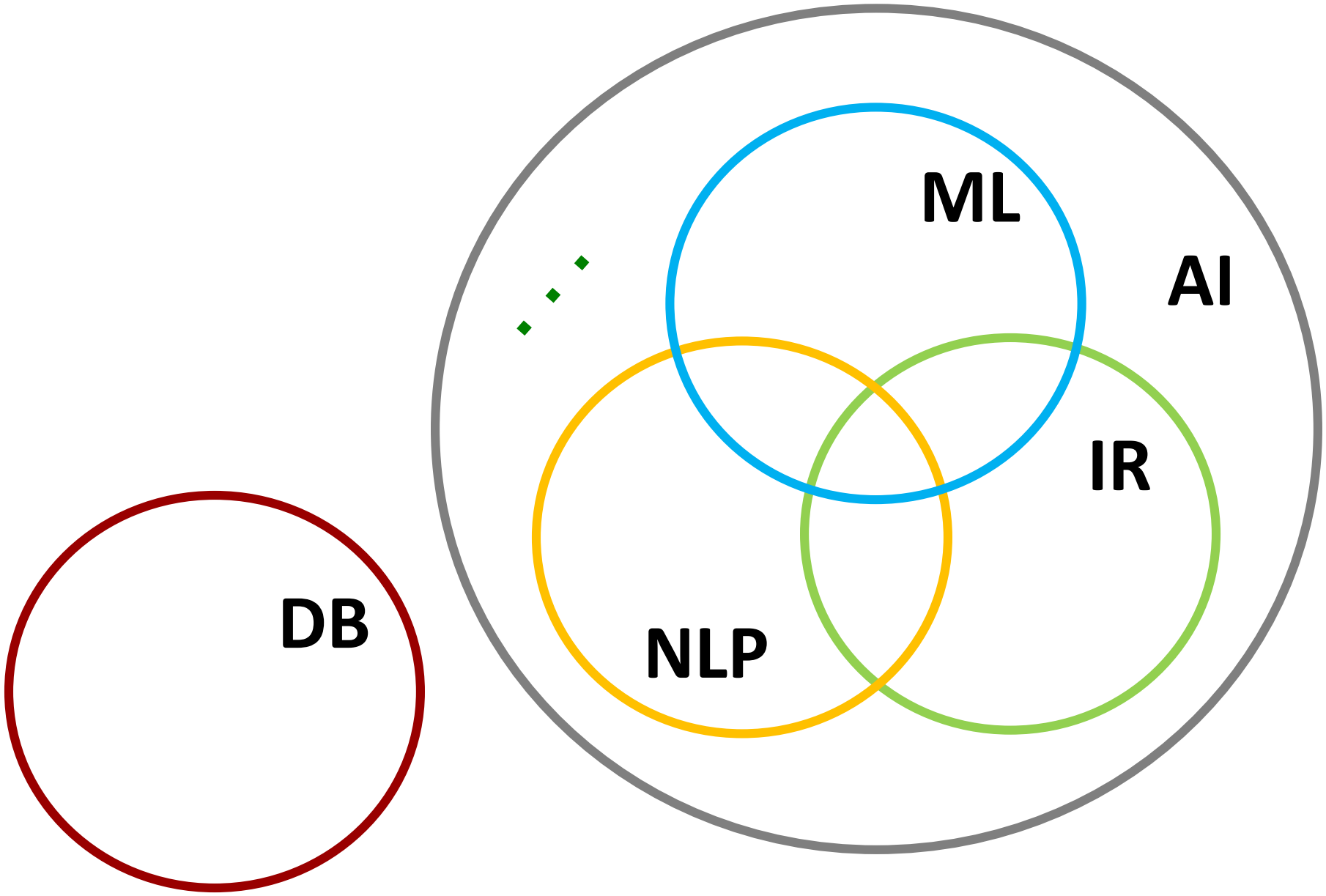
# IR is NOT "NLP"!

- Natural Language Processing (NLP) is about processing and analysis of natural languages.




