HI 744: Text Retrieval and Its Applications in Biomedicine Final Project

1. Methods

1.1 Data Pre-processing

The input data consisted of patient case files stored in JSON format. Each file contained structured metadata (e.g., patient identifiers) and unstructured clinical text describing the patient's medical history and diagnosis. To prepare the data for information retrieval, the following preprocessing steps were applied:

- Text extraction: From each JSON file, the patient narrative text and title fields were extracted and concatenated into a single document per patient.
- Lowercasing: All text were converted to lowercase to ensure that words such as "Cancer" and "cancer" were treated as the same term.
- Tokenization: The text was split into individual word tokens to remove punctuation while preserving alphanumeric tokens.
- Stop-word removal: Common English words (e.g., "the", "and", "of") that do not carry meaningful information for retrieval were removed. When standard stop-word lists were unavailable, a small fallback list was used to ensure robustness.
- Stemming: Words were reduced to their root form using Porter stemming (e.g., "diagnosed", "diagnosis" → "diagnos"), which helps group related word forms together.

These steps reduce noise in the data and improve the quality of similarity comparisons between documents.

Packages used:

- json, os, glob – file and data handling
- nltk – tokenization, stop-word removal, stemming
- numpy – numerical operations
- rank_bm25 – BM25 retrieval model
- gensim – Word2Vec embeddings
- requests – communication with a locally hosted LLM (Ollama)

1.2 Task 1: Information Retrieval Methods

Three information retrieval methods were implemented to identify similar patient cases.

1.2.1 BM25-based Retrieval (Traditional IR)

BM25 is a traditional, keyword-based retrieval method widely used in search engines. It scores documents based on:

- How often query terms appear in a document (term frequency)

- How rare those terms are across the corpus (inverse document frequency)
- Document length normalization (to avoid favoring exceedingly long texts)

In this project, each patient document was treated as both a query and a candidate. For every patient, BM25 retrieved the top 5 most similar patient documents based on lexical overlap.

1.2.2 Word2Vec-based Retrieval (Distributional Semantics)

Word2Vec is a vector-based representation method that learns word embeddings from the corpus. Words appearing in similar contexts are mapped to nearby points in a continuous vector space. The way it does this is by:

- Training Word2Vec module on the tokenized patient documents
- Representing each document by averaging its word vectors
- Compute cosine similarity between document vectors

This method captures semantic similarity beyond exact word matching (e.g., “tumor” and “neoplasm”).

1.2.3 LLM-based Retrieval via Reranking

A local large language model (LLM) running via Ollama was used as a reranking model. Rather than searching the entire corpus, the LLM received a shortlist of candidate patients (from BM25) and reordered them based on deeper semantic understanding. This approach reflects modern retrieval pipelines, where:

- A fast traditional method retrieves candidates
- A slower but more powerful model refines the ranking

1.3 Task 2: Evaluation Metrics

Two evaluation metrics were used: precision@5 - (p@5) and recall@5 - (r@5).

p@5: measures how many of the top 5 retrieved patients are actually relevant.

$$\text{p@5} = \frac{\text{Number of relevant patients in top 5}}{\mathbf{5}}$$

This means if the system shows 5 similar patients and 1 of them is truly similar, precision@5 = 1/5 = 0.20.

r@5: measures how many of all relevant patients were successfully retrieved.

$$\text{r@5} = \frac{\text{Number of relevant patients in top 5}}{\text{Total number of relevant patients}}$$

This means that if a patient has 4 known similar cases and the system retrieves 1 of them, recall@5 = 1/4 = 0.25.

2. Results

2.1 Quantitative Performance

A higher precision means the system is better at showing “good matches” at the top while a higher recall means the system is better at finding all relevant patients, even if they appear lower in the ranking. Low values often reflect limited overlap between known “gold” cases and the evaluated sample.

For patient 4006568-1 for instance, BM25 and Word2Vec retrieval results partially overlap, indicating that some patient cases are consistently identified as similar across both lexical and semantic representations. However, precision@5 and recall@5 values are zero because none of the retrieved patients overlap with the gold-labeled similar patients for the evaluated queries. This occurs primarily because the evaluation was performed on a very small random sample of patients, while the gold similar patients often lie outside this subset.

3. Challenges Encountered and How They Were Addressed

I encountered several challenges during this project. First, the dataset was very large, making it computationally expensive to run the retrieval methods, especially Ollama, on all patient files. Even to run BM25 on all the files took greater than 20 minutes. To address this, I only ran a sample of 100 patients so that all methods could be tested fairly without exceeding hardware limits. As I do not have a GPU, I ran Ollama (mistral:7b) locally on my CPU which caused several timeout errors. I had to limit the number of candidates the LLM ranked on (i.e., the candidate patient UIDs to be re-ranked came from BM25). I also had to simplify the prompt and could not even use the few-shot technique taught in the lecture.

Third, evaluation results often contained many zeros. This was not due to errors in the code (I believe), but because many patients had known similar cases that did not appear in the sampled subset. This highlighted an important limitation of evaluation on small samples.

Finally, different retrieval methods capture different types of similarity. Traditional methods rely on shared terminology, while neural methods attempt to capture meaning. Comparing these methods fairly required some nuance.

Overall, these challenges reflect real-world constraints in biomedical text retrieval and emphasize the importance of thoughtful system design rather than relying on a single method.