THE UNIVERSITY of EDINBURGH

Text Technologies for Data Science

INFR11145

Introduction

Instructor:
Walid Magdy

18-Sep-2024

1

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!

2

Text Technologies for Data Science

= documents, words, terms, …

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

Information Retrieval

Text Classification

Text Analytics

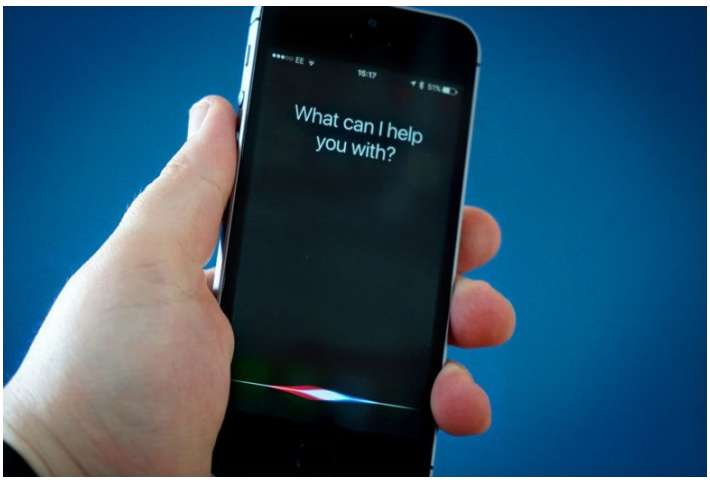
Search Engines Technologies

What is Information Retrieval (IR)?

IR is NOT just

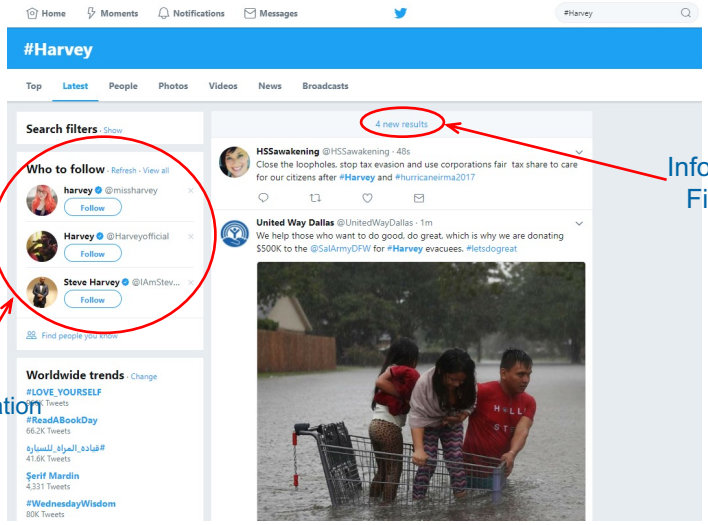
Web search

What is IR?



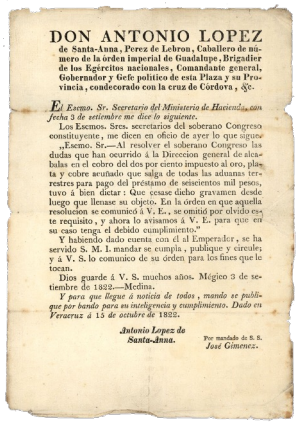
Speech - QA

What is IR?



Social search

What is IR?



DON ANTONIO LOPEZ
de Santa-Anna, Perez de Lebron, Caballero de número de la orden 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. México 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.

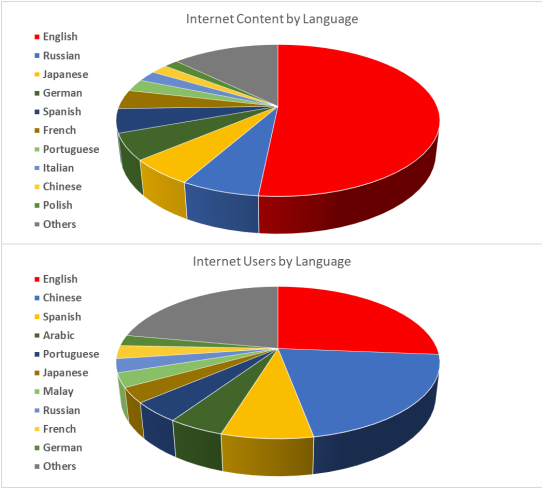
Library (book) search

1950's

What is IR?

Legal search

What is IR?



