



THE UNIVERSITY
of EDINBURGH

Text Technologies for Data Science

INFR11145

CW3: Group Project

Instructors
Walid Magdy & Björn Ross

26-Nov-2025

1

Group

- Members:
Min: 4, Max: 6
- Recommendation:
Look for diverse skills:
Planning, coding, interface, writing report
- Can't find 4 people?
 - Use Piazza to look for group members
 - Anyone left over at the end will be put into a group!

Magdy & Ross, TTDS 2025/2026



THE UNIVERSITY
of EDINBURGH

2

Objectives of the project

- Learn to work in teams effectively and efficiently
 - Planning
 - Work distribution
 - Issues managements
- Bring what you learnt over the course into real-life application
- Gain project management and software engineering skills
- This is **40%** of the overall course mark. Take it seriously!

Magdy & Ross, TTDS 2025/2026



3

What is Required

- Fully functional search engine built from scratch
 - Indexer
 - Search module
 - Retrieval model\
 - Interface
 - LARGE data collection
 - Real-time search
 - More?

Magdy & Ross, TTDS 2025/2026



4

Indexer/Search module

- Similar to CW1, but
- Optimized
 - Index is saved efficiently
 - Stop words there or not?
 - Stemming applied or not?
- Flexible\Scalable
 - Works well with long queries
 - Enables Free query or Boolean query
 - Has phrase/proximity search

Magdy & Ross, TTDS 2025/2026



5

Retrieval Model(s)

- Which one to select?
- Only one?
- Tfidf? Which formula? BM25?
- LM?
- New novel model optimized for you task?
- L2R?

Magdy & Ross, TTDS 2025/2026



6

Interface

- User will need interface to run the query
 - Web interface?
 - Mobile interface?
- How results will be displayed?
- Heading of document? Snippet?

Data Collection

- 100Ks or millions?
- One language or more?
- One level or more? (book vs. chapter vs page)
- Only text? Or multimedia?
- Links? PageRank?

Online/Offline system

- One-shot data collection?
- Live data collection
 - Continuous collection of data streaming and indexing
- One user at a time? Or multiuser?
- Should be hosted on server
 - Google cloud credit will be provided

Magdy & Ross, TTDS 2025/2026



9

More?

- PageRank applied for linked documents
- Classification of results
 - By genre, topic, sentiment ... etc.
- L2R?
- Query Expansion
 - Dictionary/word embedding
 - User/pseudo/implicit feedback
 - Display learnt terms with search
- Query suggestion / Spell checker
- Evaluation for the system? (topics+qrels)

Magdy & Ross, TTDS 2025/2026



10

Marking

$$\text{Mark}_{\text{final}} = \text{Mark}_{\text{project}} \times \text{weight}_{\text{individual}}$$

- **Mark_{project}:** 0 - 100% (same for all members)
 - Completeness and system working properly
 - Effectiveness/Efficiency
 - Innovation/Creativity/Features
 - Report
- **Mark_{individual}:** 0.0 - 1.0 (different for each member)
 - The amount of effort contributed to the project
 - Note: each member can be responsible on one part of the project (coding, data collection, UX, management ..)

Evaluation

- Search engine backend: 30%
- Real-life search scenario: 30%
- Innovative TTDS features: 30%
- Report: 10%
- Individual weight:
 - Worked well with team and achieved assigned tasks on time: 1.0
 - Didn't collaborate and left assigned tasks to last moment which led to lower quality of whole project: 0.2-0.8
 - Didn't contribute: 0

Eval – Search Engine backend (30%)

- Core IR functionalities
 - Index, search module, one retrieval model
- Advanced search
 - Phrase search (n words), proximity search, search by field
- Query expansion
 - RF, PRF, BERT
- Effective retrieval
 - Retrieval results are of high quality by relevance

Magdy & Ross, TTDS 2025/2026



13

Eval – Real-life Scenario (30%)

- Realistic search task
 - Solves a real problem, innovative tasks are appreciated
- Large collection of documents
 - 100Ks of large documents or 10Ms of short documents
- Speed
 - Fast retrieval in ms
- Nice interface
 - Easy to follow interface, results with snippets, query suggestion, ... etc

