



THE UNIVERSITY
of EDINBURGH

Text Technologies for Data Science

INFR11145

Introduction

Instructor:
Walid Magdy

23-Sep-2026

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!

Walid Magdy, TTDS 2026/2027



THE UNIVERSITY
of EDINBURGH

2

Text Technologies for Data Science



Walid Magdy, TTDS 2026/2027

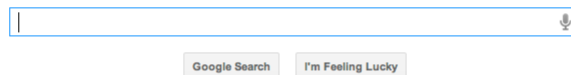


3

What is Information Retrieval (IR)?

IR is **NOT** just

Google



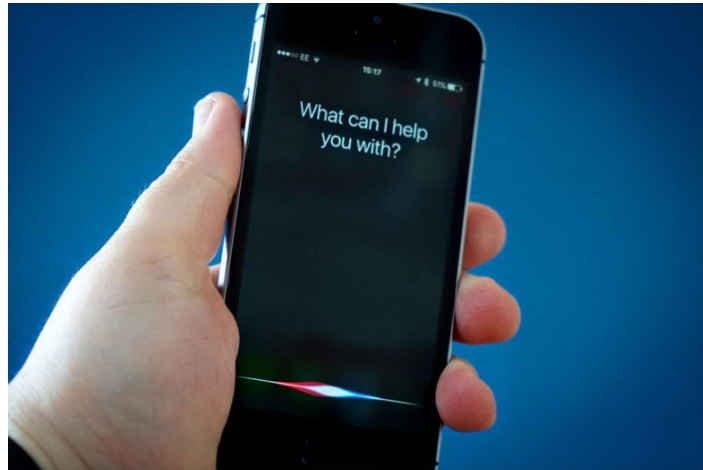
Web search

Walid Magdy, TTDS 2026/2027



4

What is IR?



Speech - QA

Walid Magdy, TTDS 2026/2027



5

What is IR?

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 66.2K Tweets
- #ReadABookDay 41.6K Tweets
- فقدوا المرأة للسياحة 41.6K Tweets
- Serif Mardin 4,331 Tweets
- #WednesdayWisdom 80K Tweets

4 new results

HSSawakening @HSSawakening · 40s
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

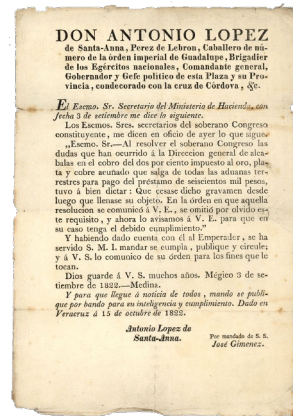
Social search

Walid Magdy, TTDS 2026/2027



6

What is IR?



Library (book) search 1950's

Walid Magdy, TTDS 2026/2027



7

What is IR?



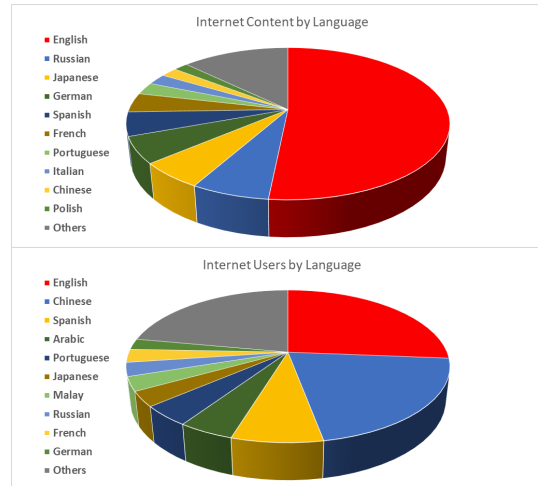
Legal search

Walid Magdy, TTDS 2026/2027



8

What is IR?



