

نظم استرجاع المعلومات Information Retrieval System

مقدمة Introduction

Afaf Al shalaby, PH.D

مخطط العرض PLAN

- الهدف من مقرر نظم استرجاع المعلومات (IR) Information Retrieval System
- تطبيقات نظم استرجاع المعلومات
- ما هي نظم استرجاع المعلومات؟
- تنظيم المقرر الدراسي Course Organization
- خارطة الطريق لليوم Today's Roadmap
- التحديات Challenges

الهدف من المقرر Goals



- تعلم المبادئ الأساسية في نظم استرجاع المعلومات.
- التعرف على الأدوات المستعملة لبناء نظم استرجاع معلومات من خلال جلسات العملي

الهدف من المقرر

Search Engines
changed the world !

Life without
GOOGLE

الهدف من المقرر





الهدف من المقرر

- تتيح لنا محركات البحث، البحث عن المعلومات بشكل آلي.
- سرعة في البحث
- جودة في المصادر
- تخفيض الكلفة في البحث
- إتاحة الوصول لمعلومات كبيرة جدا

الهدف من المقرر

Search Engines
changed the world !

IR is the field that
changed the world !

- نظم استرجاع المعلومات هي
التقنية أو العلم ما وراء
محركات البحث

مخطط العرض PLAN

- الهدف من المقرر
- تطبيقات نظم استرجاع المعلومات
- ما هي نظم استرجاع المعلومات؟
- تنظيم المقرر الدراسي Course Organization
- خارطة الطريق لليوم Today's Roadmap
- التحديات Challenges

تطبيقات نظم استرجاع المعلومات



IR is **NOT** just

Google™

- ما هي IRS؟

- ما هي تطبيقات نظم استرجاع المعلومات الموجودة في حياتنا اليومية؟

- استرجاع المعلومات هو ليس فقط جوجل

تطبيقات IRS

- ليس مجرد البحث النصي في غوغل . IR is NOT just the search box!

The Google logo is displayed in its characteristic multi-colored font.A standard Google search bar with a light blue border and a small microphone icon on the right side.

Google Search

I'm Feeling Lucky

تطبيقات IRS



• microblog search:

- البحث في المدونات الصغيرة
- كالمبحث في وسائل التواصل الاجتماعي

تطبيقات IRS



- Text :classification
- التصنيف الآلي للنصوص

تطبيقات IRS

EXPERT SEARCH CO.UK

There are 9 records matching your search request:

Area of Expertise = Information Retrieval

Your search took 0.260 seconds to perform.

Name: [Mr John Alcock](#)

Town/County: London

Organisation: Bristol, Solicitors

Occupation: Solicitor and European Patent/Trade Mark Attorney

Name: [Mr Matthew J Abba](#)

Town/County: Wigan, Lancs

Organisation: Independent Drug Monitoring Unit (IDMU)

Occupation: Drug Abuse Research & Information Consultant

Name: [Miss Annette Clancy](#)

Town/County: Slough, Berks

Organisation: BioMark Forensics Ltd

Occupation: Forensic Biologist

Name: [Mr Andrew Fox](#)

Town/County: Plymouth, Devon

Organisation: Audax Digital Forensic

Occupation: Computer Forensic Consultant

SAMUEL GINN COLLEGE OF ENGINEERING

Prospective Students Business and Government Current Students Alumni Faculty

College of Engineering > Academic Processes > 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

JAMES GILBERT
Distinguished Associate Professor
Computer Science and Software Engineering
210 Shufly Center
Phone: (334) 844-6216 Fax: (334) 844-6326
E-mail: gilbert@auburn.edu
Website: <http://www.jamesgilbert.com>

Human-Centered Computing, Human-Computer Interaction, Open-Language Systems, Database Information Management, Advanced Learning Technologies, Ethnocomputing

W. H. CARLISLE
Associate Professor
Computer Science and Software Engineering
110 Doreen Hall
Phone: (334) 844-6335 Fax: (334) 844-6326
E-mail: carlisle@auburn.edu

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

Expert Search :

- البحث عن أشخاص (ليس وثائق) خبراء في مجال معين وفي مجتمع معين وبثقه عالية

تطبيقات IRS



- **Speech Search: البحث**
عن مقاطع صوتية وليس
نصوص

تطبيقات IRS



- **Conversational Search:** الإجابة الآلية عن أسئلة المستخدم في نظم التحوار الآلي

