THE UNIVERSITY *of* EDINBURGH

Text Technologies for Data Science

INFR11145

Introduction

Instructor:
Youssef Al Hariri

20/09/2023

slides credit to Dr Walid Magdy

Lecture Objectives

- Know about the course:
  - Topic
  - Objectives
  - Requirements
  - Format
  - Logistics

- Note:
  - No much technical content today
  - Don’t assume next lectures would be the same!

Text Technologies for Data Science

= documents, words, terms, …

≠ images, videos, music *(with no text)*

Information Retrieval

Text Classification

Text Analytics

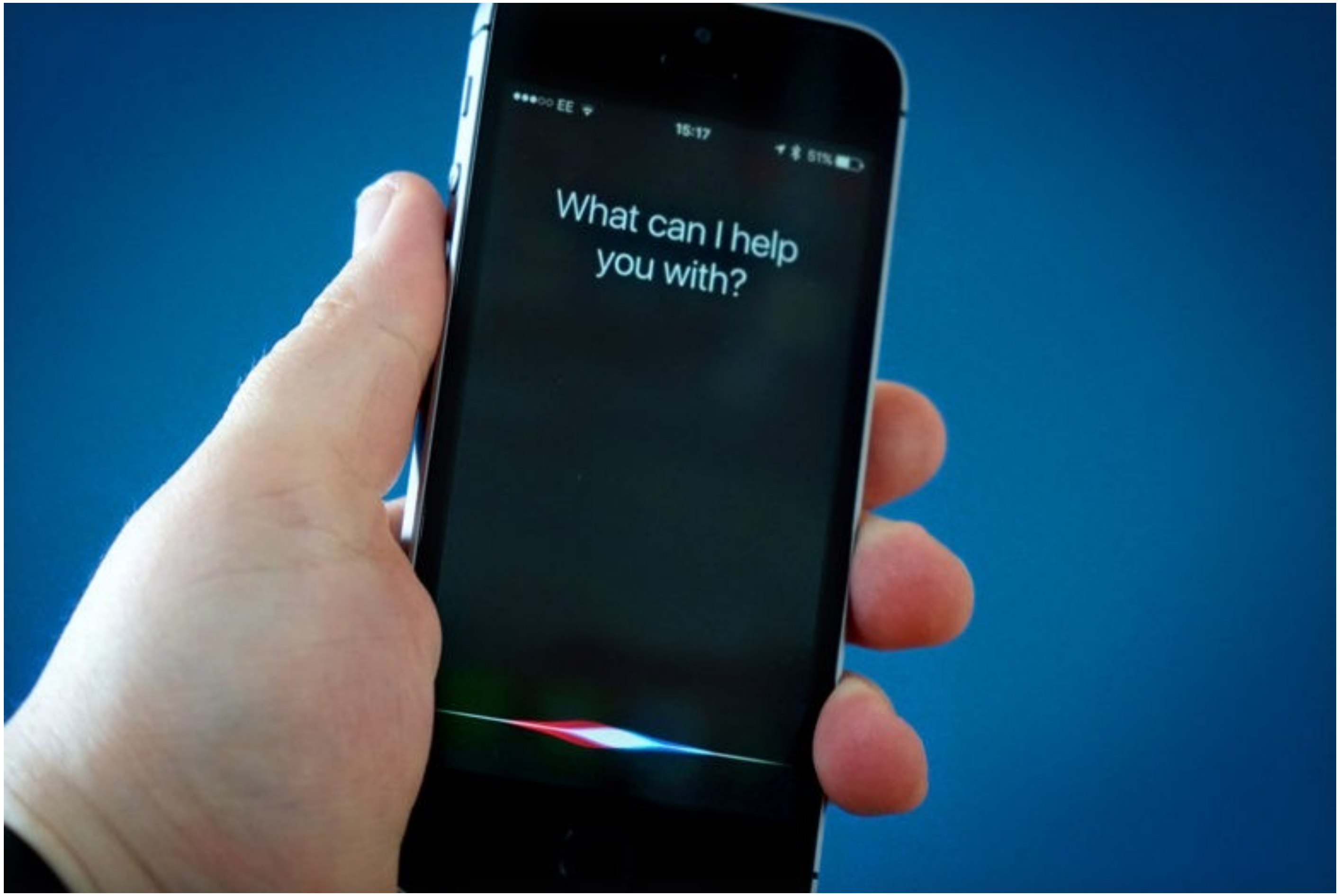
Search Engines Technologies

What is Information Retrieval (IR)?

IR is NOT just

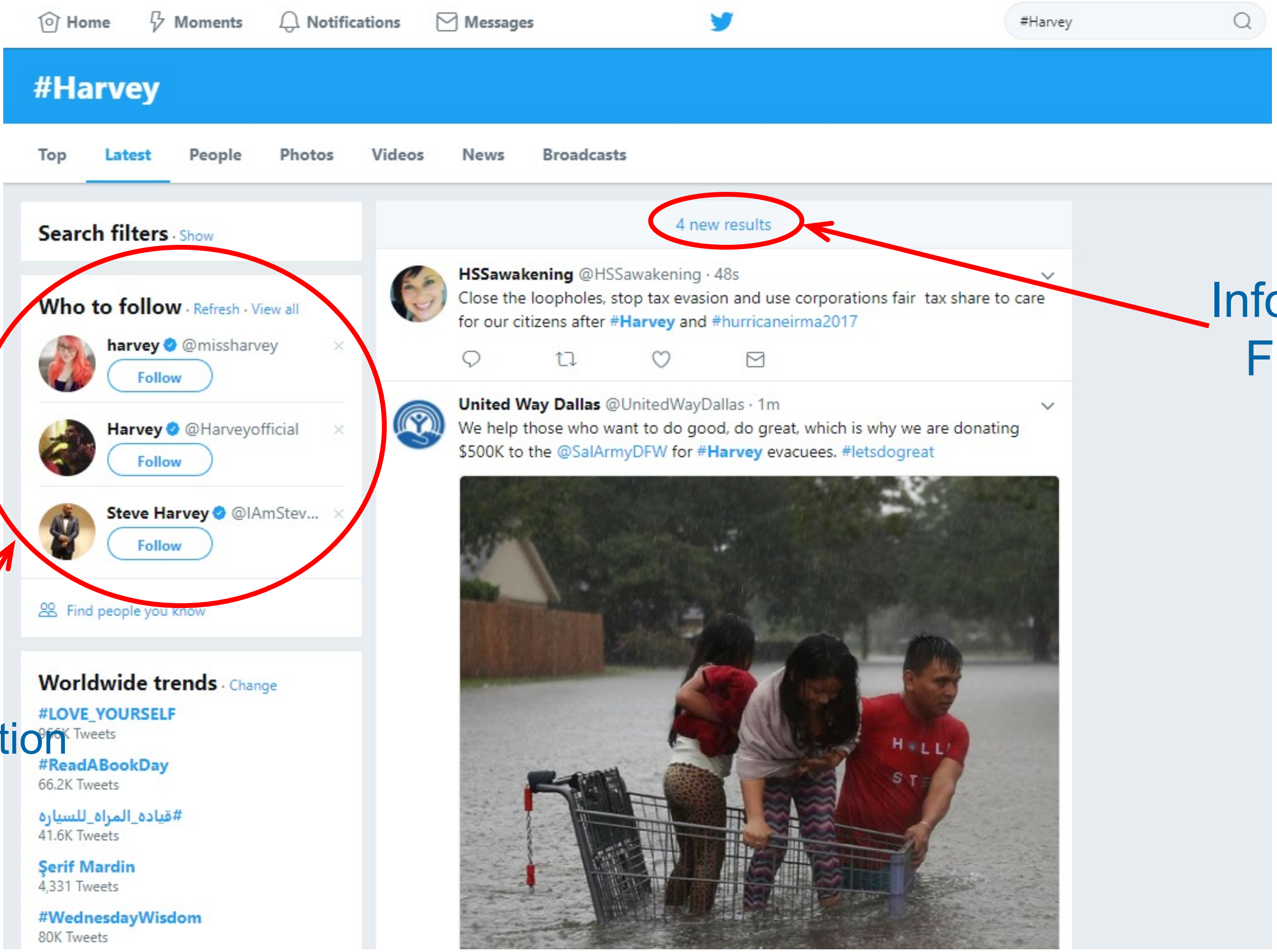
Web search

What is IR?