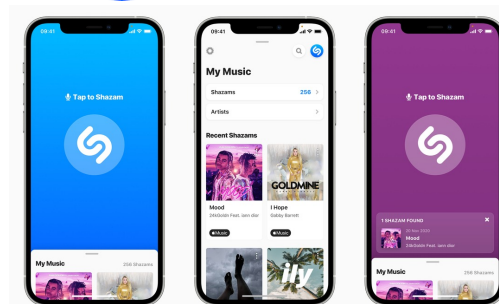
Cross-Language search

Walid Magdy, TTDS 2026/2027



9

What is IR?



Content-based music search

Walid Magdy, TTDS 2026/2027



10

What is IR?

Query Suggestion

Categorisation (search verticals)

Query correction

AI Generated Result

Retrieval Augmented Generation (RAG)

Snippet selection / summarisation

Highlights / Ads

Walid Magdy, TTDS 2026/2027

THE UNIVERSITY of EDINBURGH

11

Who needs IR anymore!

ChatGPT >

When did the Palestine/Israel conflict start?

There isn't one single start date.

The modern Israel–Palestine conflict is usually traced to the late 19th and early 20th centuries, when Zionist immigration to Ottoman Palestine grew alongside Palestinian Arab nationalism.

Key milestones:

- 1917: Britain's Balfour Declaration supported a "national home for the Jewish people" in Palestine.
- 1947–48: The UN proposed partition; war followed Israel's creation in 1948. About 700,000 Palestinians were displaced or fled—known as the Nakba. This is often treated as the beginning of the conflict in its current form.
- 1967: Israel captured the West Bank, East Jerusalem, and Gaza in the Six-Day War, beginning the occupation central to today's conflict.

Jewish and Arab ties to the land, of course, go back thousands of years; the dispute as a modern political conflict developed much later.

12:21

ChatGPT >

How many Palestinian children killed by Israel in the last 3 years?

Worked for 31s >

For the three years ending 23 September 2026, the best conservative, source-based answer is **more than 21,700 Palestinian children**—almost all in Gaza.

- Gaza:** UNICEF reported at least 21,289 Palestinian children killed from 7 October 2023 to 31 December 2025. It then reported 265 children killed after the October 2025 ceasefire by 19 June 2026, meaning the Gaza total had already passed roughly 21,450, with further deaths since. [UNICEF year-end report](#), [UNICEF June update](#)
- West Bank, including East Jerusalem:** UNICEF recorded 239 Palestinian children killed from 7 October 2023 to the end of 2025, plus at least 13 in the first four months of 2026 and further deaths afterward. [UNICEF West Bank update](#)

These are reported/identified conflict-related deaths; the true total may be higher because people remain missing or under rubble, and comprehensive independent verification takes time. Attribution in Gaza is generally reported in the context of Israeli military attacks, while West Bank deaths include Israeli forces and, in some cases, settlers.

12:23

Walid Magdy, TTDS 2026/2027

THE UNIVERSITY of EDINBURGH

12

Critical for Chatbots!



ChatGPT

Gemini



Claude

Walid Magdy, TTDS 2026/2027



THE UNIVERSITY
of EDINBURGH

13

What is IR? Find?

File Edit View Window Help

Home Tools QU-RT.pdf x

3 / 6

Find

retrieval

Previous Next

erent 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 k expansion terms, are added to the topic vector. If there is a significant drift, the topic vector is reset to the initial vector before each expansion, as shown in the following equation:

$$\vec{q}' = \vec{q} + \beta \cdot \vec{e}$$

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](https://dev.twitter.com/rest/public/search)

¹<https://dev.twitter.com/rest/public/search>

IR ≠ Find

- Sequential
- Exact match

Walid Magdy, TTDS 2026/2027



THE UNIVERSITY
of EDINBURGH

14

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

Walid Magdy, TTDS 2026/2027



15

Text classification

The screenshot shows the BBC News website. At the top, there's a navigation bar with 'BBC', 'Sign in', and a bell icon. Below that, a red banner says 'NEWS'. Underneath, a row of category links is highlighted with a yellow border: 'Home', 'UK', 'World', 'Business', 'Politics', 'Tech', 'Science', 'Health', 'Education', and 'Entertainment & Arts'. Below this, there are regional links: 'England', 'N. Ireland', 'Scotland', 'Alba', 'Wales', and 'Cymru'. A headline 'Second man held' is visible with a corresponding image. At the bottom, there's a navigation bar with 'Home', 'Moments', 'Notifications', 'Messages', and a 'Search Twitter' button. Below this, a row of category links is highlighted with a red border: 'Today', 'News', 'Sports', 'Entertainment', and 'Fun'.

