Search Engine for Medical Information Retrieval

Biasco Anna Marika [865873]
Novacco Alessia [918550]

Aim of the project: develop a search engine for medical information retrieval capable of processing natural language queries

Phases:

Pre-processing and dataset analysis

Retrieval pipelines

Improving the retrieval pipelines

1. Pre-processing and dataset analysis

Pre-processing

1. Text Standardization.
2. Data Cleaning:
    - Removed HTML tags using BeautifulSoup.
    - Stripped URLs using regex.
    - Removed numeric characters to focus on textual content.
    - Removed stopwords, punctuation, numbers, spaces, and repeated characters, keeping only the roots of meaningful words.
3. Tokenization.
4. Stopword Removal and Stemming.

Dataset analysis

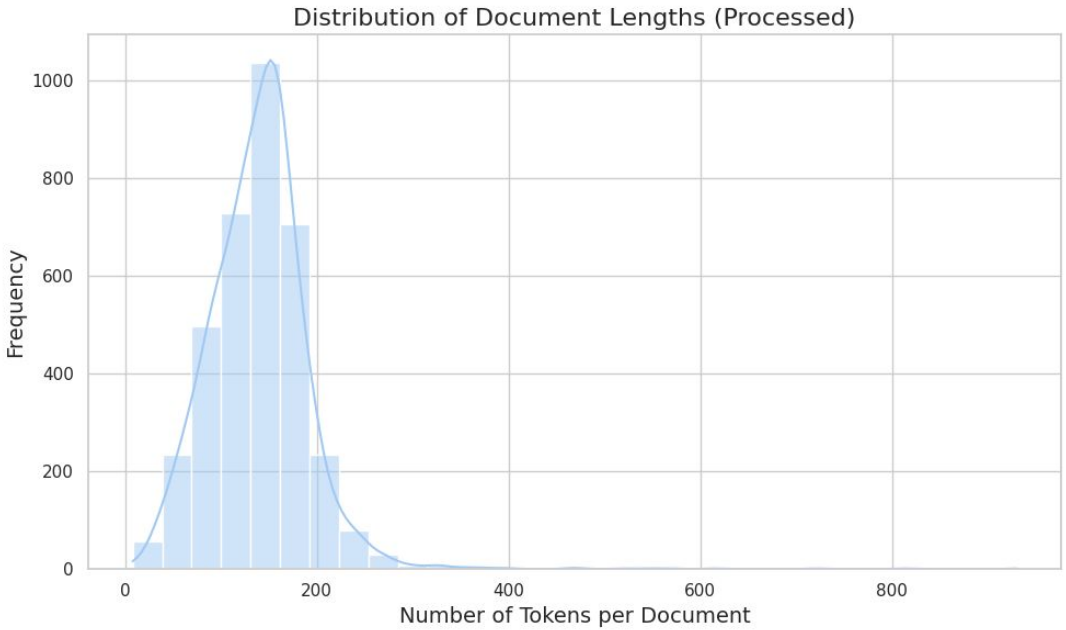
Dataset:

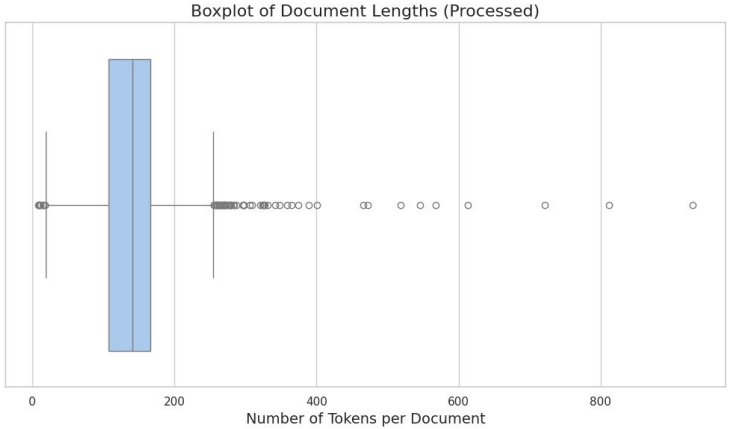
- **Queries**: 3,244 health-related questions written in non-technical English.
- **Documents**: 9,964 medical documents with complex, terminology-heavy language.
- **Relevance Judgments**: 169,756 automatically extracted qrels to support the evaluation of retrieval systems.

Analysis' steps:

1. **Document Length Analysis**: Calculated the distribution of document lengths in tokens and computed document lengths before and after preprocessing.
2. **Query Length Analysis**: Measured the average query length and analyzed its distribution.
3. **Relevance Judgments Analysis**: Assessed the number of relevant documents per query using the qrels data and valuated the distribution of relevance labels to identify underrepresented queries and terms.