Speech - QA

What is IR?



Home Moments Notifications Messages

#Harvey

#Harvey

Top Latest People Photos Videos News Broadcasts

Search filters · Show

Who to follow · Refresh · View all

harvey @missharvey
Follow

Harvey @Harveyofficial
Follow

Steve Harvey @IAmStev...
Follow

Find people you know

Worldwide trends · Change

#LOVE_YOURSELF
966K Tweets

#ReadABookDay
66.2K Tweets

#قيادة_المراه_للسياره
41.6K Tweets

Şerif Mardin
4,331 Tweets

#WednesdayWisdom
80K Tweets

4 new results

HSSawakening @HSSawakening · 48s
Close the loopholes, stop tax evasion and use corporations fair tax share to care for our citizens after #Harvey and #hurricaneirma2017

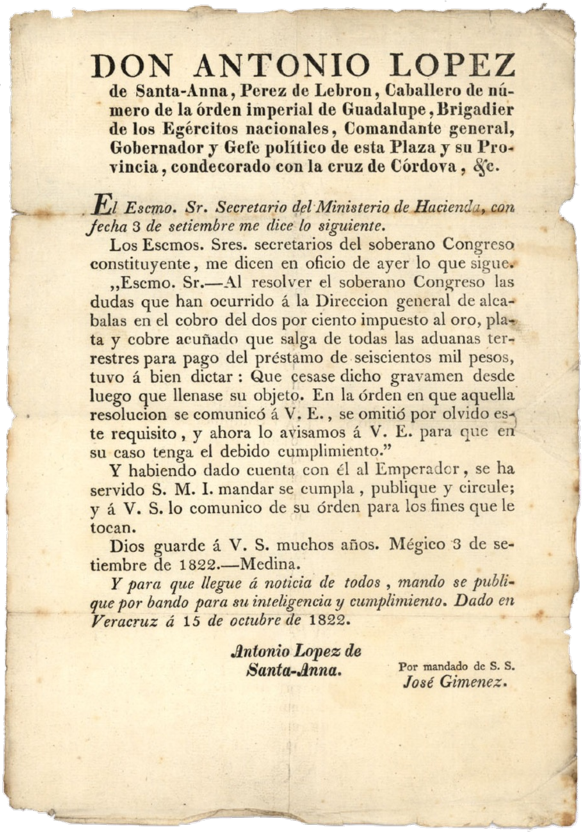
United Way Dallas @UnitedWayDallas · 1m
We help those who want to do good, do great, which is why we are donating $500K to the @SalArmyDFW for #Harvey evacuees. #letsdogreat

Information Filtering

Recommendation

Social search

What is IR?



DON ANTONIO LOPEZ

de Santa-Anna, Perez de Lebron, Caballero de número de la órden imperial de Guadalupe, Brigadier de los Egércitos nacionales, Comandante general, Gobernador y Gefe político de esta Plaza y su Provincia, condecorado con la cruz de Córdova, &c.

El Escmo. Sr. Secretario del Ministerio de Hacienda, con fecha 3 *de setiembre me dice lo siguiente.*

Los Escmos. Sres. secretarios del soberano Congreso constituyente, me dicen en oficio de ayer lo que sigue.

,,Escmo. Sr.—Al resolver el soberano Congreso las dudas que han ocurrido á la Direccion general de alcabalas en el cobro del dos por ciento impuesto al oro, plata y cobre acuñado que salga de todas las aduanas terrestres para pago del préstamo de seiscientos mil pesos, tuvo á bien dictar: Que cesase dicho gravamen desde luego que llenase su objeto. En la órden en que aquella resolucion se comunicó á V. E., se omitió por olvido este requisito, y ahora lo avisamos á V. E. para que en su caso tenga el debido cumplimiento."

Y habiendo dado cuenta con él al Emperador, se ha servido S. M. I. mandar se cumpla, publique y circule; y á V. S. lo comunico de su órden para los fines que le tocan.

Dios guarde á V. S. muchos años. Mégico 3 de setiembre de 1822.—Medina.

Y para que llegue á noticia de todos, mando se publique por bando para su inteligencia y cumplimiento. Dado en Veracruz á 15 *de octubre de* 1822.

Antonio Lopez de
Santa-Anna.

Por mandado de S. S.
José Gimenez.

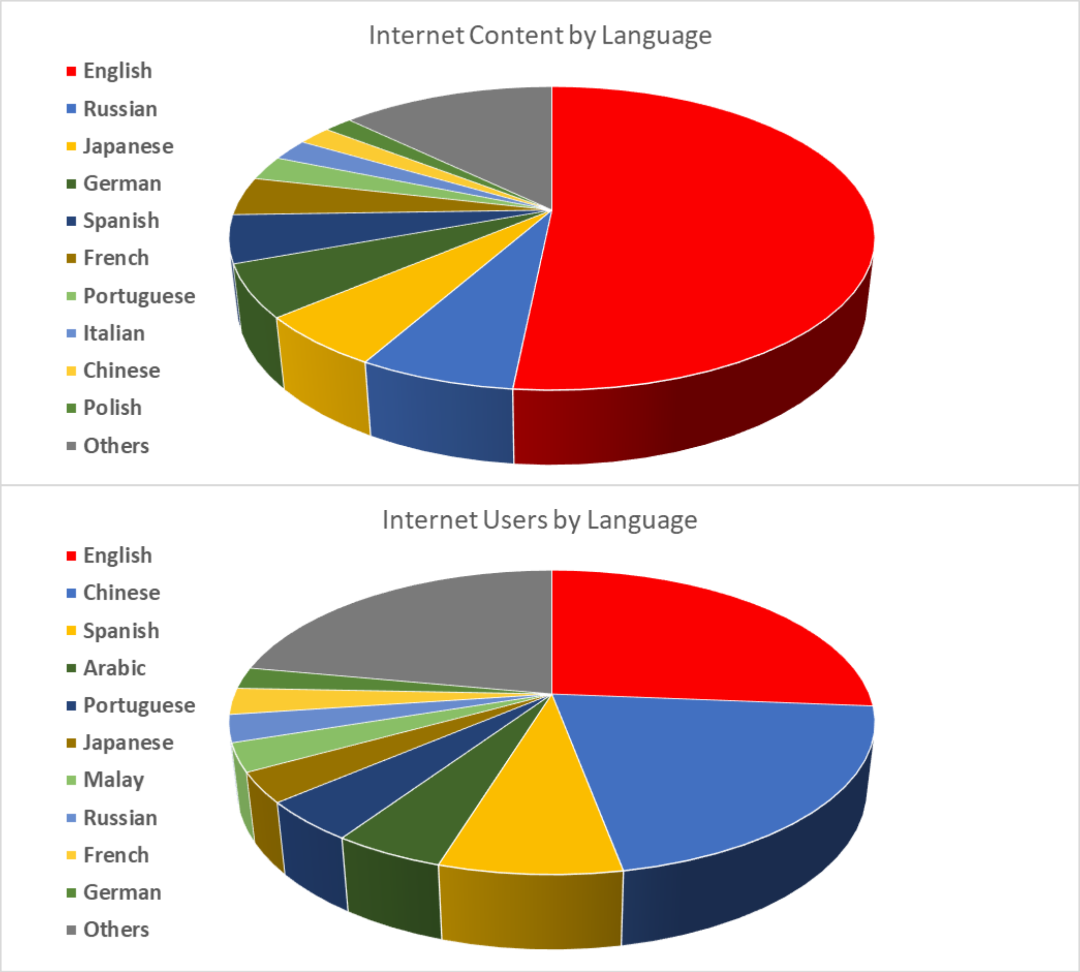
Library (book) search

1950's

What is IR?

Legal search

What is IR?