Walid Magdy, TTDS 2026/2027



16

Text classification

The screenshot displays a Gmail interface. At the top, the text 'Text classification' is visible. Below it, the Gmail header shows a search bar, a refresh button, and a 'More' dropdown. The main area lists several email categories: 'Primary' (with a folder icon), 'Social' (Google+, 1 new), 'Promotions' (Google Offers, Zagat, 2 new), and 'Updates' (Google Play, 1 new). The inbox list shows three emails: one from 'James, me (2)' with a 'Hiking' label and a subject about a Saturday trip; one from 'Hannah Cho' with a 'Thank you' subject; and one from 'Jen Bishop' with a 'School' label and a subject about conference dates. A close-up inset shows a hand cursor clicking on the 'Spam (372)' folder, which is highlighted in blue. Other folders visible in the inset are 'Sent Mail' and 'Trash'.

Walid Magdy, TTDS 2026/2027

THE UNIVERSITY
of EDINBURGH

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

Walid Magdy, TTDS 2026/2027



THE UNIVERSITY
of EDINBURGH

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

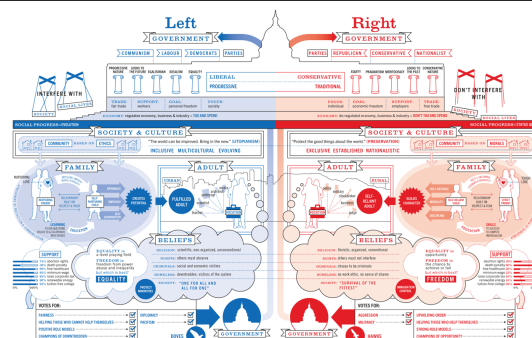
Walid Magdy, TTDS 2026/2027



19

Text Analytics

- How to compare two sets of corpora?
- What is unique about each?
- What is common?
- How to do deep analysis?
- Data driven approaches



Who are Russia's supporters?

Government responses to the war in Ukraine, March 2022

■ Condemns Russia ■ Supports Russia ■ Neutral
 ■ West-leaning ■ Russia-leaning ■ No data



Walid Magdy, TTDS 2026/2027



20

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)

Walid Magdy, TTDS 2026/2027



21

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
 - RAG systems
 - Building a full production system

Walid Magdy, TTDS 2026/2027



22

Some terms you will learn about

- Inverted index
- Retrieval models: TFIDF, BM25, LM
- Page rank
- Learning to rank (L2R)
- MAP, MRR, nDCG
- Mutual information, information gain, Chi-square
- Topic modelling: LDA, BERTopic
- Binary/Multiclass classification, ranking, regression
- Retrieval Augmented Generation (RAG)

Walid Magdy, TTDS 2026/2027



23

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

Walid Magdy, TTDS 2026/2027



24

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.

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

Walid Magdy, TTDS 2026/2027

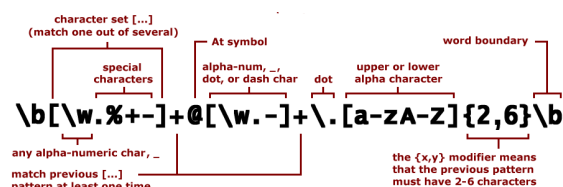


25

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!



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

Walid Magdy, TTDS 2026/2027



26

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!



Walid Magdy, TTDS 2026/2027



27

Skills to be gained !!!