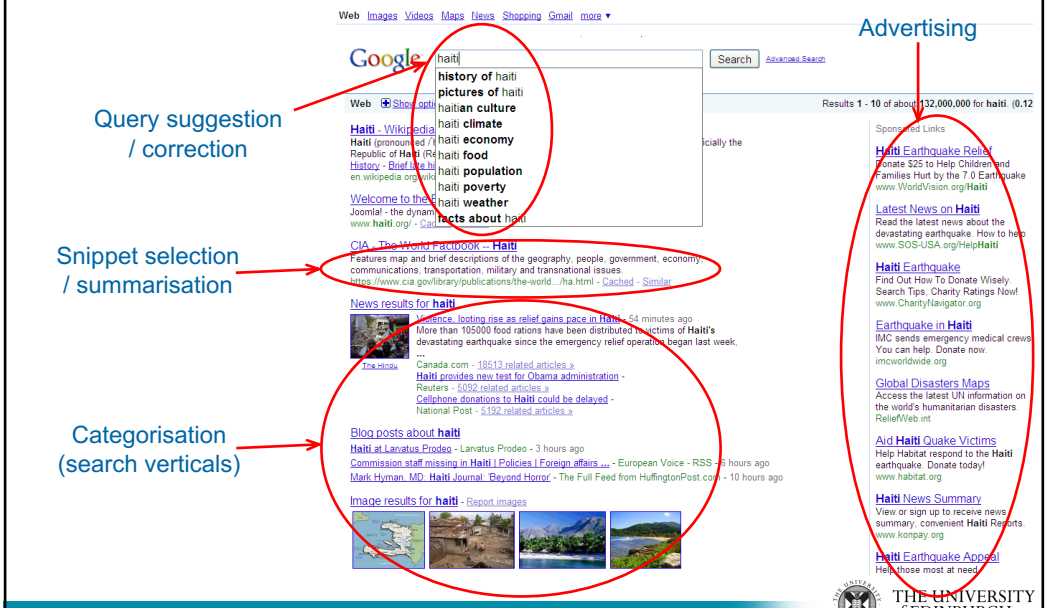
Cross-Language search

What is IR?

Content-based music search

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

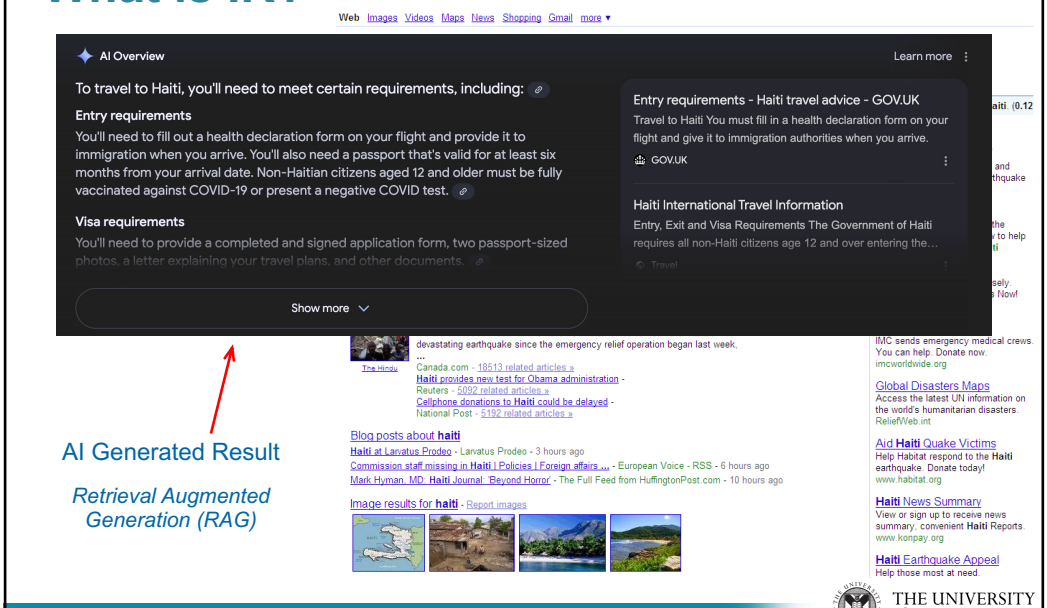
What is IR?