Cross-Language search

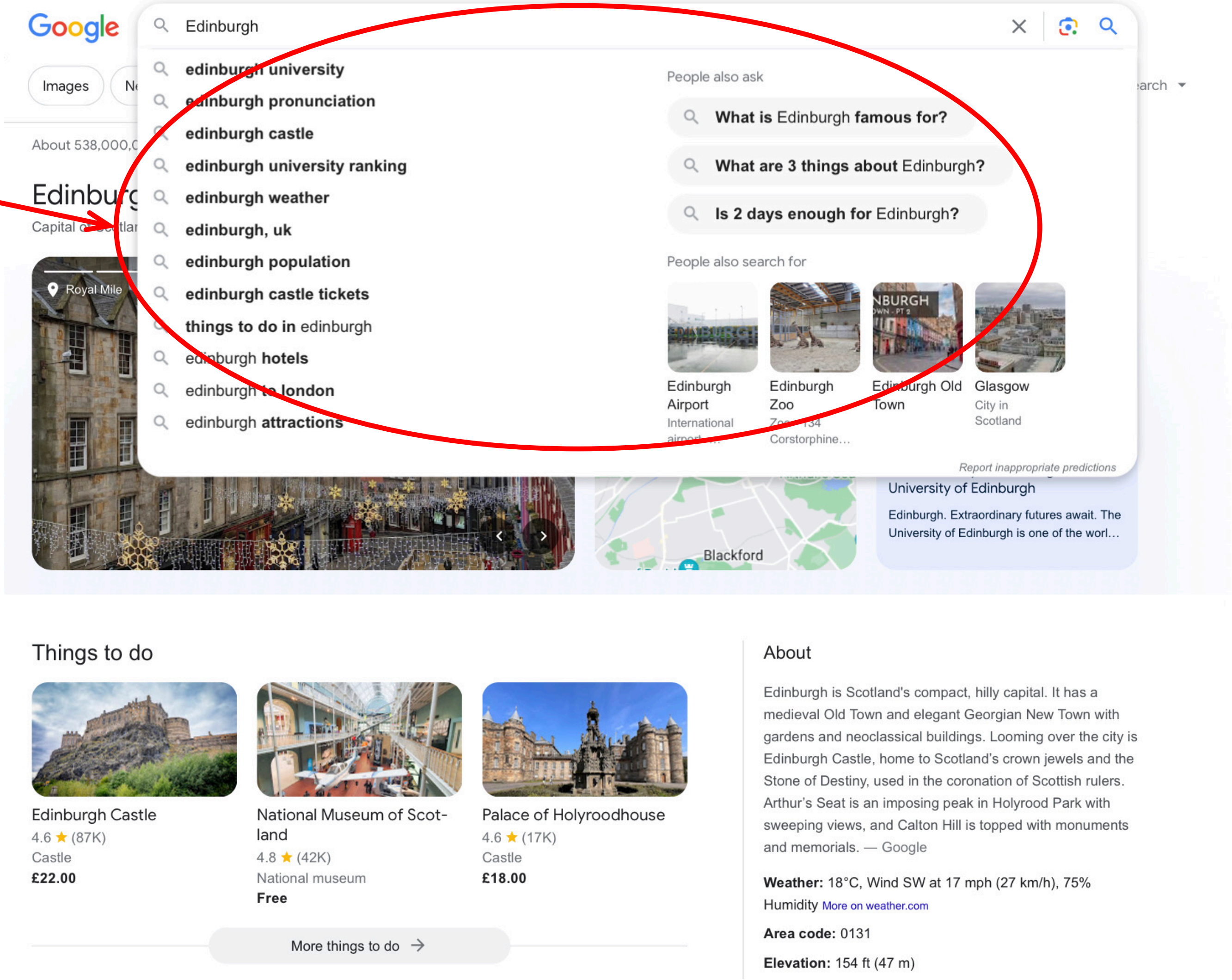
What is IR?

Content-based music search

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

What is IR?

Query suggestion / correction



Edinburgh

edinburgh university
edinburgh pronunciation
edinburgh castle
edinburgh university ranking
edinburgh weather
edinburgh, uk
edinburgh population
edinburgh castle tickets
things to do in edinburgh
edinburgh hotels
edinburgh to london
edinburgh attractions

People also ask

What is Edinburgh famous for?
What are 3 things about Edinburgh?
Is 2 days enough for Edinburgh?

People also search for

Edinburgh Airport — International airport...
Edinburgh Zoo — Zoo · 134 Corstorphine...
Edinburgh Old Town
Glasgow — City in Scotland

Report inappropriate predictions

University of Edinburgh
Edinburgh. Extraordinary futures await. The University of Edinburgh is one of the worl…

Things to do

Edinburgh Castle
4.6 ★ (87K)
Castle
£22.00

National Museum of Scotland
4.8 ★ (42K)
National museum
Free

Palace of Holyroodhouse
4.6 ★ (17K)
Castle
£18.00

More things to do →

About

Edinburgh is Scotland's compact, hilly capital. It has a medieval Old Town and elegant Georgian New Town with gardens and neoclassical buildings. Looming over the city is Edinburgh Castle, home to Scotland's crown jewels and the Stone of Destiny, used in the coronation of Scottish rulers. Arthur's Seat is an imposing peak in Holyrood Park with sweeping views, and Calton Hill is topped with monuments and memorials. — Google

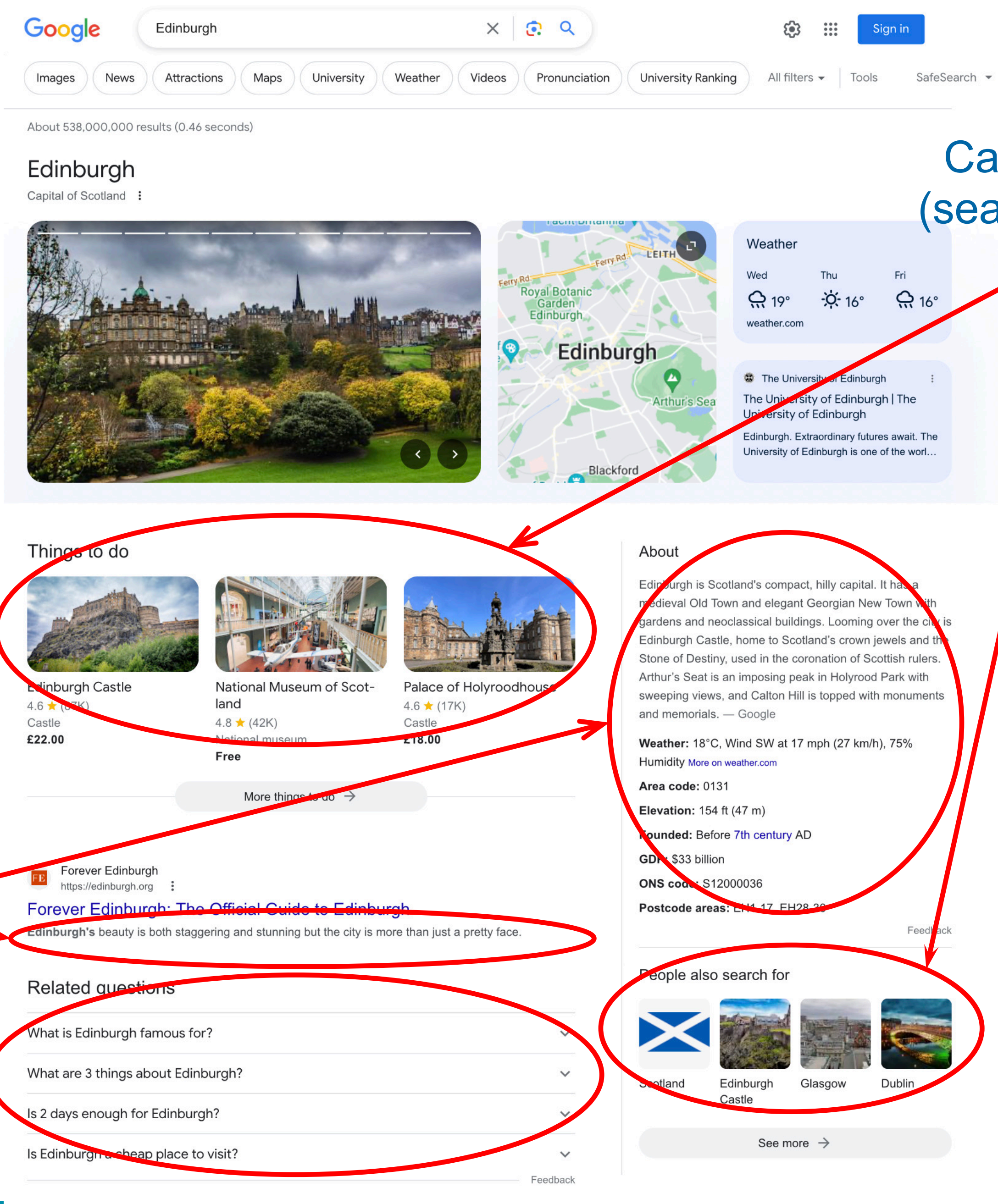
Weather: 18°C, Wind SW at 17 mph (27 km/h), 75% Humidity More on weather.com

Area code: 0131

Elevation: 154 ft (47 m)

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

What is IR?



