Text Technologies for Data Science

INFR11145

Group Project

Instructor:
Walid Magdy
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 something of what you learnt in the course into an interesting application

• Explore new or challenging data sets

• This is 30% of the mark on course. Take it seriously!
Idea 1: Inverted index + search simulator

- Create an app that takes very small collection of text (e.g. 10-50 documents) and show in visual steps how the inverted index (postings) for each term is created.
- Accepts queries from app user then simulate in steps how terms postings are loaded, then linear merge is applied for AND/OR operations.

\[
\begin{align*}
\text{ink} & \quad 3:1 \quad 4:1 \quad 5:1 \\
\text{wink} & \quad 1:1 \quad 5:1
\end{align*}
\]
Idea 2: Create a Test collection

• Crawl an interesting collection of documents
  • News/books/videos/tweets/FB pages
  • Large enough (50,000+)
• Create a set of topics that relates to the collection
  • In TREC topic format (query+desc+narrative)
  • At least 25 topics
• Run several IR setups
  • with/without stemming
  • TFIDF (different SMART notions), BM25, LM, …
  • Try again with PRF (different $n_d$, $n_t$)
• Apply pooling
Idea 3: QE simulator

• Create an app that takes an input collection and query, apply ranked retrieval (using any IR toolkit, e.g. Solr), and displays results to user

• User can select in the top retrieved documents what is relevant and what is not (create nice interface)

• Show how Rochhio’s FB formula will change the query. Show terms learnt to be positive and others to be negative, so the modified query

• User can select PRF with sets of documents. Display the terms learnt
Idea 4: PageRank Simulator

• Input in text or graph a set of pages and links between them

• Show in visual simulation how the page rank of pages will change with every step

• Change the value of $\lambda$ and observe how this would affect page rank

• Allow editing the links then update page rank accordingly
Idea 5: Text classification

• Propose an interesting text classification task
  • Classify tweets by topic (new scheme)
  • Classify news to going to be popular or not from text
• Get the required data for the task
• Label enough set of data for training and testing
• Test different features set
  • BOW
  • Semantic features, Characters, links, …
• Apply cross-validation for testing
Idea: Think of your own

• You can think of any other project as long as it is interesting and relates directly to the course

• Feel free to propose search/classification for tasks in non-English languages

• Apps that will help normal people to learn something about IR are highly appreciated
Process

• Identify your team member
• Agree on your project idea
• Draft one page about your project (idea, data, planned system/outcome)
• Send proposal to advisor (me!)
• Get feedback about the idea
• Start working
• Submit once you finish
Deadlines

• Submission of group members: Wednesday 21 Nov. 2018

• Submission of project proposal: Wednesday 28 Nov. 2018

• Project submission: Friday 8 February 2019, 5pm

• Submissions are accepted any time earlier than the mentioned above!
Proposal/Group Submission

- Each group needs to assign one member to be representative to communicate with the advisor.
- Submission is by direct email to the advisor.
- Email subject: “[TTDS-Project] <subject>”.
- Once the proposal is submitted, you should receive:
  - Feedback on your proposal (if it is OK, enough, relevant, or too passionate).
  - Your group number
- Future communication, “[TTDS-Project] Group <xx>”
Project Submission

• Running system/collection that is easy to use
  • For app projects, having your tool as online service would be highly appreciated

• Report
  • 4-8 pages for app projects
  • 6-12 pages for research projects
  • 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.

• Demo + presentation (optional)
Allowed / Not Allowed

- Not Allowed:
  - Get a ready app/project and submit
  - Using collections that are not public

- Allowed
  - Using libraries and tools
  - Discussing with other groups and sharing ideas
Marking

• 50% on the project (same for all members):
  • How challenging was it?
  • How complete and well furbished is it?
  • Impact (usage for systems, and novelty for research).

• 50% on individual contribution (separate for each member)
  • The amount of effort contributed to the project
  • Does not have to be technical (labelling data, interface design, …)
Examples from last year

• *Implementing Text Search Across Multiple Online Videos Using Captions*
  • 20K captioned TED talks in multiple languages
  • 25 queries in multiple languages + qrels
  • Tested multiple retrieval approaches for best performance

• *Creating a Search Engine for the School of Informatics*
  • All course pages of informatics collected and indexed including content
  • Allow search slides, pages, or external pages of similar courses in other universities
  • Allow different weighing schemes of queries and docs

• **Note**: last year, project was only 20% of mark. We expect more this year!
Advices

• 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