Clinical trials search engine development

SIMONA ZLATANOVA 504574
FRANCESCA VISINI 502176
MARIE ELYSE BASSIL 503962

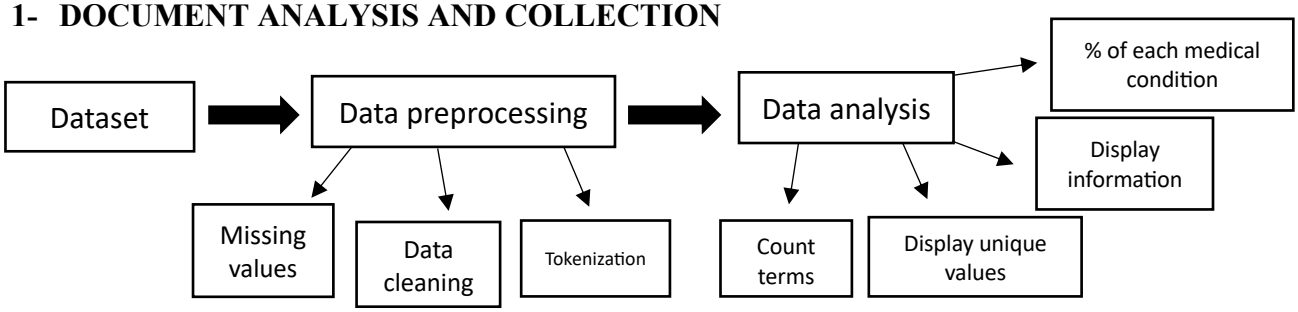
TASK DESCRIPTION: Our project aims to develop a specialised search engine capable of retrieving relevant clinical trials based on written summaries of patients' medical conditions; it should be able to proces and interpret descriptive medical information. Hence the ultimate goal is connecting the eligible patients with the appropriate clinical trials for them, improving their accessibility to medical research opportunities.

LIMITATIONS:

1. Dependency on the quality and specificity of patient summaries.
2. Limited Availability of labelled data which could impact the training of relevance models.
3. Challenges in achieving high precision due to the complexity of medical information.
4. Leveraging semantics and domain knowledge to better understand the context and nuances of medical information.
5. Considering the impact of bias and time pressure on the perception of relevance.

METHOD OUTLINE :

1- DOCUMENT ANALYSIS AND COLLECTION



In the initial phase, we embark on document collection by obtaining the dataset clinical trial documents for our search engine, leveraging the ClinicalTrials.gov dataset. Following this, the data preprocessing stage ensues, involving the removal of irrelevant information and standardization of formats.
For the analysis, we adopted a multifaceted approach:

- Employment nlp techniques to tokenize the text and quantify term occurrences, fostering a comprehensive understanding of the document content.
- Identification of unique values in pivotal fields, such as medical conditions and interventions, to facilitate the creation of facets for refining search results.
- Analysis of basic statistical information about the dataset, encompassing metrics like document count, date range, and the distribution of medical conditions.

2- INDEXING: Utilizing the Terrier Information Retrieval platform for indexing, we will create an inverted index to enable efficient document retrieval. The following components should be indexed:

- **a.** Patient Information Section: this includes descriptive medical information, summaries of patients' conditions, and eligibility criteria.
- **b.** Title and Abstract: could contain crucial information about the medical conditions being studied.
- **c.** Keywords: indexing the keywords.

3- **TERM STATISTICS:** We obtain the occurrences and frequency from the obtained index, this is crucial for term weighting and relevance.

4- **EXPERIMENTATION :**
    a. Searching an index: we utilized PyTerrier to search the index and retrieve the documents based on a manual query we put using the term: “heart-attack”.
    b. Experimenting with query language: we investigated the impact of different query formulations, term weights, and Boolean operators on search results by experimenting with different queries.

5- **QUERY PROCESSING:**
    a. **Loading queries and qrels:** We then load queries and qrels, using an IR-dataset, which is crucial for evaluating the performance of an IR system by assessing its alignment with standardized queries and relevance judgments, facilitating benchmarking and refinement.
    b. **KeyBert**: we utilize the KeyBERT library to perform keyword extraction on the queries to enable the generation of informative keywords that are valuable for our tast. We perform this using two methods:
        i. Regular keybert extraction.
        ii. KeyBert with Flair embeddings.
    c. **Analysis of the methods:** Doing both methods allows for a comparative analysis of the effectiveness of these approaches in generating meaningful and relevant keywords by:
        i. Calculation of average keywords in each
        ii. Calculation of overlap of keywords between the two
        iii. Manual evaluation

6- **RANKING ALGORITHM:** We employ Terrier's built-in ranking models such as BM25 and BM5+RM3, we will experiment with the 2 different algorithms to optimize result relevance. This is to determine how the documents will be ranked based on their elevance to the user query.

7- **DOCUMENT RETRIEVAL :** The document retrieval is performed using BatchRetrieve to search the index and retrieve a batch of relative documents to the users’ queries using the models.

8- **EVALUATION**: We evaluate the performance of our retrieval system using:
    a. Reciprocal rank RR(rel=2)@1000: Measures the rank of the first relevant document.
    b. Precision at K(P) P(rel=2)@1,5,10,25,30 and 75 : Measures the proportion of relevant documents among the top K retrieved documents.
    c. R-precision Rprec: Compares the number of relevant documents retrieved to the total number of relevant documents in the collection.
    d. Recall at K(R): R(rel=2)@10,25: Measures the proportion of relevant documents retrieved out of the total number of relevant documents in the collection.

RESOURCE REFERENCES:

- "Terrier Information Retrieval Platform" documentation.
- Official respository for the Medical Workshop in The 14th European Summer School onInformation Retrieval.
- GitHub Repository for Information Extraction in Medical IR.
- The SIGIR 2018 Health Search Tutorial
- KEYBERT Information extraction
- ExaCT : automatic extraction from clinical trial characteristics

TECHNICAL ASPECTS:

- Terrier Information Retrieval Platform: for indexing, query processing, and ranking algorithms.
- Python Programming Language: for scripting and interfacing with Terrier, as well as some data preprocessing. NLP libraries: such as spaCy, NLTK, and scikit-learn.
- Evaluation Metrics: Incorporating evaluation metrics from the Terrier platform for continuous performance assessment.
- Data Visualization: Utilizing Matplotlib and/or Seaborn for creating visualizations, such as precision-recall curves or other relevant plots, to facilitate a comprehensive understanding of our system's performance.