Google | Edinburgh

About 538,000,000 results (0.46 seconds)

Edinburgh

Capital of Scotland

Weather

Wed	Thu	Fri
19°	16°	16°

weather.com

The University of Edinburgh

The University of Edinburgh | The University of Edinburgh

Edinburgh. Extraordinary futures await. The University of Edinburgh is one of the worl…

Things to do

Edinburgh Castle
4.6 ★ (87K)
Castle
£22.00

National Museum of Scotland
4.8 ★ (42K)
National museum
Free

Palace of Holyroodhouse
4.6 ★ (17K)
Castle
£18.00

More things to do →

Forever Edinburgh
https://edinburgh.org

Forever Edinburgh: The Official Guide to Edinburgh

Edinburgh's beauty is both staggering and stunning but the city is more than just a pretty face.

Related questions

- What is Edinburgh famous for?
- What are 3 things about Edinburgh?
- Is 2 days enough for Edinburgh?
- Is Edinburgh a cheap place to visit?

Feedback

About

Edinburgh is Scotland's compact, hilly capital. It has a medieval Old Town and elegant Georgian New Town with gardens and neoclassical buildings. Looming over the city is Edinburgh Castle, home to Scotland's crown jewels and the Stone of Destiny, used in the coronation of Scottish rulers. Arthur's Seat is an imposing peak in Holyrood Park with sweeping views, and Calton Hill is topped with monuments and memorials. — Google

Weather: 18°C, Wind SW at 17 mph (27 km/h), 75% Humidity More on weather.com

Area code: 0131

Elevation: 154 ft (47 m)

Founded: Before 7th century AD

GDP: $33 billion

ONS code: S12000036

Postcode areas: EH1-17, EH28-30

Feedback

People also search for

Scotland
Edinburgh Castle
Glasgow
Dublin

See more →

Categorisation (search verticals)

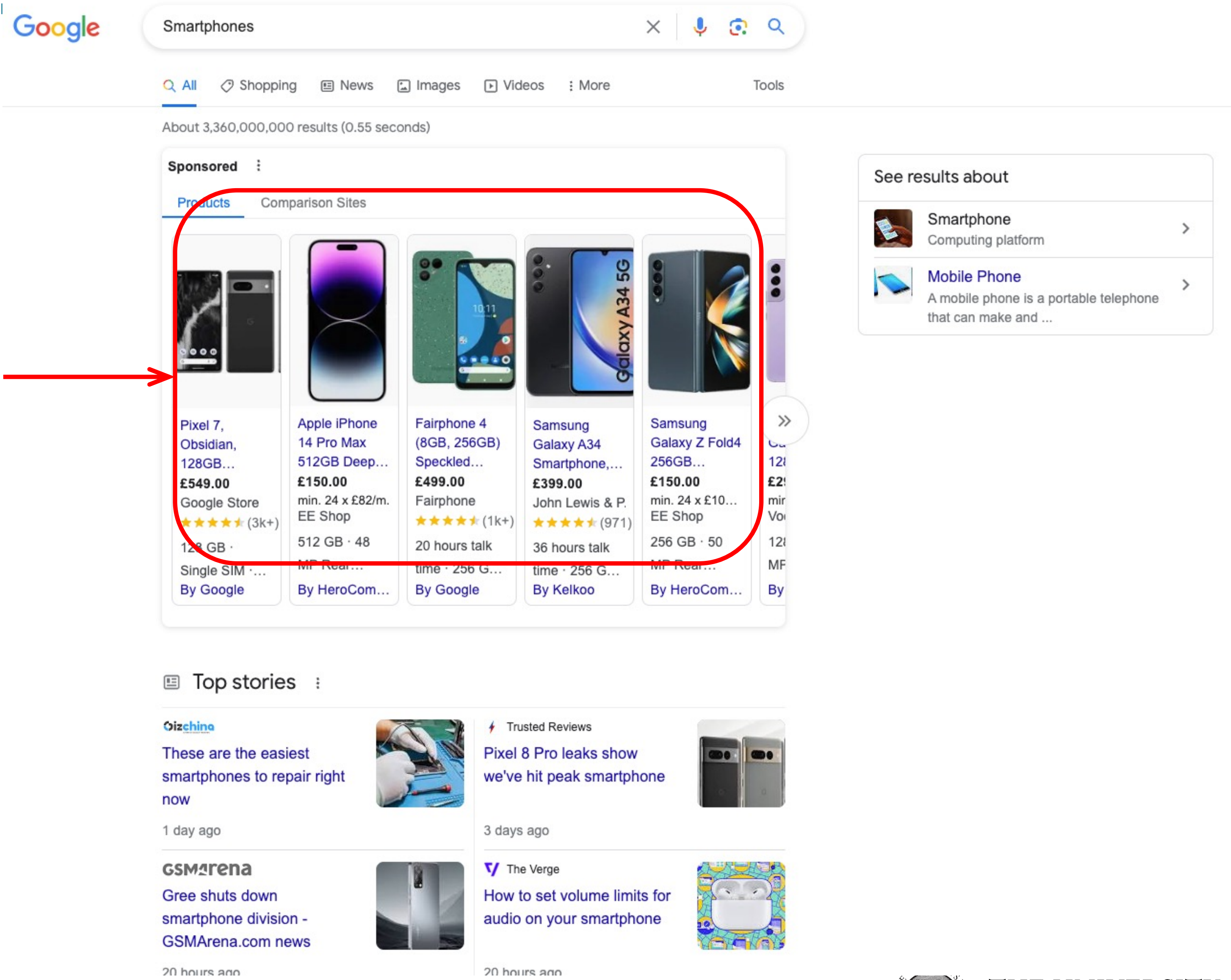
Snippet selection / summarisation

Query suggestion

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

What is IR?

Advertising



Google

Smartphones

All · Shopping · News · Images · Videos · More · Tools

About 3,360,000,000 results (0.55 seconds)

Sponsored

Products · Comparison Sites

Pixel 7, Obsidian, 128GB…
£549.00
Google Store
★★★★★ (3k+)
128 GB · Single SIM ·…
By Google

Apple iPhone 14 Pro Max 512GB Deep…
£150.00
min. 24 x £82/m.
EE Shop
512 GB · 48 MP Rear…
By HeroCom…

Fairphone 4 (8GB, 256GB) Speckled…
£499.00
Fairphone
★★★★★ (1k+)
20 hours talk time · 256 G…
By Google

Samsung Galaxy A34 Smartphone,…
£399.00
John Lewis & P.
★★★★★ (971)
36 hours talk time · 256 G…
By Kelkoo

Samsung Galaxy Z Fold4 256GB…
£150.00
min. 24 x £10…
EE Shop
256 GB · 50 MP Rear…
By HeroCom…