تطبيقات IRS



- **Recommendation**
: التوصيات سواء التتبع
للأشخاص أو التوصيات
ببعض المنتجات

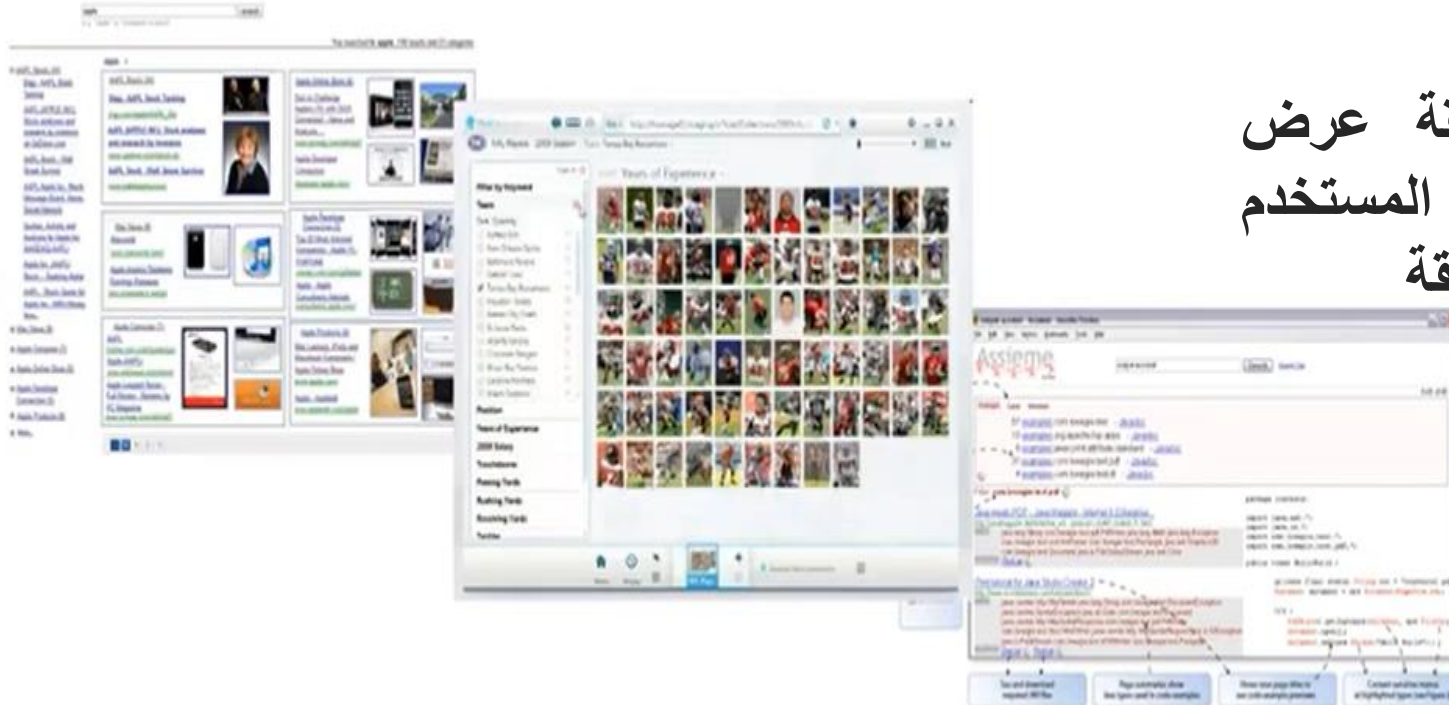
تطبيقات IRS

Stuff on Search Results Page!

The image shows a Google search results page for the query "haiti". Red circles and arrows highlight specific features of the search results page, categorized into three main areas:

- Query suggestion / correction:** This area points to the search bar and the dropdown menu showing suggestions such as "history of haiti", "pictures of haiti", "haitian culture", "climate", "economy", "food", "population", "poverty", "weather", and "facts about haiti".
- Snippet selection / summarisation:** This area points to the search results snippets, which include a brief description of the country, a map, and a list of related articles.
- Categorization (search verticals):** This area points to the "Sponsored Search" section on the right side of the page, which displays various sponsored links related to Haiti, including "Haiti Earthquake Relief", "Latest News on Haiti", "Haiti Earthquake", "Earthquake in Haiti", "Global Disasters Maps", "Aid Haiti Quake Victims", "Haiti News Summary", and "Haiti Earthquake Appeal".

تطبيقات IRS



Information •
Visualization: طريقة عرض
نتائج البحث بحيث يستفيد المستخدم
من هذه النتائج بأفضل طريقة

مخطط العرض PLAN

- الهدف من المقرر
- تطبيقات نظم استرجاع المعلومات
- ما هي نظم استرجاع المعلومات؟
- تنظيم المقرر الدراسي Course Organization
- خارطة الطريق لليوم Today's Roadmap
- التحديات Challenges

ما هي نظم استرجاع المعلومات؟

- هي التقنية أو العلم الذي يربط الناس بالمعلومات

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

ما هي نظم استرجاع المعلومات؟



مخطط العرض PLAN

- الهدف من المقرر
- تطبيقات نظم استرجاع المعلومات
- ما هي نظم استرجاع المعلومات؟
- تنظيم المقرر الدراسي Course Organization
- خارطة الطريق لليوم Today's Roadmap
- التحديات Challenges

تنظيم المقرر الدراسي Course Organization

- **Duration & Daily Schedule**

- 3 “intensive” months!
- Every Sunday from today (Sunday October 14th) to Sunday January 15th

Time	Activity
On Sunday 8:00 am-9:30 am	Lecture Part 1
On Sunday 11:00 am-12:30 pm	Lecture Part 1
On Sunday 9:30 am-11:00 am	Practical Session



