Task Description

The task involves developing a search engine for clinical trials using PyTerrier. The goal is to provide an easy to use and understandable platform that allows researchers and healthcare professionals to efficiently search and access information related to clinical trials based on textual description of patients. This includes details such as trial descriptions, eligibility criteria, outcomes, and relevant publications.

Method Outline

The Dataset

The dataset is made up by different columns, each containing different information related to a specific medical trail based on how this information are structured. The definite combination of columns that we are going to select will be inferred by the result that we will obtain during the implementation and experiments of the search engine.

Preprocessing of the dataset

With knowledge of NLP, we will use different text preprocessing techniques to get most suitable representation of our data. Analysis and observations will be made to see how the results will differ between preprocessing techniques. In a nutshell, we will lowercase the text, irrelevant special characters will be removed, remove punctuations, numbers and stop-words. If appliable, other methods will be considered, depending on the results obtained.

Indexing

In our approach, we will try to use two different indexes in order to filter out unwanted documents. The main one, will be created upon the columns decided a priori, the second one, with the columns including the exclusion criteria of the medical trial. Our idea is to use the 'ids' of the documents retrieved from our "exclusion criteria" index, to remove these "bad" documents that also appear in the main retrieval result. We believe that in this way, we can remove those medical trials for which the patients would be excluded.

The query

Representation & Content Analysis

We will define how user queries will be represented in the system, considerations will be based on keywords, natural language or a combination of both.
An analysis will be made on how different contextual information such as, past medical history, current medical conditions, family descriptions and textual units such as, negations, numbers etc... will be incorporated into the search process. Furthermore, we will determine how different types of content will be combined to provide a comprehensive search result.

Preprocessing & LLM adoption

An LLM will be implemented to carry out content selection and entities recognition on the query. NER will be adapted to identify and extract specific entities from the queries, enhancing search accuracy by reducing the number of descriptive units. We will also see the possibility of getting different query representation to get a higher generalisation of the results and apply a selection technique to achieve the average results. Afterwards we will apply lowercasing and removing characters. Medical terms are sensitive, so in-order not to lose their meaning, we place consideration on some whole word as single tokens; example, f-MRI as a single token.

Query Form & Ranking

Creating a detailed process for handling single queries to ensure that users can effectively search for information. Additionally, develop a system for handling multiple queries and furthermore incorporating ranking techniques to present the most relevant results first. For ranking we will be experimenting on:

- TF-IDF and its variations
- BM25
- Cosine similarity
- Possibly other methods if applicable

Also weighting techniques will be considered in such a way, terms that we or the LLM, consider more descriptive will eventually guide the retrieval process to meaningful and related documents.