11

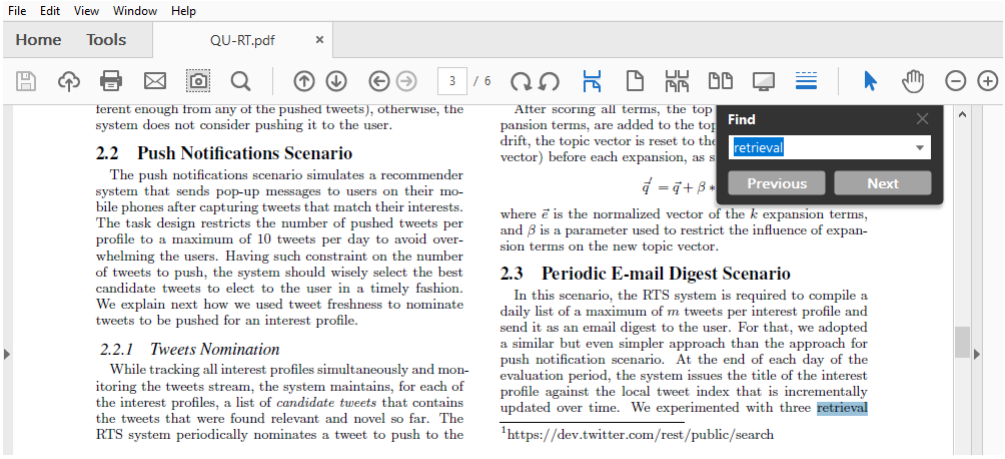
*Source: Matt Lease (IR Course at U Texas)

What is IR?



12

What is IR? Find?



IR ≠ Find

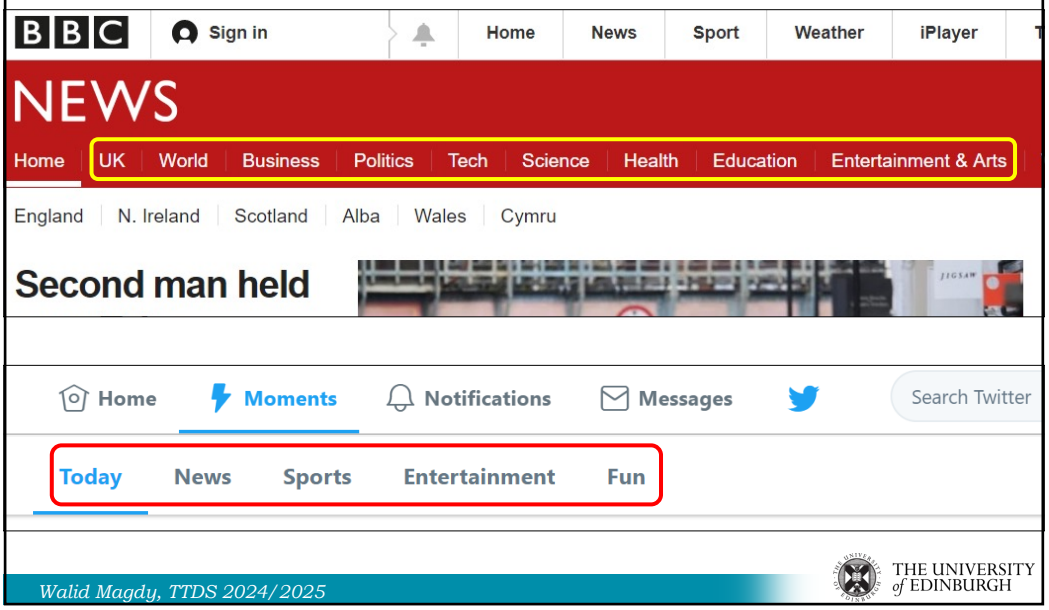
- Sequential
- Exact match

What is IR?