Resources المصادر

- **Lecture slides**
- **Readings**
 - **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.
- **Lab notebooks**

Prerequisites المتطلبات المسبقة

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

مخطط العرض PLAN

- الهدف من المقرر
- تطبيقات نظم استرجاع المعلومات
- ما هي نظم استرجاع المعلومات؟
- تنظيم المقرر الدراسي Course Organization
- خارطة الطريق لليوم Today's Roadmap
- التحديات Challenges

خارطة الطريق لليوم Today's Roadmap

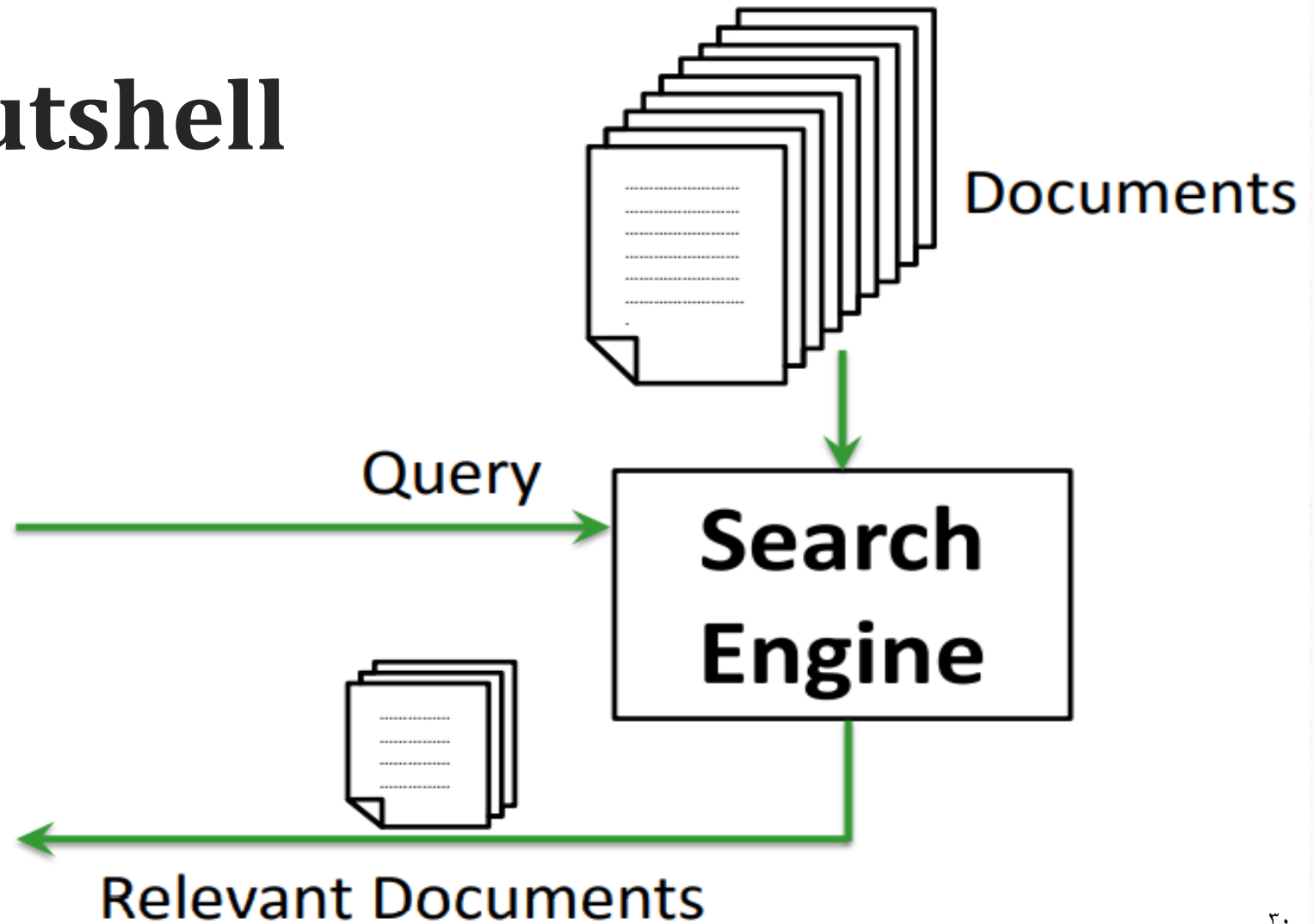
- Introduction to IR
- How IR “sees” documents?
- Boolean retrieval



Introduction to IR



IR in a nutshell



IR, basic form

Given **Query q**, find **relevant documents** ← ?
search results D ←

The image shows a Google search interface for the query "donald trump". Red arrows and brackets illustrate the information retrieval process:

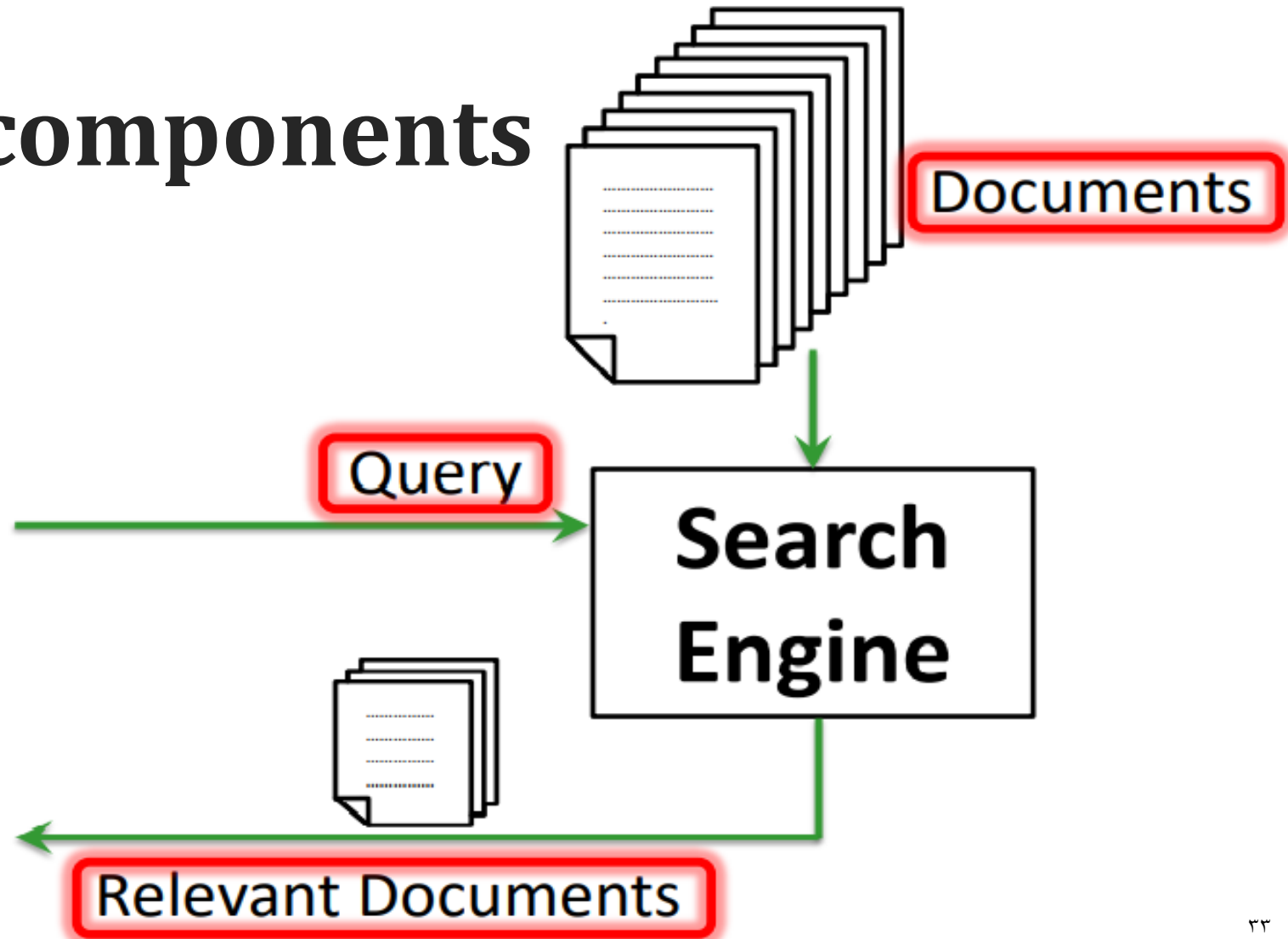
- Query q:** "donald trump" (highlighted in a red box).
- Search Results D:** The entire page of results is bracketed on the right.
- Top stories:** A section of three news cards is bracketed on the left. The cards are:
 - Trump on Irma: 'We've never seen anything like this' (CNN.com - 1 hour ago)
 - Bound to No Party, Trump Upends 150 Years of Two-Party Rule (The New York Times - 19 h...)
 - Hurricane Irma: Florida Keys hit by 'most catastrophic storm ever' - latest news (The Telegraph - 20 mins ago)
- Donald Trump Biography:** A link to "Trump Organization Hotel's Real Estate Golf ..." is bracketed on the left.
- Donald J. Trump (@realDonaldTrump) · Twitter:** A link to his Twitter profile is bracketed on the left.
- Donald Trump Profile:** A detailed profile card on the right is bracketed on the right. It includes:
 - 45th U.S. President
 - donaldjtrump.com
 - Biography: Donald John Trump is the 45th and current President of the United States, in office since January 20, 2017. Before entering politics, he was a businessman and television personality. Wikipedia
 - Born: 14 June 1946 (age 71), Jamaica Hospital Medical Center
 - Height: 1.88 m
 - Net worth: 3.5 billion USD (2017) Forbes
 - Spouse: Melania Trump (m. 2005), Marla Maples (m. 1993–1999), Ivana Trump (m. 1977–1992)
 - Education: Wharton School of the University of Pennsylvania (1968), MORE