Magdy & Ross, TTDS 2025/2026



14

Eval – Additional Features (30%)

- Live indexing
 - Documents are continuously collected and added to index
- Classification
 - Results are classified based on a trained model
- PageRank
 - PR is calculated for links among docs in the collection
- Innovative models
 - Using advanced retrieval models, or newly developed ones (e.g. integrate recency of docs into the model), or L2R approach

Magdy & Ross, TTDS 2025/2026



15

Eval – Report (10%)

- Well written report that describes the developed system well.

Magdy & Ross, TTDS 2025/2026



16

A Basic Project (~30%)

- Use CW1 code
- Improve a little bit
- Implement some basic interface
- Select a collection of 100K document

Magdy & Ross, TTDS 2025/2026



17

An OK Project (~50%)

- Use some code from CW1, but reimplement to be highly optimized in storage + speed
- Implement a nice interface for query submission and results display
- Select an interesting collection of large amount of documents
- Host online (and potentially live indexing)
- Add few features to your engine (check the slide "More?").

Magdy & Ross, TTDS 2025/2026



18

An Excellent Project (~70+%)

- Same points as in OK project +
- Innovative search task or data collection
- Live/Robust/Scalable
- Multiple additional features

Magdy & Ross, TTDS 2025/2026



19

Process

- Identify your team members
 - Search for different skills
- Agree on your general project idea
- Draft a title for your project (OK to change later)
- Elect a contact person for the group
- Contact person → submit the list of group members (include student ID) + title of project
- Start working
- Submit once you finish

Magdy & Ross, TTDS 2025/2026



20

Proposal/Group Submission

- 1 Team member should fill out sign-up form (link will be posted on Piazza)
- Includes:
 - List of all team members (select 1 as contact person)
 - Team name (optional)
 - Project title
 - Project abstract (up to 1 page)
- You will receive a group ID via email
 - Future communication, "[TTDS-Project] Group <ID>"
- We might give feedback if proposed project looks irrelevant

Magdy & Ross, TTDS 2025/2026



21

Deadlines

- Submission of project group + title:
Friday 12 December 2025
- Project submission:
Friday 6 March 2026, noon UK time
- Submissions are accepted any time before the deadline!

Magdy & Ross, TTDS 2025/2026



22

Project Submission

- **Link** to your live search engine
- **Report**
 - 6-8 pages for project description (explain each component in your project and how it works what method/tool used to implement)
 - This is used for the group mark
 - 1-2 pages: each member of the group should write a paragraph/section on his/her contribution clearly in the report. Which role was taken, and what work was done.
 - This is used for individual marks
 - Appendices can be added at end of report, but be aware that markers are not required to read them

Magdy & Ross, TTDS 2025/2026



23

Allowed / Not Allowed

- **Not Allowed:**
 - Get a ready app/project and submit
 - Using data collections that are not public
 - Using IR toolkits (such as Solr)
- **Allowed**
 - Using libraries for adding more features
 - More ready libraries → more expected features
 - Discussing with other groups and sharing ideas

Magdy & Ross, TTDS 2025/2026



24

Advice

- Have the role of each member **very** well-defined from the beginning
- Agree on each single step before you start
- Use *Trello*
- Elect a team leader
 - Has the right to have final decision when no agreement could be reached by members
 - Organises work among members and follows progress
- If X can have outcome A
team of 5X should have an outcome of >> 5A

Magdy & Ross, TTDS 2025/2026



25

SCRUM

- Clearly defined project management method
- Key points
 - Defined roles (e.g., product owner)
 - Split your time into sprints (set internal deadlines)
 - Keep a product & sprint backlog
 - Work iteratively (get a basic version up asap, then improve)
 - Hold sprint retrospective (what went well? what can be improved?)
- More information: <https://scrumguides.org/scrum-guide.html>

Magdy & Ross, TTDS 2025/2026



26

Good luck!

- Any questions?