See results about

Smartphone
Computing platform

Mobile Phone
A mobile phone is a portable telephone that can make and …

Top stories

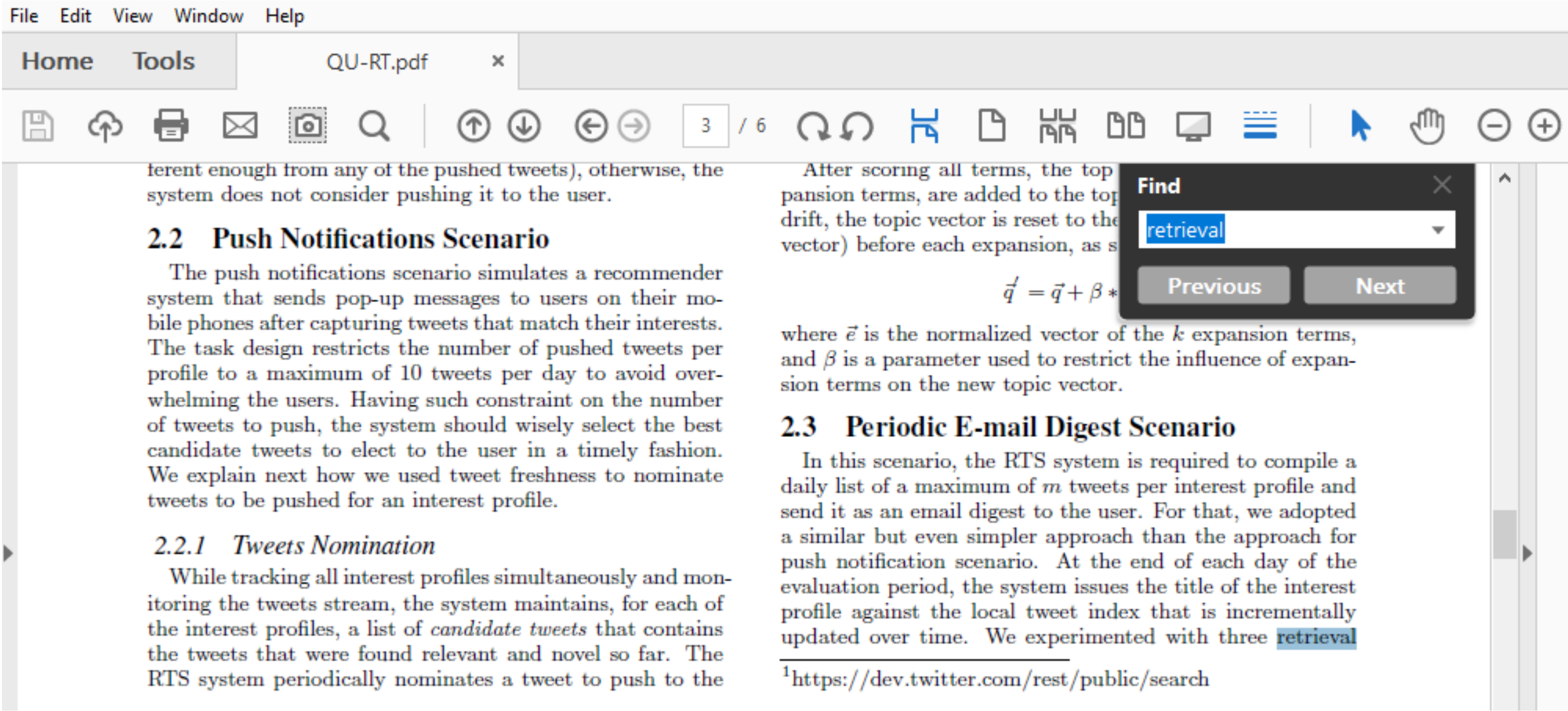
Gizchina
These are the easiest smartphones to repair right now
1 day ago

Trusted Reviews
Pixel 8 Pro leaks show we've hit peak smartphone
3 days ago

GSMArena
Gree shuts down smartphone division - GSMArena.com news
20 hours ago

The Verge
How to set volume limits for audio on your smartphone
20 hours ago

What is IR? Find?



ferent enough from any of the pushed tweets), otherwise, the system does not consider pushing it to the user.

2.2 Push Notifications Scenario

The push notifications scenario simulates a recommender system that sends pop-up messages to users on their mobile phones after capturing tweets that match their interests. The task design restricts the number of pushed tweets per profile to a maximum of 10 tweets per day to avoid overwhelming the users. Having such constraint on the number of tweets to push, the system should wisely select the best candidate tweets to elect to the user in a timely fashion. We explain next how we used tweet freshness to nominate tweets to be pushed for an interest profile.

2.2.1 *Tweets Nomination*

While tracking all interest profiles simultaneously and monitoring the tweets stream, the system maintains, for each of the interest profiles, a list of *candidate tweets* that contains the tweets that were found relevant and novel so far. The RTS system periodically nominates a tweet to push to the

After scoring all terms, the top
pansion terms, are added to the top
drift, the topic vector is reset to the
vector) before each expansion, as s

$$\vec{q}' = \vec{q} + \beta *$$

where $\vec{e}$ is the normalized vector of the k expansion terms, and β is a parameter used to restrict the influence of expansion terms on the new topic vector.

2.3 Periodic E-mail Digest Scenario

In this scenario, the RTS system is required to compile a daily list of a maximum of m tweets per interest profile and send it as an email digest to the user. For that, we adopted a similar but even simpler approach than the approach for push notification scenario. At the end of each day of the evaluation period, the system issues the title of the interest profile against the local tweet index that is incrementally updated over time. We experimented with three retrieval

[1] https://dev.twitter.com/rest/public/search

IR ≠ Find

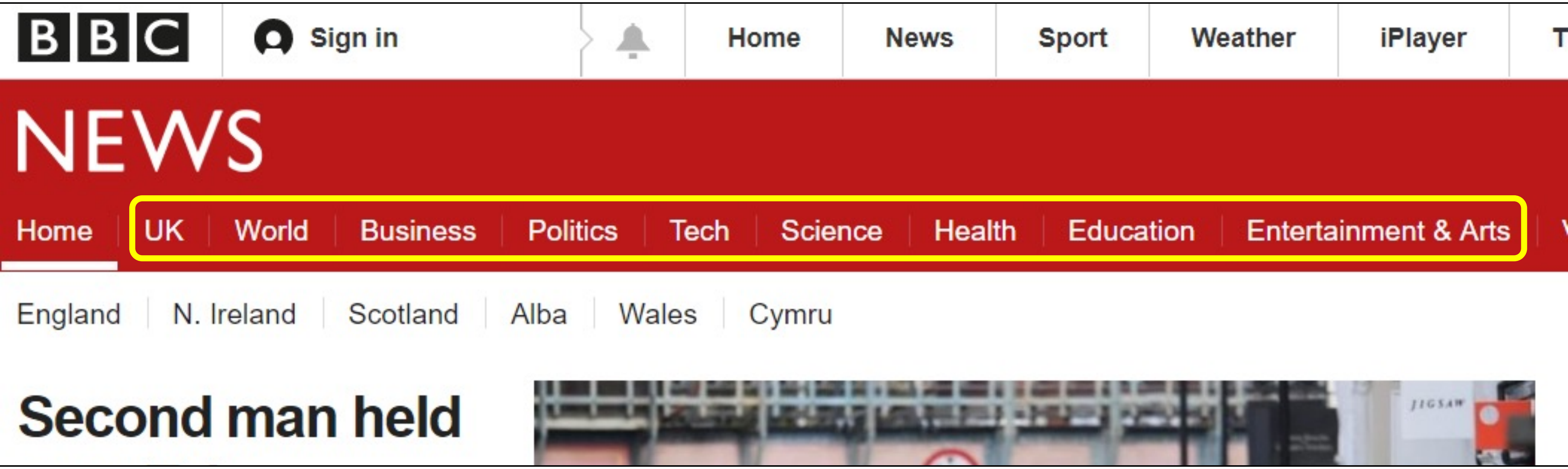
- Sequential
- Exact match

What is IR?