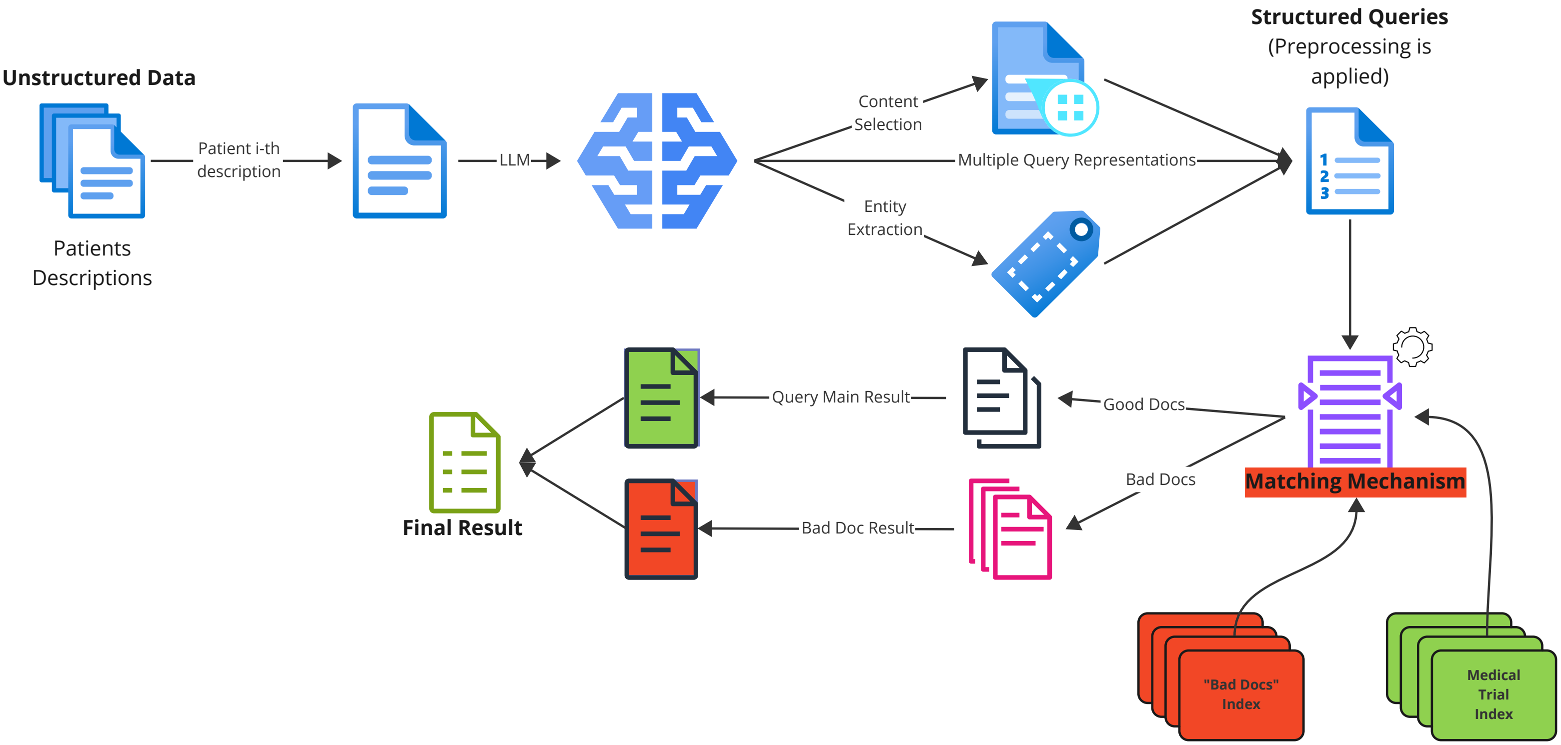
Resource References

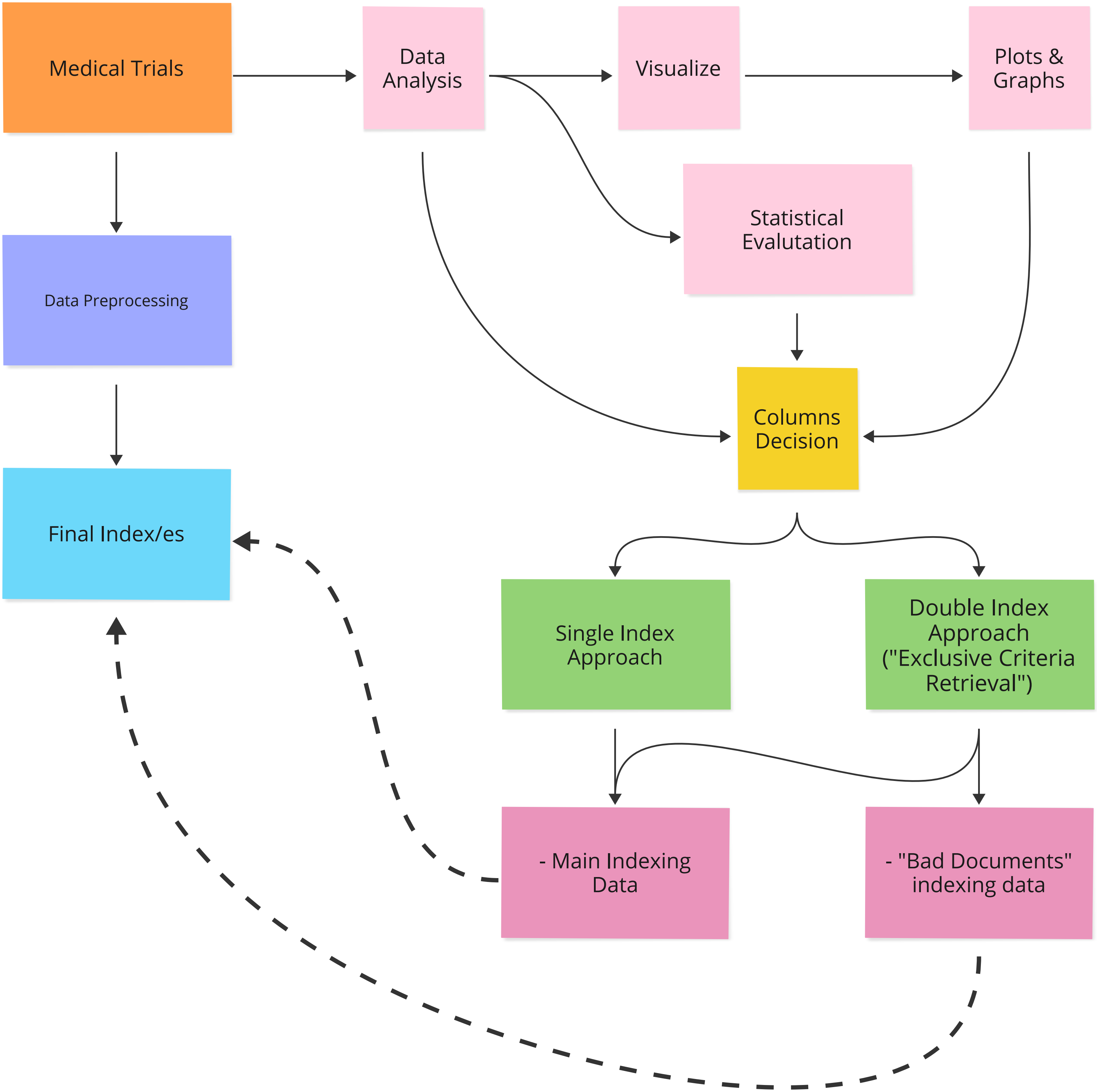
Official Clinical Trials Website: A comprehensive database of privately and publicly funded clinical studies conducted around the world.