Two main issues in IR

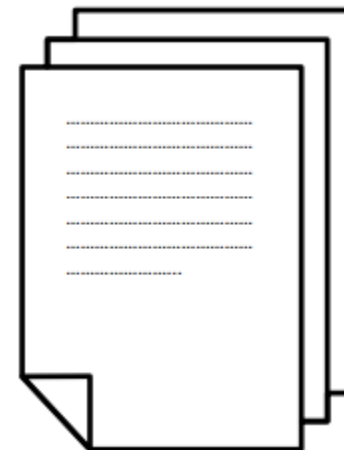
About 293,000,000 results (0.79 seconds)

- Effectiveness (هل هي ذات صلة)
 - need to find relevant documents
 - needle in a haystack
 - very different from relational DBs (SQL)
- Efficiency (سرعة البحث أو سرعة استجابة محرك البحث وسرعة الرد)
 - 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



الوثائق Documents



- Document = the **element** to be retrieved العنصر المراد استرجاعه
 - Unstructured nature عادة تكون غير مهيكلة
 - Unique ID لها رقم فريد
 - N documents --> Collection عددها كبير وتدعى مدونة أو مجموعة
- web-pages, emails, book, page, sentence, tweets
- photos, videos, musical pieces, code
- answers to questions
- product descriptions, advertisements
- people

الاستعلامات Queries

- لدى المستخدم مهمة يريد البحث عنها نعبر عنها ب حاجة المعلومات وهي موجودة داخل عقل المستخدم، أعبّر عن هذه الحاجة داخل صندوق البحث باستعلام
- الاستعلام ليس هو Information need وإنما هي تمثيل لحاجة المعلومات
- Free text to **express** user's **information need**
- Same information need can be described by different queries مثلاً للبحث عن الخصوصية في تطبيقات الدردشة
 - هل تطبيقات الدردشة آمنة؟ Are chatting Apps secure?
 - حماية الدردشة الحية Live chat protection
 - الخروقات في الدردشة عبر الانترنت Breaches in online chat
- Same query can represent different information needs
 - Apple
 - Jaguar



Queries – different forms

• هناك العديد من الأشكال لكتابة الاستعلام:

- Web search e.g.: keywords, narrative ...
- Image search e.g. keywords, sample image
- QA e.g. question
- Music search e.g. humming a tune
- Filtering/recommendation e.g. user's interest/history
- Scholar search e.g. 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))`

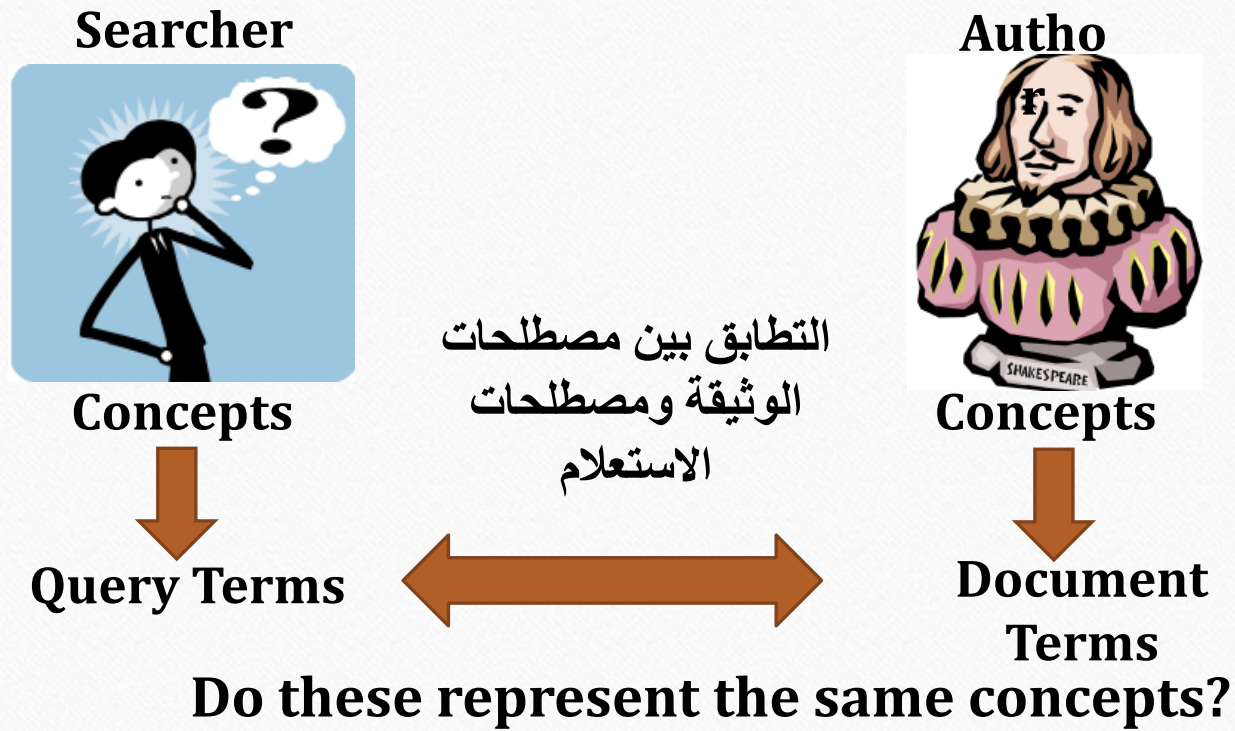
الموائمة Relevance

- في المستوى التجريدي، مهمة محركات البحث إعادة نتائج (وثائق) توافق استعلام معطى:
- 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 and similarity يتحول موضوع الموائمة إلى قياس درجة التشابه
 - i.e. d,q share similar “meaning” بين الوثيقة والاستعلام
 - about the same topic / subject / issue التشابه بالموضوع

