University of Pavia
Artificial Intelligence BSc
Information Retrieval and Recomender Systems
Michele Ventimiglia, Manuel Dellabona

Clinical Trials Search Engine

Final Project Proposal

Task Description

The primary goal of our search engine is to facilitate the retrieval of relevant clinical trials by interpreting written summaries of patients' medical conditions. The engine will be designed to process descriptive medical information, translating it into effective search queries that will be obtained by translation in english, query expansion, primal medical condition (PCM) extraction.
Those will retrieve and rank the appropriate clinical trial data using live user feedback to rerank the results.
This methods should make more accessible the information retrieval to both who knows english and specific medical terms and patients.

Challenges

- **Natural Language Processing**: one of the key challenges will be the development of NLP algorithms capable of understanding and processing the complex language often found in medical summaries, including technical terminology and varied ways of expressing similar conditions;
- **Performance**: ensuring that the search engine can quickly and accurately retrieve relevant results from a vast database of clinical trials;
- **User Experience**: creating an intuitive interface that can guide users in providing the necessary information and presenting results in an accessible manner.

Method Outline

Our proposed methodology integrates NLP and Information Retrieval (IR) techniques to develop a search engine tailored for the domain of clinical trials. The overview of our method is visualized in the accompanying flowchart, which breaks down the process into distinct phases:

Query Formulation

- **Traduction**: convert patient information into search queries;
- **Expansion**: Pyterrier built-in expansion;
- **Primal Medical Condition Extraxtion**: use LLMs (Language Models) to identify the PMC and make a query exploiting the condition paragraph in the documents to improve retrieval performance.

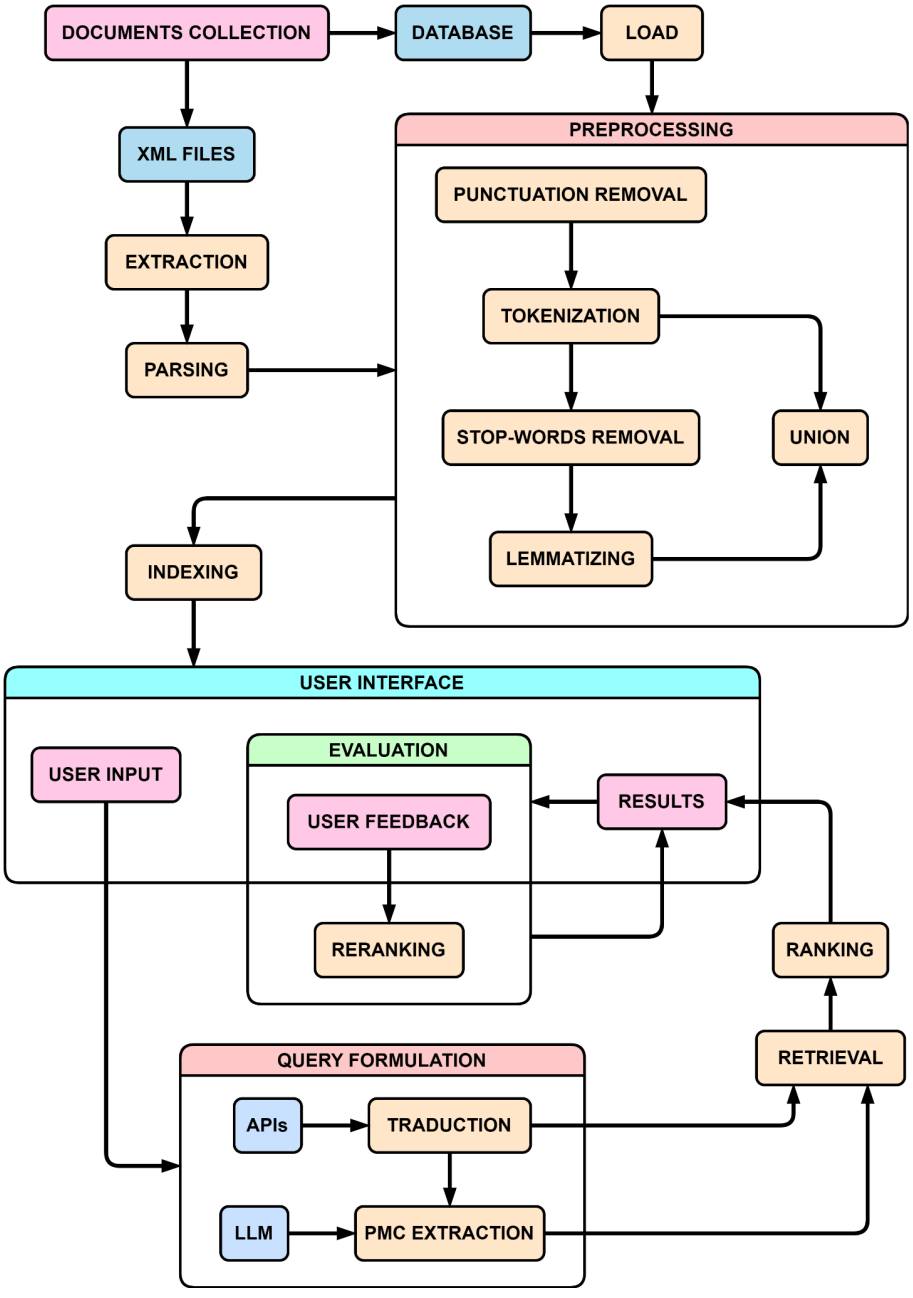
Retrieval & Ranking & Re-Ranking

Implement an IR system, to retrieve and rank clinical trials based on query relevance, then with a user feedback we rerank the results with a neural model (ClinicalBert) that calculates the similarity between the documents.

User Interface

- **Input**: user-friendly interface that allows for easy input of patient data;
- **Results**: ordered results with title and description;
- **Feedback**: feedback loop where users can provide input on the relevance of retrieved documents.

Pipeline



Resource References

- Course Literature: we have referred to the comprehensive academic curriculum provided throughout our coursework, which has been instrumental in laying the theoretical foundation for the project;
- https://sease.io/2021/12/using-bert-to-improve-search-relevance.html
- https://arxiv.org/abs/2306.02077

Technical Aspects

Programming Languages, Frameworks and Tools

- **Python**: our primary development language, renowned for its extensive libraries and community support in data science and machine learning;
- **Java**: utilized as the backend for PyTerrier, ensuring robust information retrieval capabilities;
- **C++**: powers the backend of llama.cpp framework.

Core Libraries and Tools

- **PyTerrier**: an IR research platform that facilitates our indexing and retrieval operations, chosen for its integration with Python and comprehensive IR functionalities;
- **PyTorc**h: our chosen machine learning framework, which offers flexibility and a dynamic computation graph for NLP model development;
- **spaCy**: an industrial-strength NLP library that we employ for NLP tasks;
- **Pandas**: a data manipulation library essential for handling and analyzing our datasets efficiently;
- **NiceGUI**: an advanced framework for a reactive Graphical User Interface.