Course 2022-2023

Guidelines for the IRTM Project

General

Students can select one (or more) environments and implement, based on either provided text corpora or on other open-source collections, one of the practical exercises provided.

Implementation should include a number of relevant information retrieval and text-mining operations.

Students can work individually or in groups of two. If you work together, you will both have to do two different implementations of each of the line items hereunder. E.g. one with NLTK and one with BERT and compare the two. Or one topic modeling with LDA and one with NMF.

Proposals of work will have to be submitted within one week (Wednesday, April 19, 5pm) after which they will be reviewed. After approval, the students can start the implementation of their proposals.

Submit on Canvas (assignments).

25% of your grade

3 Types of Project to Choose From

To give you some more guidance, I would like you to choose from the following 3 types of projects:

1. Take a collection of books, extract names, locations, times, sentiments, relations, events, anomalies, etc. Combine these and visualize them. See: https://textmining.nu/ for videos and blogs on former projects. For this project you have to follow the following steps:
    a. implement: text-extraction,
    b. pre-processing,
    c. basic entity extraction (PERSON, COMPANY, LOCATION, ORGANIZATION, DATES, TIMES, ….),
    d. link these basic entities together with each other or with SENTIMENTS or EMOTIONS.
    e. Also think about detecting special events such as WAR, VIOLENCE, SPELLS, PARTIES, INTIMIDATION, PRESSURE, etc and link them to the exacted BASIC Entities.
    f. Discuss linguistic complexities such as boundary-detection, co-reference handling and negation handing if applicable
    g. You can use Topic Modeling to find more relations.
    h. Quantitatively measure the results (P, R, F1, 11 Points PR, Kappa-Cohen for annotated data) either by using a ground-truth or by checking a small random subset (25-50 data points).
    i. Visualize the results (word clouds do not count, you have to do something more advanced).

2. Optimize a search engine such as Lucene:
    a. add additional lookup tables in the index for forms of semantic search (e.g. PERSON and “travel” and LOCATION where PERSON and LOCATION are collections of extracted entities).
    b. Use Word-Embeddings for better relevance ranking.
    c. Use Word-embeddings for better relevance feedback. E.g. allow a user to select a relevant paragraph and return documents with exactly that text in content.

3. Do a better job integrating a chatbot with a search engine than Microsoft:
    a. Convert an NLP query to a keyword search.
    b. Execute keyword search on traditional search engine.
    c. Take top results.
    d. Make a better search integration by:
        i. Create a knowledge graph from the relevant text in these results.
        ii. Use this to manage the conversation (prompt generation in GPT like models).
        iii. Add XAI (where does the info come from).
        iv. Have a more meaningful conversation

    In this project, the idea is to show examples were BING/GPT goes wrong and how you can solve that with text mining techniques you learning in the IRTM course. An example where BING goes wrong: if there are two individuals with the same name, GPT will mix up facts about them. It has no understanding that these are actually 2 different real-world entities. Same for companies, locations or products.

In all 3 cases you will get extra points for a video of your project.

Proposal for IRTM Project

The proposal should be:

- Fit on one to two pages.
- The title of your project
- Contain your name(s) and student ID.
- The project type you choose (Text-mining, Search Engine or better Chatbot integration).
- Which text or benchmark you use.
- The steps you will take for the implementation. Address the point listed above in the 3 project examples.
- Do not forget to mention how you will QUANTITATIVELY evaluate the results

Report IRTM Project

Guidelines for the project report:

- 8-15 pages.
- Write it like a scientific paper.
- Describe the process you followed: what worked / did not work / how did you address this? Compare results
- Teams of 2: General aspects can be repeated (which corpus, how did you get it). But explain in detail who did what (preferably in a table).
- Data sets used (text corpus and training data)
- Quality measurement methods & results.
- Screenshots of the visualizations.
- Recommendations for future work
- Do not forget references!
- Include source code and (links to) data sets and final results in the submission you make to canvas. DO NOT MAKE SUBMISSIONS larger than 10 Mb to Canvas. If you have more to share: give us a link to you Google drive or another external storage location.

Presentation IRTM Project

For the presentation of you project:

- Individuals: 5 minutes. Teams: 10 minutes (including setup, test your laptop in advance!).
- Slides should contain: explain data set, problem, approach, algorithms (don't forget anaphora, negations, …), what is standard in libraries you used, what did you do yourself, training data used, results (P, R, F1, Kappa, Rand, Silhouette, SMI, …), improvements, XAI, Distillation, conclusions, future research.
- Video (cool) or live demo