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



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

IR vs DB vs NLP vs ML

IR == DB?

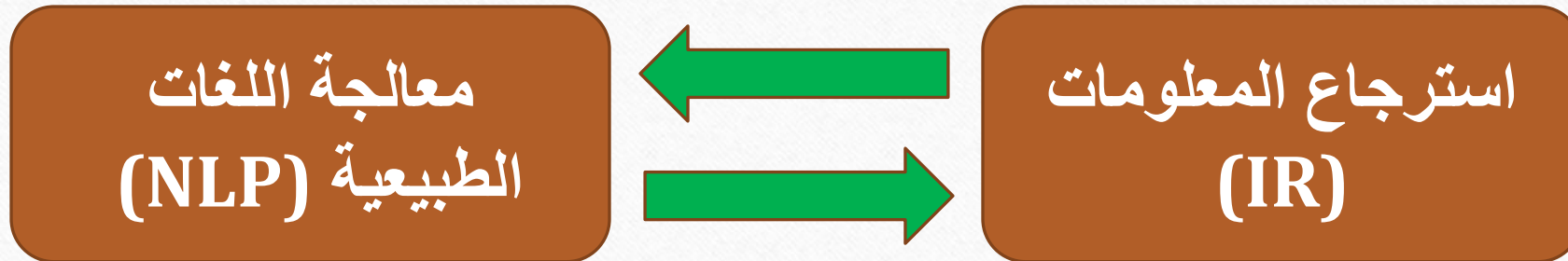
IR is NOT “DB”

	DB	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 vs DB vs NLP vs ML

IR == NLP?

IR is NOT “NLP”!