*"IR makes NLP useful.*
*NLP makes IR interesting."*

- Jimmy Lin

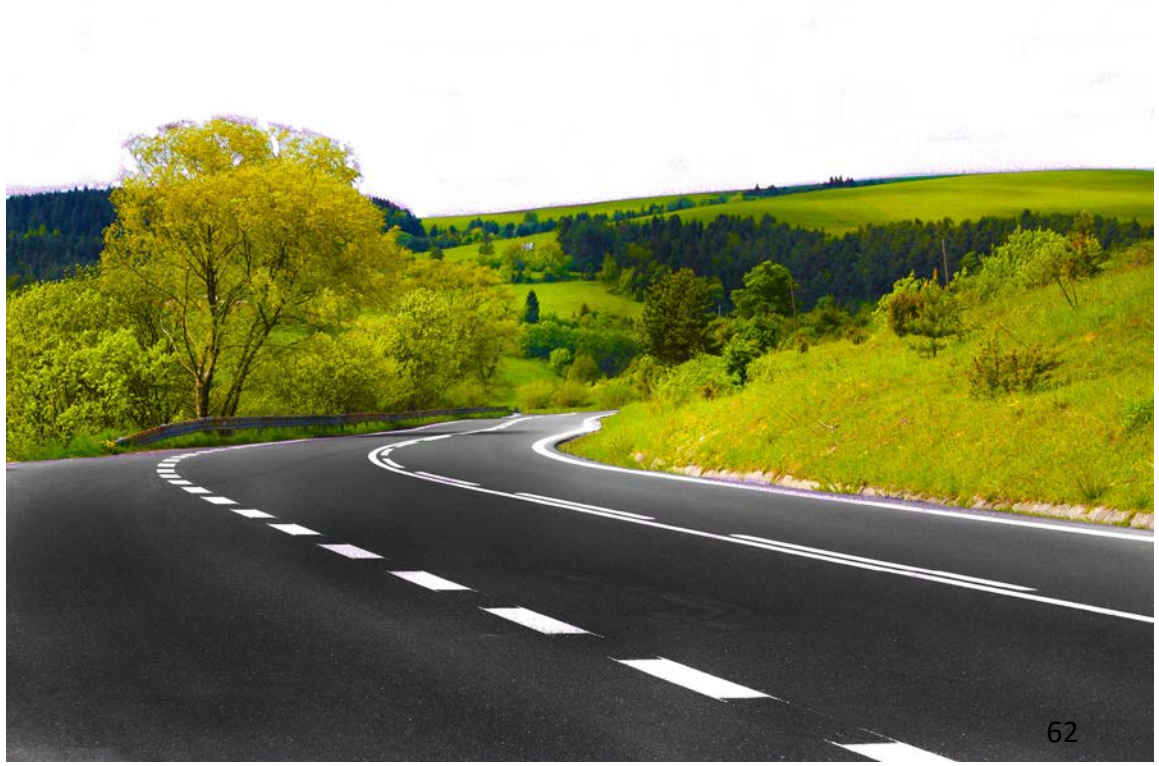

ML
AI
IR
NLP
DB

# Today's Roadmap

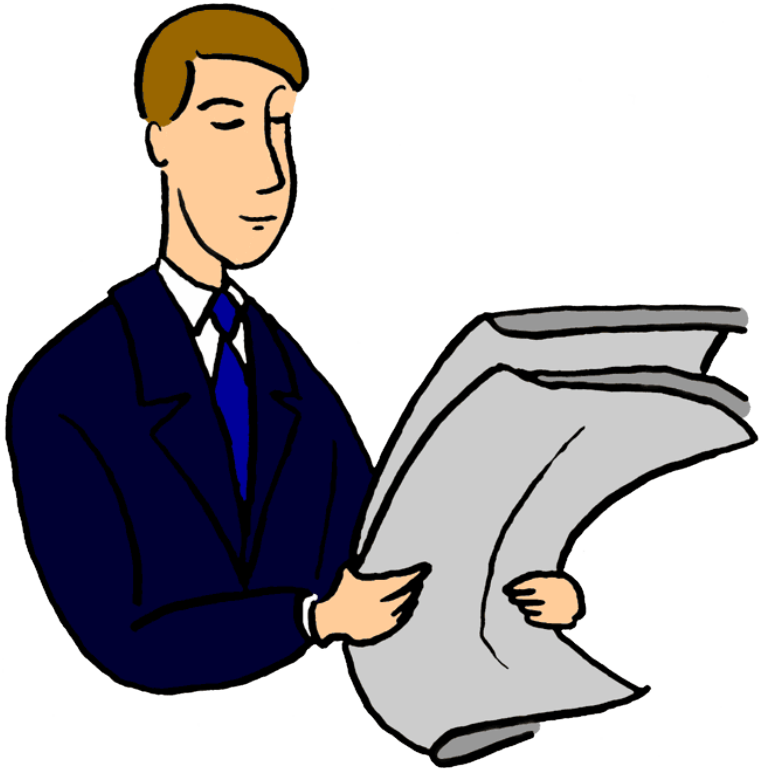
- Introduction to IR …

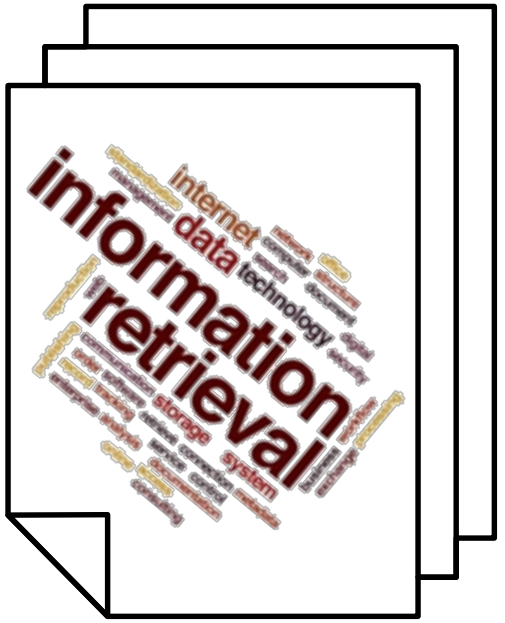
- How IR "sees" documents?

- Boolean retrieval



# How IR “sees” documents?





# Bag-of-Words trick

- Can you guess what this is about:
  - per is salary hour £5,594 Neymar's

    Neymar's salary per hour is £5,594
  - obesity French is of full cause and fat fries

    French fries is full of fat and cause obesity

- **Main idea**: Re-ordering doesn't destroy the topic
  - individual words are "building blocks"
  - "bag" of words: a "composition" of "meanings"

# Simplest: Bag-of-Words trick

- Most search engines use BOW
  - treat documents and queries as bags of words
- A “bag” is a set with repetitions
  - match = “degree of overlap” between ***d***, ***q***

- **Retrieval models**
  - statistical models (functions): usually use words as features
  - decide which documents most likely to be relevant
- BOW makes these models tractable (and also effective!)


# BOOLEAN RETRIEVAL

# What's the Simplest IR System?

- Given a collection of documents and a "free text" query
- How can we get some search results in a simple way?

- ***grep-like***: a "sequential scan"
- Simple but …
  - very inefficient
- Is it effective?

**How can we make it more effective AND efficient?**

# Boolean Retrieval Model

- **Queries**: Users express queries as a ***Boolean expression***
  - AND, OR, NOT
  - Can be arbitrarily nested
- Ex. query: ***information*** *AND* ***retrieval*** *AND NOT* ***technology***

- **Documents**: Views each document as a ***"bag" of words***

- Return only documents that satisfy the Boolean query.