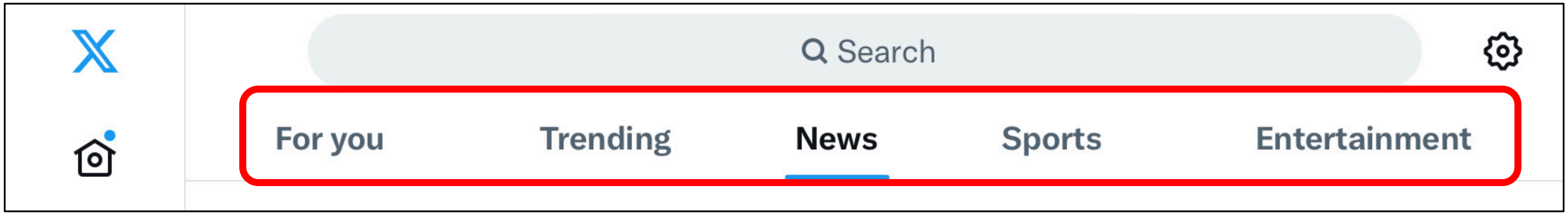
- **IR** is finding material of an unstructured nature that satisfies an information need from within large collections

- Find → Task
- Unstructured → Nature
- Information need → Target
- Satisfies → Evaluation

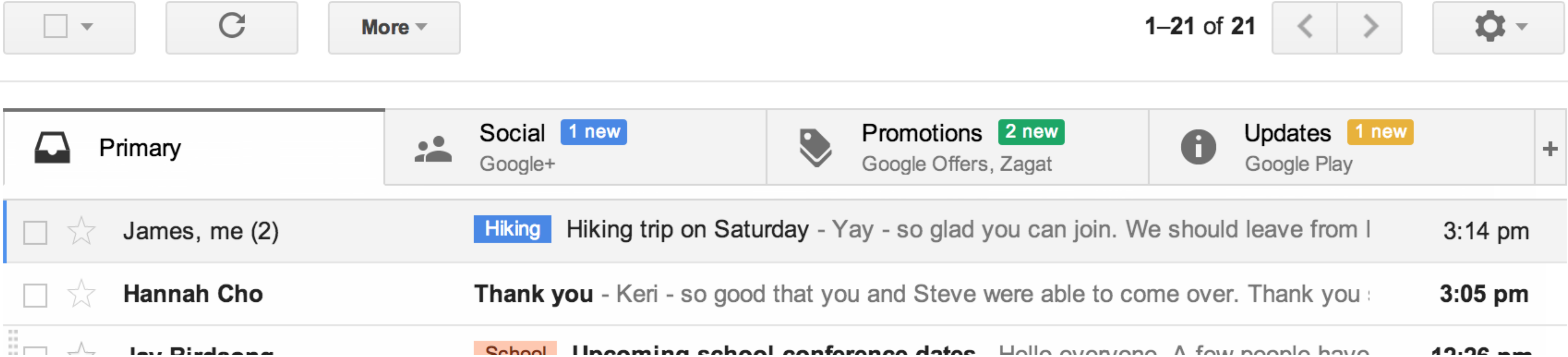
Text classification



Second man held



Text classification

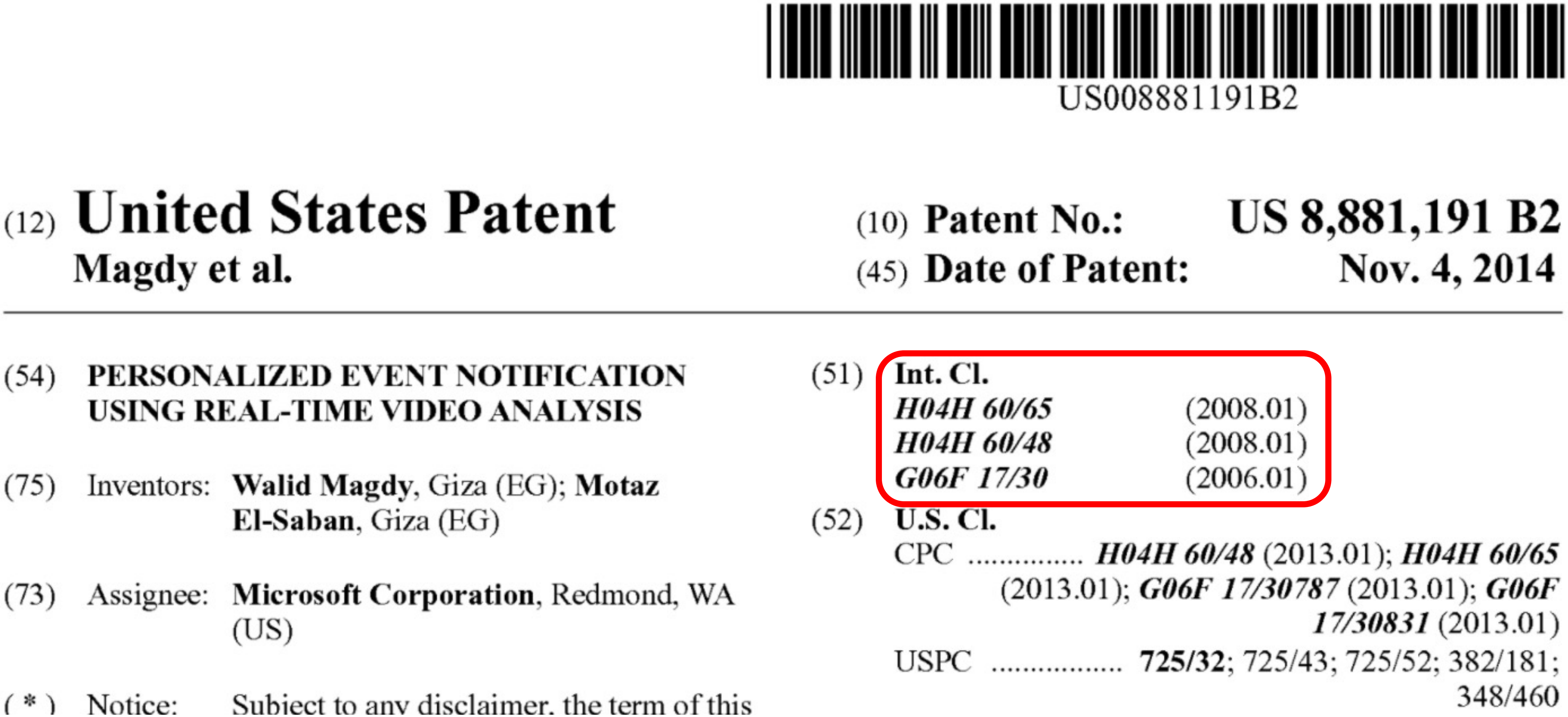


□ ☆ James, me (2) — Hiking — Hiking trip on Saturday - Yay - so glad you can join. We should leave from I — 3:14 pm

□ ☆ **Hannah Cho** — **Thank you** - Keri - so good that you and Steve were able to come over. Thank you : — **3:05 pm**

□ ☆ **Jay Birdsong** — School — **Upcoming school conference dates** - Hello everyone. A few people have — **12:26 pm**

Text classification



US008881191B2

(12) **United States Patent**
Magdy et al.

(10) **Patent No.: US 8,881,191 B2**
(45) **Date of Patent: Nov. 4, 2014**

(54) **PERSONALIZED EVENT NOTIFICATION USING REAL-TIME VIDEO ANALYSIS**

(75) Inventors: **Walid Magdy**, Giza (EG); **Motaz El-Saban**, Giza (EG)

(73) Assignee: **Microsoft Corporation**, Redmond, WA (US)

(*) Notice: Subject to any disclaimer, the term of this

(51) **Int. Cl.**

H04H 60/65	(2008.01)
H04H 60/48	(2008.01)
G06F 17/30	(2006.01)

(52) **U.S. Cl.**
CPC ***H04H 60/48*** (2013.01); ***H04H 60/65*** (2013.01); ***G06F 17/30787*** (2013.01); ***G06F 17/30831*** (2013.01)
USPC **725/32**; 725/43; 725/52; 382/181; 348/460

What is text classification?

- **Text classification** is the process of classifying documents into predefined categories based on their content.

- Input: Text (document, article, sentence)
- Task: Classify into one/multiple categories
- Categories:
  - Binary: relevant/irrelevant, spam .. etc.
  - Few: sports/politics/comedy/technology
  - Hierarchical: patents

In this course, we will learn to

- How to build a search engine
  - which search results to rank at the top
  - how to do it fast and on a massive scale
- How to evaluate a search algorithm
  - is system A really better than system B
- How to work with text
  - two tweets talk about the same topic?
  - handle misspellings, morphology, synonyms
- How to classify text
  - into categories (sports, news, comedy, …)
  - features to use
  - evaluate classification quality
- Apply text analytics
  - Find what makes a set of document different from others

How this course is different from others?