- Working with large text collections
- Few shell commands
- Some Python programming
- Software engineering skills
- Build large systems at production level
- TEAM WORK
 - Project management
 - Time management
 - Task assignment + system integration

Walid Magdy, TTDS 2026/2027



28

Course Structure

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

Walid Magdy, TTDS 2026/2027



29

Course Instructors



Walid Magdy
Professor
(14 lectures)



Bjorn Ross
Reader
(6 lectures)

+ 1 potential guest lecture

Walid Magdy, TTDS 2026/2027



30

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)

Walid Magdy, TTDS 2026/2027



31

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 **EdStem** (tag question by lab number)
 - Produced output → Share on **EdStem** (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, 10am, 11pm
- Demonstrators:
Nataliya Stepanova & Zahra Bokaei

Walid Magdy, TTDS 2026/2027



32

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

Walid Magdy, TTDS 2026/2027



33

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%**

Walid Magdy, TTDS 2026/2027



34

Group Project

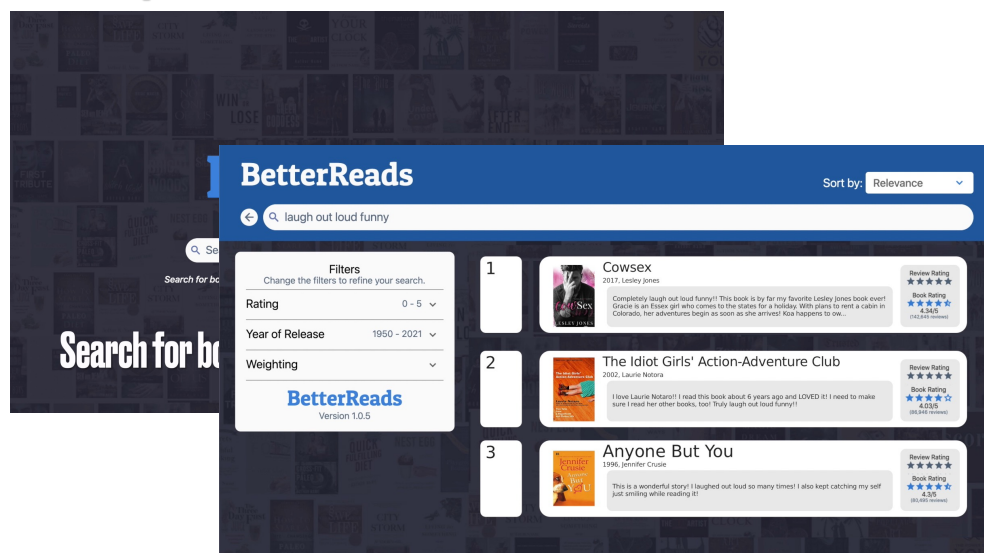
- The largest weight: **40%** of the total mark
- Teamwork → Group 5-6 (you select your own group)
- Design a full end-to-end search engine that searches a large collection of documents with many functionalities.
- Mark = $\text{Mark}_{\text{project}} \times \text{weight}_{\text{individual}}$
 - $\text{Mark}_{\text{project}}$ → 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 might be awarded to outstanding project

Walid Magdy, TTDS 2026/2027



35

Example: BetterReads



Walid Magdy, TTDS 2026/2027



36

Example: BetterReads

- 11.5M Book reviews from Good reads
- Average query time < 1 sec
- 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*

Walid Magdy, TTDS 2026/2027



37

Example: SearchFDA

FDA Device Search

Search by manufacturer, material, device name

Search Bar

Advanced Search

500+ matching devices

Sort: Relevance Show: 10

MiniMed 780G insulin pump

Matched by device name match keyword search

Manufacturer: Medtronic Minimed, Inc. | Decision Date: Jul 1, 2025

Decision: Substantially Equivalent | Panel: Clinical Chemistry | Product Code: QFG | Class: 2

EXCERPT FROM SUMMARY

Trade/Device Name: MiniMed 780G insulin pump Regulation Number: 21 CFR 880.5730

Regulation Name: Alternate Controller Enabled Infusion Pump Regulatory Class: Class II Product Code: QFG, NDC Dated: Apr... [Read more](#)

Materials: polycarbonate

Feature Signals

Clinical evidence Sterilization Biocompatibility Software lifecycle Electrical safety

Safety Signals: 1 recall

Add to comparison View Full Device Record View on FDA.gov

ACCU-CHEK INSULIN PUMP CONFIGURATION SOFTWARE...