Results - documents



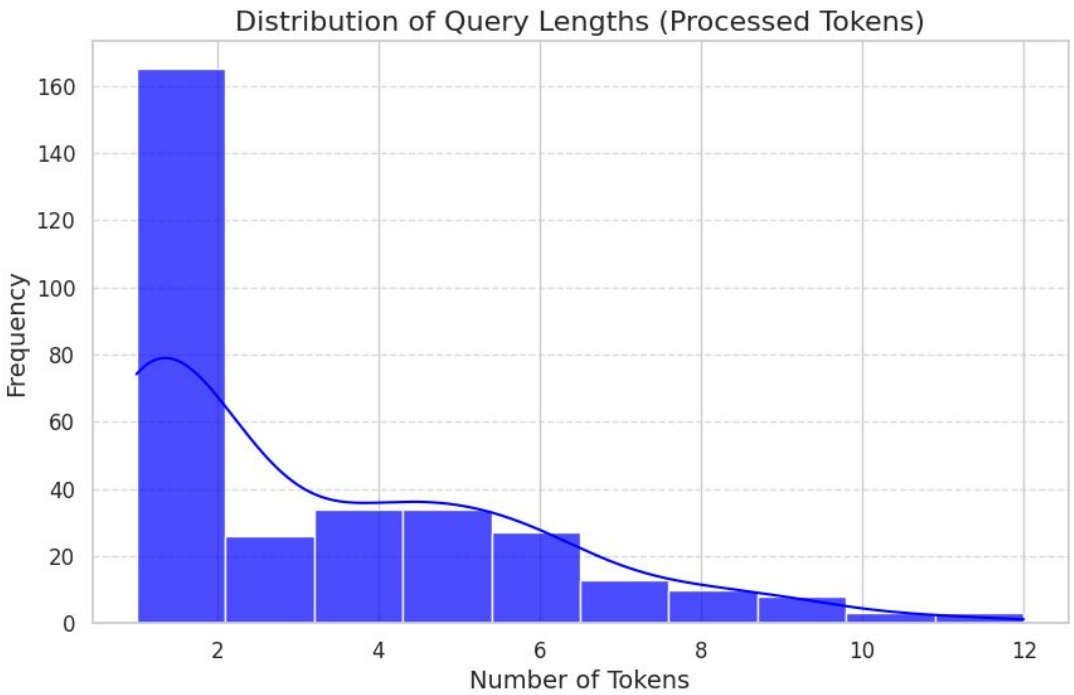


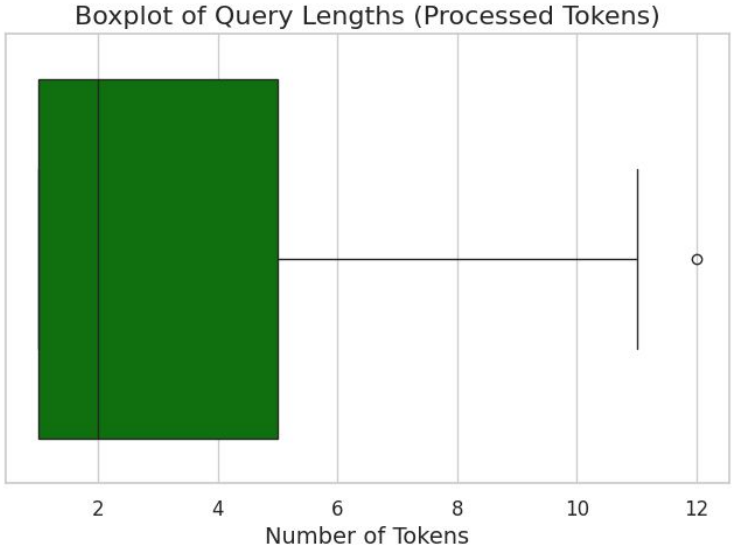
Number of outliers: 67

Longer documents may contain more relevant information but can also introduce noise if not properly indexed.