- ANLP, FNLP
  - Some text processing
  - Text laws
  - No NLP (word/phrase level vs document level)

- ML practical
  - Text classification
  - No ML (using off-the-shelf ML tool)

- It does not overlap with others on:
  - Search engines
  - IR methods/models
  - IR evaluation
  - Text analysis
  - Processing large amount of textual data

Some terms you will learn about

- Inverted index
- Vector space model
- Retrieval models: TFIDF, BM25, LM
- Page rank
- Learning to rank (L2R)
- MAP, MRR, nDCG
- Mutual information, information gain, Chi-square
- binary/multiclass classification, ranking, regression

This Course is Highly Practical

- 70% of the mark is on practical work
- You will implement **50+%** of what you learn
- By W5, you should have developed a basic working Search Engine from scratch
- Practical Lab every week
- Two coursework, mostly coding
- A course group project to develop a full system

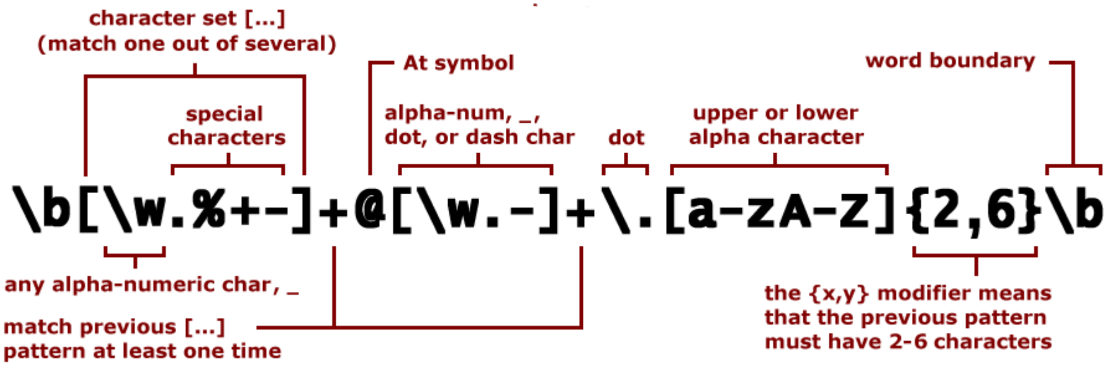
Pre-requests (1/3)

- Maths requirements:
  - Linear algebra: vectors/matrices (addition, multiplication, inverse, projections ... etc).
  - Probability theory: Discrete and continuous univariate random variables. Bayes rule. Expectation, variance. Univariate Gaussian distribution.
  - Calculus: Functions of several variables. Partial differentiation. Multivariate maxima and minima.
  - Special functions: Log, Exp, Ln.

$$\text{BM25}(D,Q) = \sum_{i=1}^{n} \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \cdot \left[\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta \right]$$

Pre-requests (2/3)

- Programming requirements:
  - Python
  - Knowledge in regular expressions
  - Shell commands (cat, sort, grep, uniq, sed, ...)
  - Data structures and software engineering for course project.
- We **DO NOT** teach coding skills in this course!
  We assume you can code!



character set [...]
(match one out of several)

special characters

At symbol

alpha-num, _, dot, or dash char

dot

upper or lower alpha character

word boundary

```
\b[\w.%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b
```

any alpha-numeric char, _

match previous [...] pattern at least one time

the {x,y} modifier means that the previous pattern must have 2-6 characters

Parse: username@domain.TLD (top level domain)

Pre-requests (3/3)

- Team-work requirement:
  - Final course project would be in groups of 5-6 students.
  - Working in a team for the project is a requirement.
  - No exceptions will be allowed!

Skills to be gained !!!

- Working with large text collections
- Few shell commands
- Some Python programming
- Software engineering skills
- Build text classifier in few mins
- TEAM WORK
  - Project management
  - Time management
  - Task assignment + system integration

Course Structure

- 20 Lectures:
  - 2 lectures → Introduction (today)
  - 14 lectures → IR (50% practical lectures)
  - 4 lectures → Text Analytics/Classification
- 8-10 Labs:
  - Practice what you learn
- No Tutorials
- Some self-reading
- <u>Lots of system implementation</u>
- Few online videos

Course Instructors

Youssef Al Hariri

Lecturer

(15 lectures)?

Bjorn Ross

Lecturer

(4 lectures)

+ 1 guest lecture

Lecture Format

- 2 Lectures at a time
- Questions are allowed any time. Feel free to interrupt
- 5-10 mins break after L1
    - Feel free to go out and come back
    - Discuss 1st lecture with friends
    - Questions on L1 are allowed before starting L2
- Some lectures are interactive. Please participate
- Some lectures will include demos (running code)

Labs

- How it works:
  - Relevant lab will be announced with each lecture on Wednesday
  - You should implement lab directly after lecture
  - Any issues → ask on Piazza (tag question by lab number)
  - Produced output → Share on Piazza (publicly)
  - Demonstrators → answer questions + validate your output
  - TA → answer questions about the course
  - DO NOT ask a question before checking if it was asked before
  - Tuesdays → Optional in-person labs for those still require support
- Optional in-person labs:
  - Location: AT 6.06 (TBC)
  - Times: Tuesday, 10:00, 11:00, 13:00 (TBC)
- Demonstrators:
  TBA

Lab Zero (Lab 0)

- Please check Lab 0 before next week lectures
- Lab 0 is designed for one purpose:
  Help you decide to take TTDS or not
- Lab content:
  - Read a text file word by word, lower-case letters, print
  - Count the number of occurrence of few words
- If Lab 0 challenging →
  → Probably, TTDS would be very challenging to you
  → You will need much extra effort to implement labs and CW
  → Think wisely before you decide to take the course