Matched by device name match keyword search

Manufacturer: Disetronic Medical Systems AG | Decision Date: Jun 15, 2006

Filters

Panel (Category)

- General Hospital (349)
- Clinical Chemistry (107)
- Anesthesiology (39)
- Gastroenterology, Urology (1)
- Physical Medicine (1)
- Immunology (1)

Product Code

Device Class

- Class II - Moderate Risk (478)
- Class I - Low Risk (21)

Decision

- Substantially Equivalent (491)
- ST (5)
- De Novo Granted (3)

Suggestions

PCR instrument clinical chemistry analyzer

Walid Magdy, TTDS 2026/2027



38

Example: SearchFDA

Database	Records	Typical Length	Structural Characteristics
510(k) & De Novo	141 000	1–20 pages (1,000–10,000 words)	Variable structure; OCR scans often low quality with handwritten artefacts.
MAUDE	4 200 000	100 words typical; up to 3,000 words	Structured short fields plus one free-text narrative.
Recalls	9500	1–3 paragraphs (700 words)	Short descriptive paragraphs with variable format.
Device Product Classification	1700	Short structured entries (device names, product codes, CFR numbers)	Highly structured regulatory metadata.



searchfda.vercel.app

- Search around 4.5M medical FDA documents of different natures**
- Hybrid Retrieval Engine:** Combines BM25 lexical search and PubMedBERT dense vectors via RRF.
- Unified Data Model:** Merges submissions, MAUDE adverse events, and recall records.
- Material & Feature Extraction:** Detects polymers, phthalates, and regulatory signals from OCR text.
- Interactive Predicate Lineage:** Maps 102,000+ lineage nodes using citation-weighted PageRank scoring.
- Sub-Second Latency:** Memory-mapped indices serve complex queries under 200ms.

Tested by doctors on a test set while achieving over 90% precision

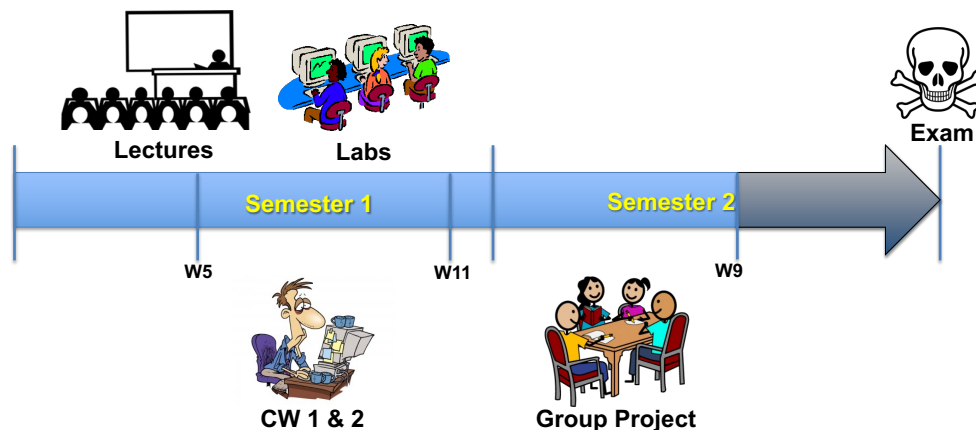
Walid Magdy, TTDS 2026/2027



39

Timeline

- 2 Semesters (or one?)



Walid Magdy, TTDS 2026/2027



40

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

Walid Magdy, TTDS 2026/2027



41

EdStem

- 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 on Learn

Walid Magdy, TTDS 2026/2027



42

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?

Walid Magdy, TTDS 2026/2027



THE UNIVERSITY
of EDINBURGH

43

Next Lecture

- Definitions of IR main concepts
(more introduction)

Walid Magdy, TTDS 2026/2027



THE UNIVERSITY
of EDINBURGH

44