# Exercise

Build a **Term-Document Incidence Matrix**

- Which term appears in which document
- Rows are terms
- Columns are documents

**Given example collection:**

$d_1$: He likes to wink, he likes to drink

$d_2$: He likes to drink, and drink, and drink

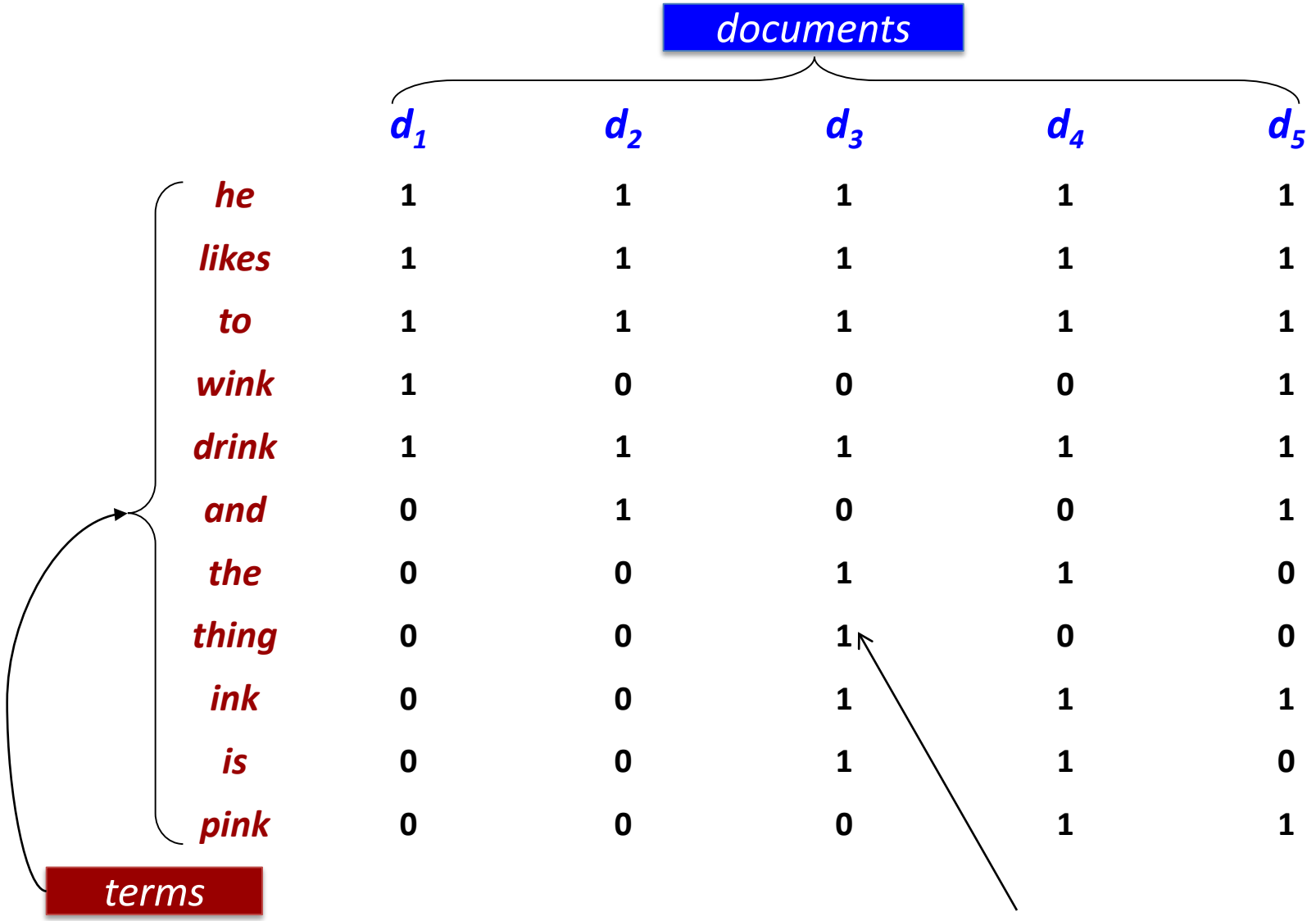
$d_3$: The thing he likes to drink is ink