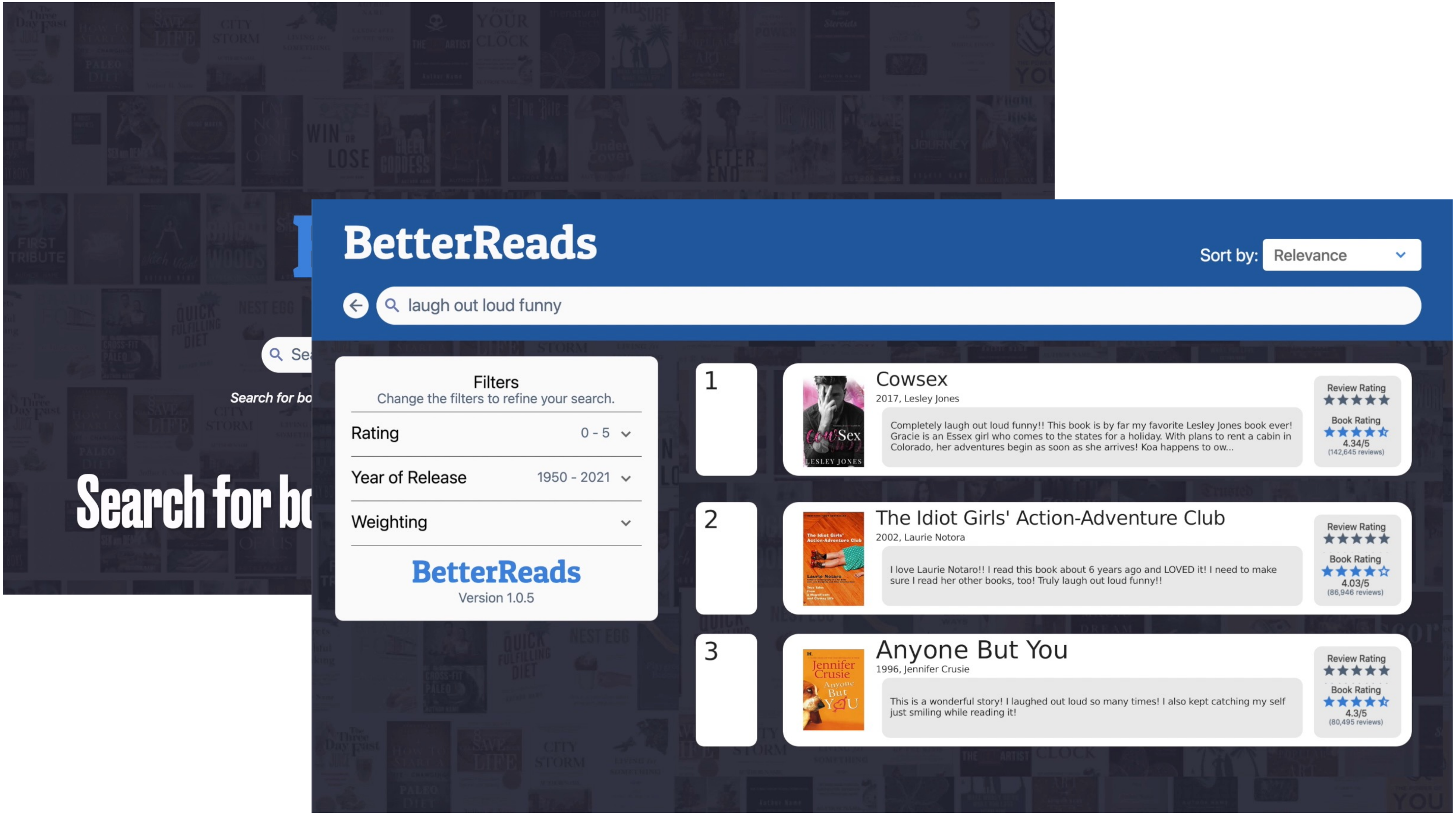
Assessments

- Coursework 1: **10%**
  The same as labs 1-3 → Build your first search engine
- Coursework 2: **20%**
  IR Evaluation, Text classification/analytics
- Group project: **40%**
  A full running search engine supported by text technologies
- Final Exam: **30%**

Group Project

- The largest weight: 40% of the total mark
- Teamwork → Group 5-6 (you select your own group)
- Design a full end-to-end search engine that searches a large collection of documents with many functionalities.
- Mark = $\text{Mark}_{\text{project}}$ x $\text{weight}_{\text{individual}}$
  - $\text{Mark}_{\text{project}}$ → the same for all team members
    - How complete/effective/fast/nice is your search engine?
  - $\text{weight}_{\text{individual}}$ → weight for individual contribution.
    - ranges from 0 to 1. It should be 1.0 by default but can be different for each member according to their contribution.
- Project prize → a prize will be awarded to best project

Example: BetterReads



BetterReads

Sort by: Relevance

laugh out loud funny

Filters
Change the filters to refine your search.

Rating	0 - 5
Year of Release	1950 - 2021
Weighting	

BetterReads
Version 1.0.5

1 — **Cowsex**
2017, Lesley Jones

Completely laugh out loud funny!! This book is by far my favorite Lesley Jones book ever! Gracie is an Essex girl who comes to the states for a holiday. With plans to rent a cabin in Colorado, her adventures begin as soon as she arrives! Koa happens to ow...

Review Rating ★★★★★
Book Rating ★★★★☆
4.34/5
(142,645 reviews)

2 — **The Idiot Girls' Action-Adventure Club**
2002, Laurie Notora

I love Laurie Notaro!! I read this book about 6 years ago and LOVED it! I need to make sure I read her other books, too! Truly laugh out loud funny!!

Review Rating ★★★★★
Book Rating ★★★★☆
4.03/5
(86,946 reviews)

3 — **Anyone But You**
1996, Jennifer Crusie

This is a wonderful story! I laughed out loud so many times! I also kept catching my self just smiling while reading it!

Review Rating ★★★★★
Book Rating ★★★★☆
4.3/5
(80,495 reviews)

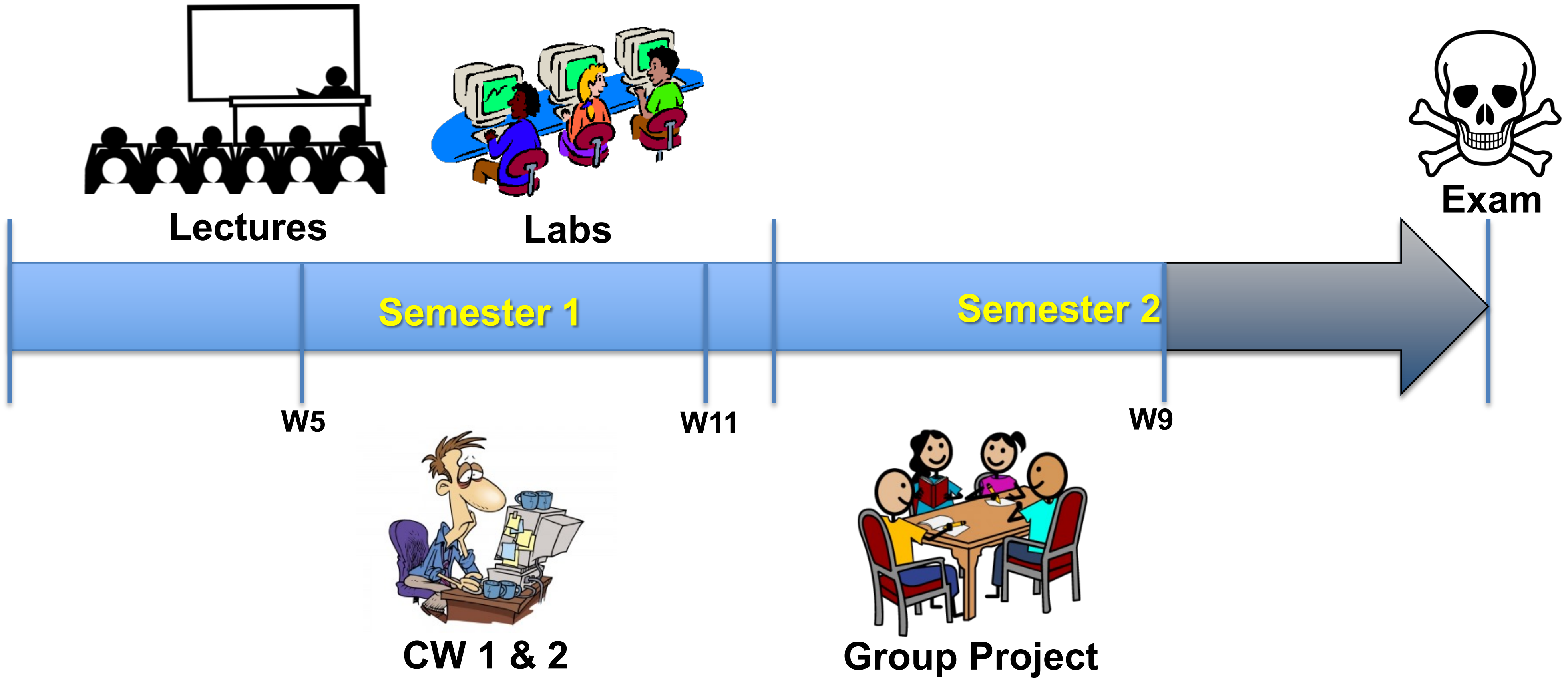
Example: BetterReads

- 11.5M Book reviews from Good reads
- Average query time: 1 secs
- New reviews are crawled and indexed automatically every day
- Ranking: Relevance + Sentiment
- Engine hosted on Google cloud compute

- *Note: we will provide credit to Google cloud to host your engine*

Timeline

- 2 Semesters (or one?)



Lectures

Labs

Exam

Semester 1

Semester 2

W5

W11

W9

CW 1 & 2

Group Project

Logistics

- Lectures:
  - Two lectures on Wednesdays, 15.00-17.00
  - Recording will be available
  - Handouts to be posted on the day of the lecture
- Course webpage:
  - Link: https://opencourse.inf.ed.ac.uk/ttds/
  - Handouts, Labs, CW details
- Learn:
  - Lecture recordings
  - Deadlines
- Note: all course materials are made public, including recordings. Feel free to share with anyone interested.

Piazza

- All communication will be there
- Questions about lectures/labs/CW are there
- Feel free to answer each other questions
- Lab support will be mainly there
- Please share your lab answers there
- Tag each question/post by its relevant topic (lab, CW … etc)

- Join NOW: link

FAQ

- How the project would be managed? What if one member does not work?
- I am not that solid in programming, should I take this course?
- Can I audit the course?
- Anything else?

Next Lecture

- Definitions of IR main concepts (more introduction)

Credits

- These slides are originally created by previous lecturers Walid Magdy and Bjorn Ross.