Clinical Trials Search Engine Development: Course Project

Assignment created by: Georgios Peikos

1 Goal

Develop a search engine to retrieve relevant clinical trials based on written summaries of patients' medical conditions. The engine should be able to process and interpret descriptive medical information to retrieve appropriate clinical trials.

2 Project Description and Material

This project allows students to experiment with and practice various Natural Language Processing (NLP) and Information Retrieval approaches. The first task involves thoroughly analyzing a collection of documents related to clinical trials. Students can leverage knowledge acquired from previous courses and external resources to explore various aspects of the collection, such as term occurrences and the percentage of collection documents related to specific medical conditions. Moreover, pertinent material can be provided if a team would like to leverage a specific approach (e.g. keyword extraction, LLMs, etc.).

After this initial analysis, teams can examine available patient information, which serves as queries in this search task. This stage may include tasks like identifying a patient's primary medical condition based on their information. This extracted information can be used as a query instead of a relatively big provided description.

With a solid understanding of document collection and patient queries, teams will create a search engine. This process involves several decisions, such as selecting which parts of the documents to index or processing the documents in a specific way before indexing them. Additionally, teams are encouraged to explore innovative ways to process queries, for instance, extracting essential information from patient data to refine search queries.

This project aims to enhance students' technical skills in NLP and information retrieval and improve their ability to apply them in practical, real-world scenarios. Students are strongly encouraged to think creatively and innovatively in addressing this search task. The project is designed to be collaborative, creating an environment where students can develop and enhance essential soft skills. These include teamwork, effective communication within a team, and understanding team dynamics.

Students are encouraged to utilize the discussions and materials presented in the labs to conceptualize and develop their projects. Additionally, they should use the following resources, which have been carefully selected to aid in developing and executing of potential project ideas. These resources include a variety of academic papers, online tutorials, that provide a comprehensive understanding of NLP, information retrieval, and clinical trials data handling.

- **Official Clinical Trials Website:** A comprehensive database of privately and publicly funded clinical studies conducted around the world.
- **TREC Clinical Trials Track:** Provides information and datasets related to the TREC's clinical trials track.
- **Tutorial in Health-Related Information Retrieval:** A detailed tutorial covering various aspects of information retrieval in the health domain.
- **Tutorial in Precision Health:** Offers insights into the evolving field of precision health and its applications.
- **Research Paper on Medical IR:** An academic paper discussing advanced topics in medical information retrieval.
- **GitHub Repository for Information Extraction in Medical IR:** A repository containing code and resources related to information extraction in medical information retrieval.
- **PyTerrier Documentation and Experiments:** Comprehensive documentation and guides for conducting experiments using PyTerrier, an IR research platform.
- **GitHub Repository for Clinical trials retrieval:** Repository for Clinical trials retrieval.
- **Additional material will be provided as teams progress with their chosen project ideas.**

3 Participant Guidelines

- **Eligibility:** Open to **all** students enrolled in the course.
- **Team Formation:** Teams are required to be formed with a composition of either 2 or 3 members. Each team must appoint one member to act as the communication lead. The responsibility of the communication lead includes registering the team and its members with the instructor. This registration must be completed by November 27. To register, send an email listing all the team members and indicating the appointed communication lead at georgios.peikos@unimib.it.

4 Steps for Completion

4.1 Timeline

1. **Team Formation and Registration:** Deadline - **November 27**, via email.
   (a) After registration, teams should start working on the Initial Proposal Submission (see Section 4.2).
   (b) Teams can start working on Development Phase I (see Section 4.3).
2. **Initial Proposal Submission:** Deadline - **December 4**. Teams have to send via email the PDF file described in Section 4.2.
3. **Ideas Evaluation:** By **December 6**, each team will receive an email with further materials and comments on their proposed ideas.
   (a) After the email communication, teams can continue working on Development Phase I and start working on Phase II (see Sections 4.3 and 4.4).
4. **Final Project Submission:** Deadline - To be specified; mid January. This will include a single PDF that contains the whole project and the python source code.
5. **Project Presentations:** To be specified; mid January.

4.2 Initial Proposal Submission

This initial proposal is the first stage of the project. Your submission should be a concise yet comprehensive document, spanning 1-2 pages in standard A4 format. The PDF should be send by the team's communicator via email and encompass the following components:

- **Task Description:** Provide a brief explanation of the task. Define the ultimate goal of the search engine you plan to develop and discuss any perceived limitations or challenges.
- **Method Outline (i.e. your idea/s):** Present an overview of your proposed methodology. You are encouraged to use visual aids such as diagrams or flowcharts to enhance clarity. For creating these, `https://app.diagrams.net/` can be a useful tool.
- **Resource References:** Include citations or references to the resources that have informed your understanding of the problem and inspired your ideas. This may include academic papers, online tutorials, or any other relevant material.
- **Technical Aspects:** List and describe the technical elements and tools you intend to utilize in your project. This could include, libraries, frameworks, or algorithms, or tools you would like to learn. Based on these, you will receive further resources.

4.3 Development and Implementation: Phase I

The initial phase of development focuses on the implementation of three retrieval runs, which will serve as baselines for further development. Teams are encouraged to explore and report different approaches for these runs, which may include:

- **Collection and Query Analysis:** Teams can present several statistics related to the document collection and the queries (See Project Description and Material section.).

- **Indexing Strategies:** Experiment with various indexing techniques. For instance, one approach could involve indexing the entire content of documents, while another could focus on indexing only titles or specific sections.
- **Query Formulations:** Develop different formulations of search queries. This might involve standard query formats, as well as more innovative or complex query structures.
- **Experiments:** The indexing must be conducted in the whole collection and evaluation should be based on P@5 (precision at top-five documents), and P@10.
- **Resource:** Each team can start their experiments based on the colab notebook here: `https://drive.google.com/drive/folders/15_zASJ3fnqdHhpEJeNxpeZrzjg2xq6yT?usp=sharing`.

These baseline runs should align with the methodologies and strategies discussed in the last Information Retrieval lab, i.e. simple approaches. This phase is important for setting a foundational understanding of the retrieval process, upon which more advanced techniques will be built in subsequent phases of the project. Please ensure that your report goes beyond merely stating the results. It is important to provide an explanation of the obtained results.

4.4 Development and Implementation: Phase II

During this second phase, teams will move forward from the foundational work established in Phase I to the development and evaluation of their unique ideas. The indexing must be conducted in the whole collection and evaluation should be based on P@5 (precision at top-five documents), and P@10.

5 Assessment Criteria

Credits. The entire project carries a total of 3 credits. The initial part, comprising the proposal of your ideas and the development in Phase I, contributes to 1.5 credits. The subsequent Phase II is allocated the remaining 1.5 credits.

In Phase I, the focus is on evaluating your understanding and application of the fundamental concepts covered in the labs. Phase II presents an opportunity for each team to delve into an area of their personal interests. It encourages exploration and innovation, as teams have the freedom to choose and develop their own ideas.

The projects will be evaluated based on their:

- **Functionality:** Efficiency in retrieving relevant clinical trials.
- **Innovation:** Originality of the proposed solutions.
- **Technical Implementation:** Quality of coding, algorithm design, and data handling.
- **Team Collaboration:** Communication and teamwork.
- **Presentation:** Written and oral presentations.

6 Additional Information and Support

For any additional information or support, please send an email at georgios.peikos@unimib.it.