$d_4$: The ink he likes to drink is pink

$d_5$: He likes to wink, and drink pink ink

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| ***he*** | 1 | 1 | 1 | 1 | 1 |
| ***likes*** | 1 | 1 | 1 | 1 | 1 |
| ***to*** | 1 | 1 | 1 | 1 | 1 |
| ***wink*** | 1 | 0 | 0 | 0 | 1 |
| ***drink*** | 1 | 1 | 1 | 1 | 1 |
| ***and*** | 0 | 1 | 0 | 0 | 1 |
| ***the*** | 0 | 0 | 1 | 1 | 0 |
| ***thing*** | 0 | 0 | 1 | 0 | 0 |
| ***ink*** | 0 | 0 | 1 | 1 | 1 |
| ***is*** | 0 | 0 | 1 | 1 | 0 |
| ***pink*** | 0 | 0 | 0 | 1 | 1 |

# Term-Document Incidence Matrix




| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| *he* | 1 | 1 | 1 | 1 | 1 |
| *likes* | 1 | 1 | 1 | 1 | 1 |
| *to* | 1 | 1 | 1 | 1 | 1 |
| *wink* | 1 | 0 | 0 | 0 | 1 |
| *drink* | 1 | 1 | 1 | 1 | 1 |
| *and* | 0 | 1 | 0 | 0 | 1 |
| *the* | 0 | 0 | 1 | 1 | 0 |
| *thing* | 0 | 0 | 1 | 0 | 0 |
| *ink* | 0 | 0 | 1 | 1 | 1 |
| *is* | 0 | 0 | 1 | 1 | 0 |
| *pink* | 0 | 0 | 0 | 1 | 1 |

1 if ***document*** contains ***term***, 0 otherwise

# Term-Document Incidence Matrix

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| *he* | 1 | 1 | 1 | 1 | 1 |
| *likes* | 1 | 1 | 1 | 1 | 1 |
| *to* | 1 | 1 | 1 | 1 | 1 |
| *wink* | 1 | 0 | 0 | 0 | 1 |
| *drink* | 1 | 1 | 1 | 1 | 1 |
| *and* | 0 | 1 | 0 | 0 | 1 |
| *the* | 0 | 0 | 1 | 1 | 0 |
| *thing* | 0 | 0 | 1 | 0 | 0 |
| *ink* | 0 | 0 | 1 | 1 | 1 |
| *is* | 0 | 0 | 1 | 1 | 0 |
| *pink* | 0 | 0 | 0 | 1 | 1 |

**Query**: ***wink*** *AND* ***drink*** *AND NOT* ***ink***

**Apply on rows**: **10001** *AND* **11111** *AND* !(**00111**) = **10000**

# Boolean Retrieval Model

- Any given query divides the collection into two sets:
  - retrieved (matching)
  - not-retrieved (not matching)
- Returns a ***set*** of documents that "**exactly**" satisfy the query (Boolean expression)
  - Called "**Exact-Match**" retrieval
- Used?
  - Many search systems still in-use are Boolean
  - e.g., Email, library catalog, Mac OS X Spotlight, legal search

# Google?

## Advanced Search

| Find pages with... | | To do this in the search box. |
|---|---|---|
| all these words: | | Type the important words: `tri-colour rat terrier` |
| this exact word or phrase: | | Put exact words in quotes: `"rat terrier"` |
| any of these words: | | Type `OR` between all the words you want: `miniature OR standard` |
| none of these words: | | Put a minus sign just before words that you don't want: `-rodent, -"Jack Russell"` |
| numbers ranging from: | to | Put two full stops between the numbers and add a unit of measurement: `10..35 kg, £300..£500, 2010..2011` |


3
POLL


3

## "Term-Document Incidence Matrix" shows ...

- for each term, the documents that have it
- for each document, the terms that appear in it
- both

## Boolean retrieval always find relevant documents

- Yes, always!
- Sometimes
- Never!

# Bigger Collections …

- Consider *N* = 1 million documents, each with about 1000 words.
- Say there are *M* = 500K *distinct* terms among these.
- 500K x 1M matrix has half-a-trillion 0's and 1's.

- But it has no more than one billion 1's. ?
  - matrix is extremely sparse.

**What's a better representation?**


*Will Term-Doc Incidence Matrix "works" for large collections?*

*If not, how can we make retrieval efficient?*

*How documents are preprocessed?*

*Is "Car" == "Cars"?*