Text Technologies for Data Science
INFR11145

Group Project

Instructor
Walid Magdy & Björn Ross

Group

• Members:
  Min: 4, Max: 6

• Recommendation:
  Look for diverse skills:
  Planning, coding, interface, writing report

• Couldn’t reach 4:
  - Two groups could be merged into one
  - Exception could apply in very limited situation
  e.g. only one group remaining with less than 4
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 mark on course. Take it seriously!

What is Required

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

Retrieval Models

• Which one to select?
• Only one?
• Tf\idf? Which formula? BM25?
• LM?
• New novel model optimized for you task?
• L2R?
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

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)
Marking

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

- \( \text{Mark}_{\text{project}} \): 0 - 100% (same for all members)
  - Completeness and system working properly
  - Effectiveness/Efficiency
  - Innovation/Creativity/Features
  - Report

- \( \text{Mark}_{\text{individual}} \): 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 less 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

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
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

Eval – Report (10%)

• Well written report that describes the developed system well.
A Basic Project (~30%)

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

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?”).
An Excellent Project (~70+%)

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

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
Proposal/Group Submission

- 1 Team member should fill out this form:
  - 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

https://forms.office.com/r/DYkVFk6h6

Deadlines

- Submission of project group + title: Friday 17 Dec. 2021
- Project submission: Friday 18 March 2022, 5pm UK time
- Submissions are accepted any time earlier than the mentioned above!
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)
  - 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.
  - Appendices can be added at end of report, but be aware that markers are not required to read them

Allowed / Not Allowed

- **Not Allowed:**
  - Get a ready app/project and submit
  - Using data collections that are not public
  - Use 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
Logistics: Rooms and TA support

- Rooms and TA support available

AT 5.04, 5.05, 5.07 (112 spaces)
Tuesdays 12:10-15:30
TA (Youssef Al Hariri) present in the first hour

Logistics: Web hosting

- Google cloud credits available (especially for running a virtual machine in Compute Engine)

- Step 1: Retrieve coupon
  - Check Learn for the link (only UoE students taking TTDS get free credits)
  - Use your student email
Logistics: Web hosting

• Step 2: Redeem coupon code
  • Use any Google account (e.g. Gmail)
  • You must redeem your coupon code before 20 January 2022!

Logistics: Web hosting

• Step 3: Pool credits in group
  • Create a project in Google Cloud for the coursework
  • Make the other group members owners of the project
  • When you run out of credit, change the billing account for the project
  • Make sure that your group mates have redeemed their coupons before 20 January 2022!

• Links to full details on the process can also be found on Learn
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

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](https://scrumguides.org/scrum-guide.html)
Example Project

• The 2nd best project from AY 19/20

• Presenter: Maysara Hammouda, co-founder & CTO of Predictiva

(check video recording of lecture)