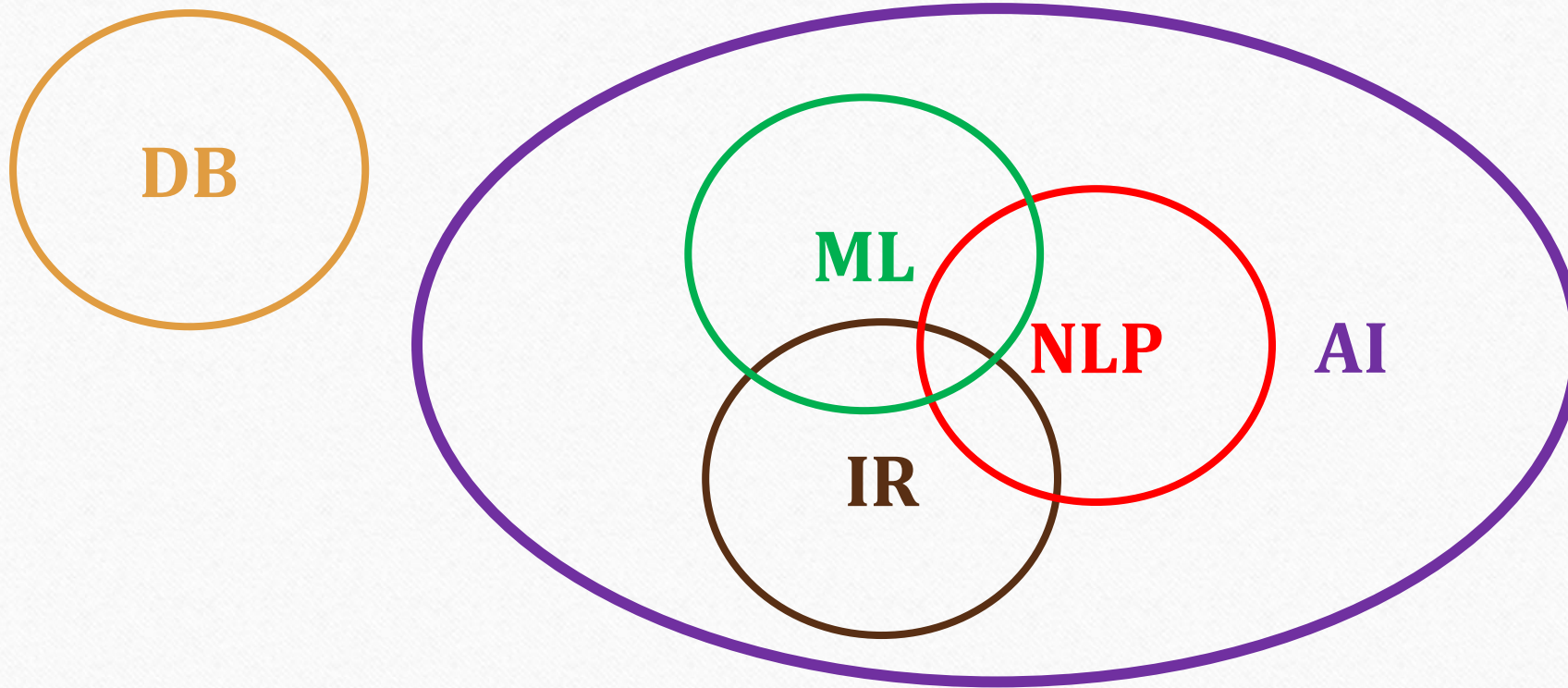


**“IR makes NLP useful. NLP makes IR interesting.”
- Jimmy Lin**

IR vs DB vs NLP vs ML

IR == ML?

AI

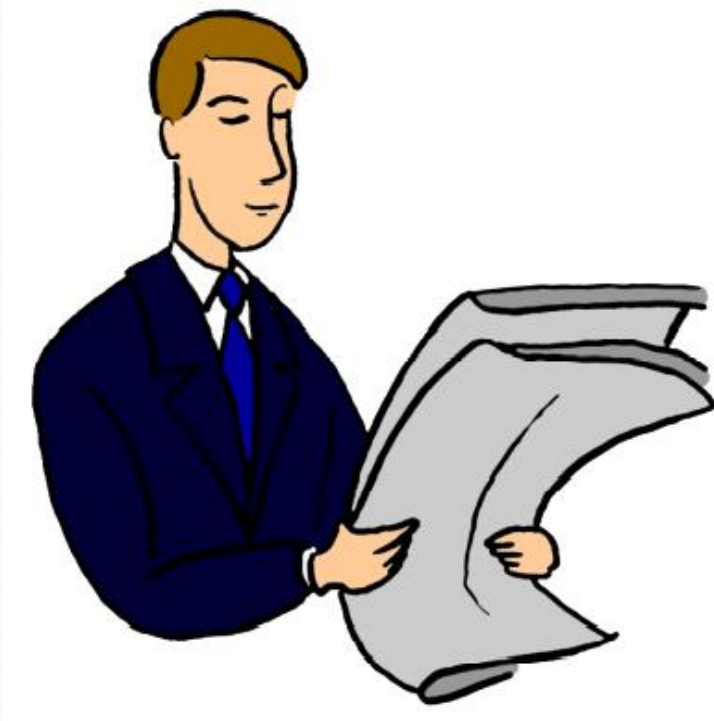


خارطة الطريق لليوم Today's Roadmap

- 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!) (دقة في البحث وسرعة)

Retrieval Models

BOOLEAN RETRIVAL



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?**

The goal is

جعل طريقة البحث أكثر فاعلية more effective
وأكثر سرعة more efficient

Boolean Retrieval Model

- يهمننا في النموذج طريقة تمثيل الاستعلام والوثائق
 - الاستعلام يكتب كتعبير بولياني
 - الوثائق تمثل بمجموعة من BOW
- **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:
 - d1: He likes to play, he likes to eat
 - d2: He likes to eat, and eat, and eat
 - d3: The thing he likes to eat is apple
 - d4: The apple he likes to eat is red
 - d5: He likes to play, and eat red apple

	d1	d2	d3	d4	d5
he	1	1	1	1	1
likes	1	1	1	1	1
to	1	1	1	1	1
play	1	0	0	0	1
eat	1	1	1	1	1
and	0	1	0	0	1
the	0	0	1	1	0
thing	0	0	1	0	0
is	0	0	1	1	0
apple	0	0	1	1	1
red	0	0	0	1	1