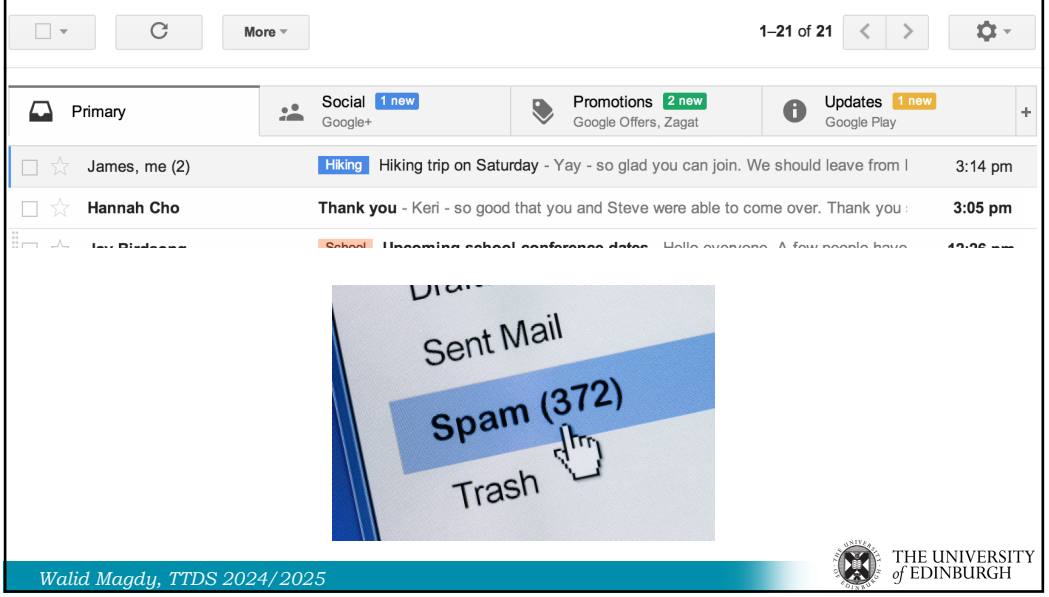
- **IR** is <u>finding</u> material of an <u>unstructured</u> nature that <u>satisfies</u> an <u>information need</u> from within large collections

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

Text classification



Text classification



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, …)
  - evaluate classification quality
- Apply text analytics
  - Find what makes a set of document different from others
- RAG systems (LLMs with IR)

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 <u>not</u> 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
- RAG

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)

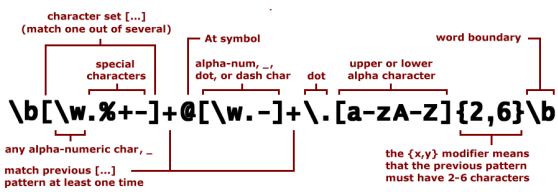
- Maths requirements:
  - Linear algebra: vectors/matrices (addition, multiplication, inverse, projections ... etc).
  - Probability theory: Discrete and continuous univariate random variables. Bayes rule. Expectation, variance. Univariate Gaussian distribution.
  - Calculus: Functions of several variables. 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)

python

- 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!



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/analyser in few mins
- TEAM WORK
  - Project management
  - Time management
  - Task assignment + system integration

Course Structure

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

Course Instructors

Walid Magdy
Reader
(11 lectures)

Bjorn Ross
Lecturer
(5 lectures)

Tj Elmas
Lecturer
(3 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
  - Mind teaser math problem (for fun)
- 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
  - 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
  - Times: Tuesday, 11am, 12pm
- Demonstrators:
  Wendy Zheng and Zahra Bokaee

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 →
  → 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 (<u>you select your own group</u>)
- 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

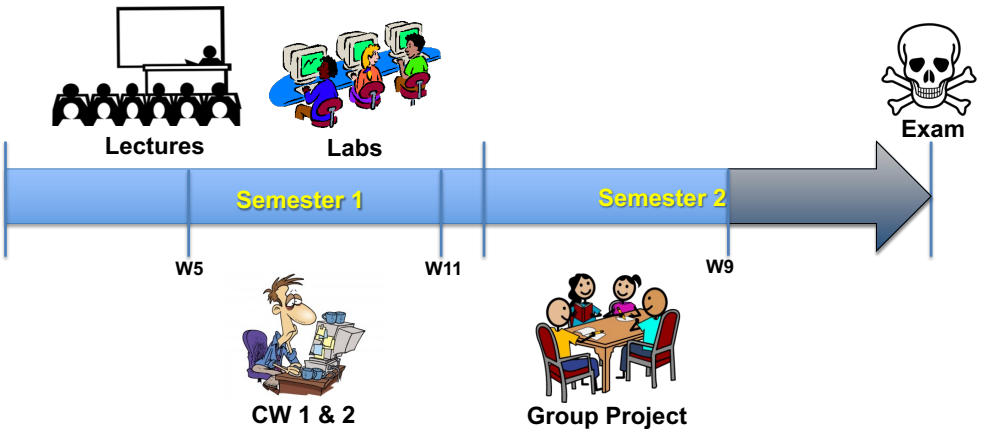
Example: BetterReads

- 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?)



Logistics

- Lectures:
  - Two lectures on Wednesdays, 15.00-17.00
  - Recording will be available on learn
  - 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)