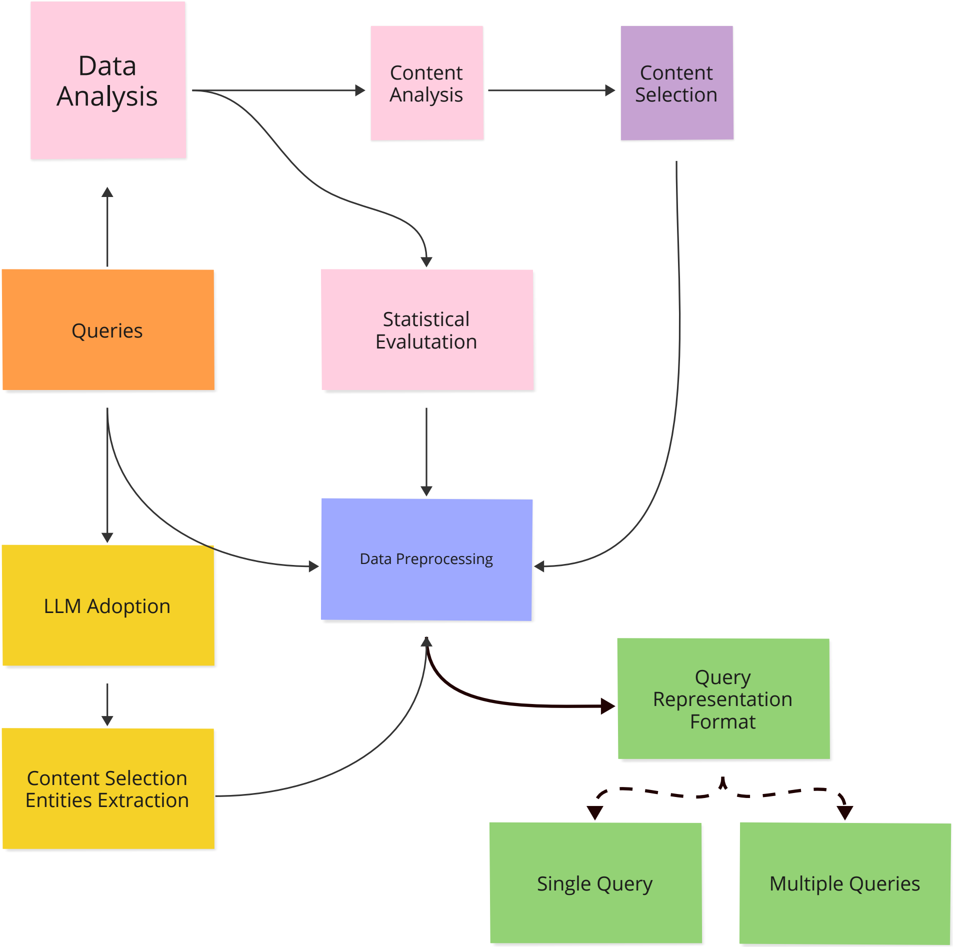
Research Paper on Medical IR: An academic paper discussing advanced topics in medical information retrieval

PyTerrier Documentation and Experiments: Comprehensive documentation and guides for conducting experiments using PyTerrier, an IR research platform.

For the tools and libraries to be used, we will decide during the implementation and execution of the project.


Unstructured Data
Patients Descriptions
Patient i-th description
LLM
Content Selection
Multiple Query Representations
Entity Extraction
Structured Queries (Preprocessing is applied)
1
2
3
Matching Mechanism
Good Docs
Bad Docs
Query Main Result
Bad Doc Result
Final Result
"Bad Docs" Index
Medical Trial Index


Medical Trials
Data Analysis
Visualize
Plots & Graphs
Statistical Evalutation
Data Preprocessing
Columns Decision
Final Index/es
Single Index Approach
Double Index Approach ("Exclusive Criteria Retrieval")
- Main Indexing Data
- "Bad Documents" indexing data


Data Analysis
Content Analysis
Content Selection
Queries
Statistical Evalutation
Data Preprocessing
LLM Adoption
Content Selection Entities Extraction
Query Representation Format
Single Query
Multiple Queries