Term-Document Incidence Matrix

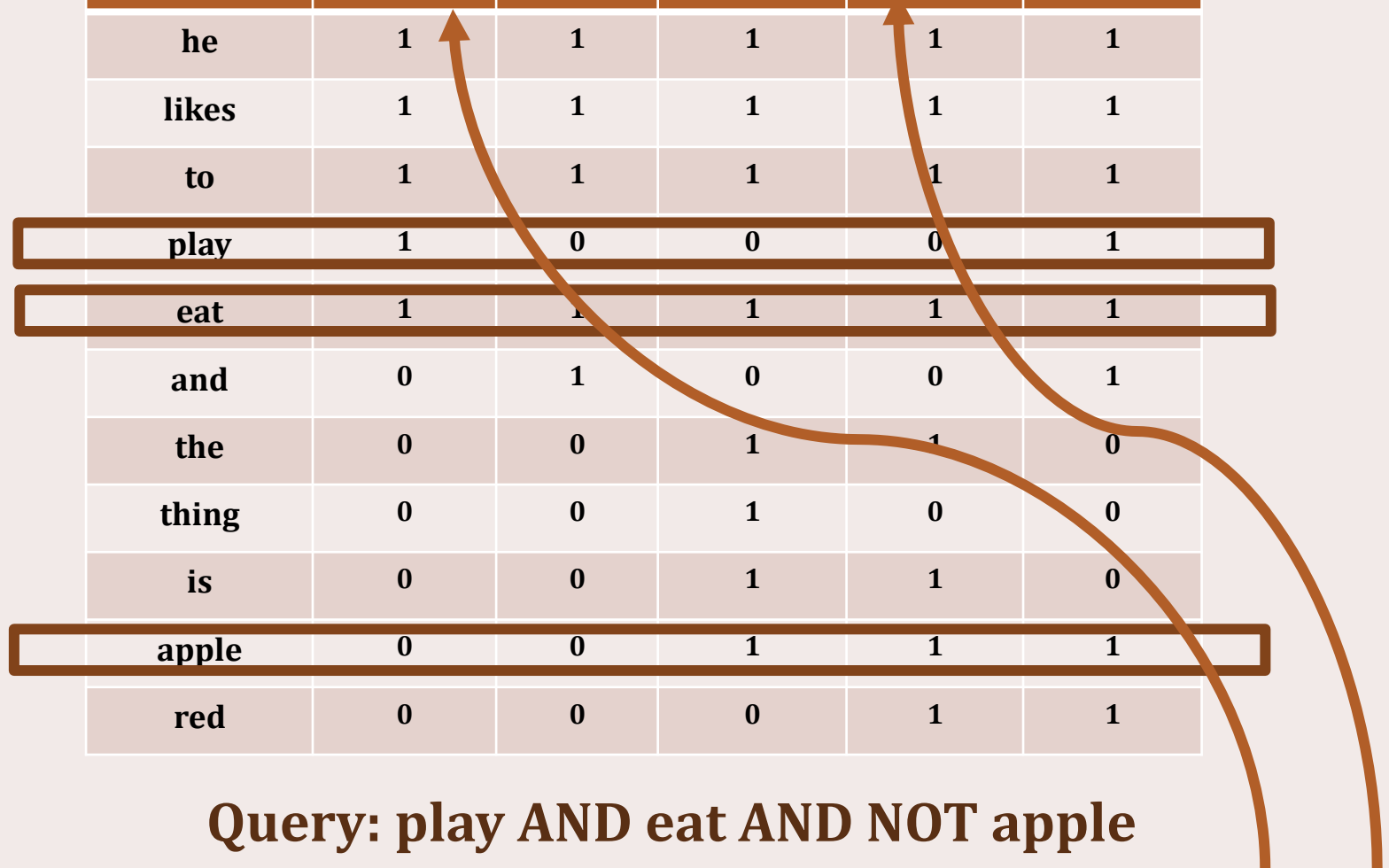
The diagram illustrates a Term-Document Incidence Matrix. The matrix is a 10x6 grid. The columns are labeled d1, d2, d3, d4, and d5, representing documents. The rows are labeled with terms: he, likes, to, play, eat, and, the, thing, is, apple, and red. A green bracket on the left groups the rows under the label 'TERMS'. A green bracket at the top groups the columns under the label 'Documents'. An orange arrow points from the text '1 if document contains term, 0 otherwise' to the cell containing the value 1 for the term 'the' in document d3.

	Documents				
	d1	d2	d3	d4	d5
he	1	1	1	1	1
likes	1	1	1	1	1
to	1	1	1	1	1
play	1	0	0	0	1
eat	1	1	1	1	1
and	0	1	0	0	1
the	0	0	1	1	0
thing	0	0	1	0	0
is	0	0	1	1	0
apple	0	0	1	1	1
red	0	0	0	1	1

1 if document contains term,
0 otherwise

Term-Document Incidence Matrix

	d1	d2	d3	d4	d5
he	1	1	1	1	1
likes	1	1	1	1	1
to	1	1	1	1	1
play	1	0	0	0	1
eat	1	1	1	1	1
and	0	1	0	0	1
the	0	0	1	1	0
thing	0	0	1	0	0
is	0	0	1	1	0
apple	0	0	1	1	1
red	0	0	0	1	1



Query: play AND eat AND NOT apple

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...

all these words:

this exact word or phrase:

any of these words:

none of these words:

numbers ranging from:

to

To do this in the search box.

Type the important words: tri-colour rat terrier

Put exact words in quotes: "rat terrier"

Type OR between all the words you want: miniature OR standard

Put a minus sign just before words that you don't want:
-rodent, -"Jack Russell"

Put two full stops between the numbers and add a unit of
measurement: 10..35 kg, £300..£500, 2010..2011

Bigger Collections

- Consider $N = 1$ million documents, each with about 1000 words. عدد كبير من الوثائق.
- Say there are $M = 500K$ distinct terms among these. في كل وثيقة عدد كبير من المصطلحات المميزة.
- $500K \times 1M$ matrix has half-a-trillion 0's and 1's. عدد كبير جدا من الأصفار والواحدات.
- But it has no more than one billion 1's. عدد قليل من الواحدات.
- matrix is extremely sparse. المصفوفة ستكون مبعثرة.

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”?

Thank you for Attention
Afaf Al Shalaby, Ph.D.