Results – documents

Results - queries

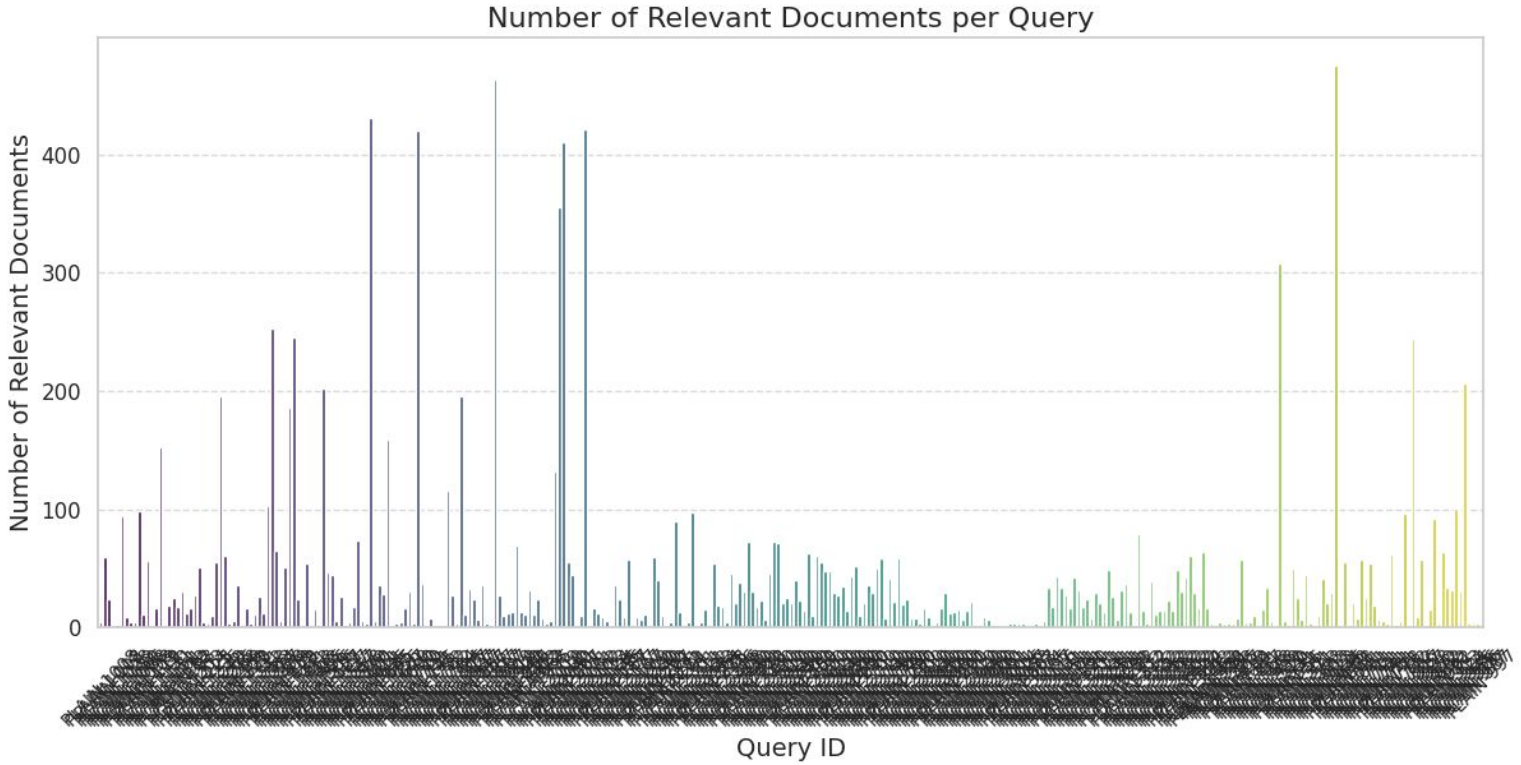




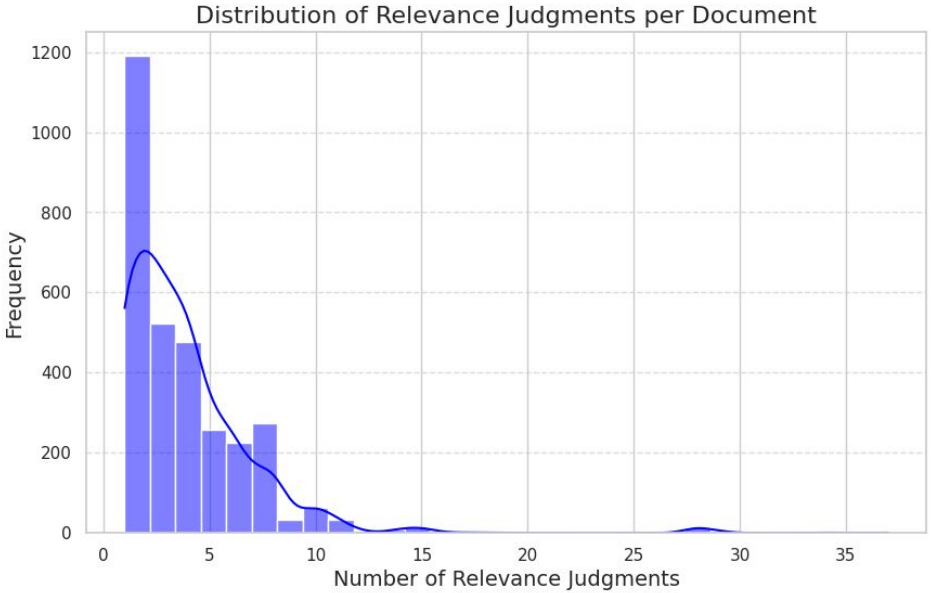
Queries are often concise, posing potential challenges for effective retrieval.

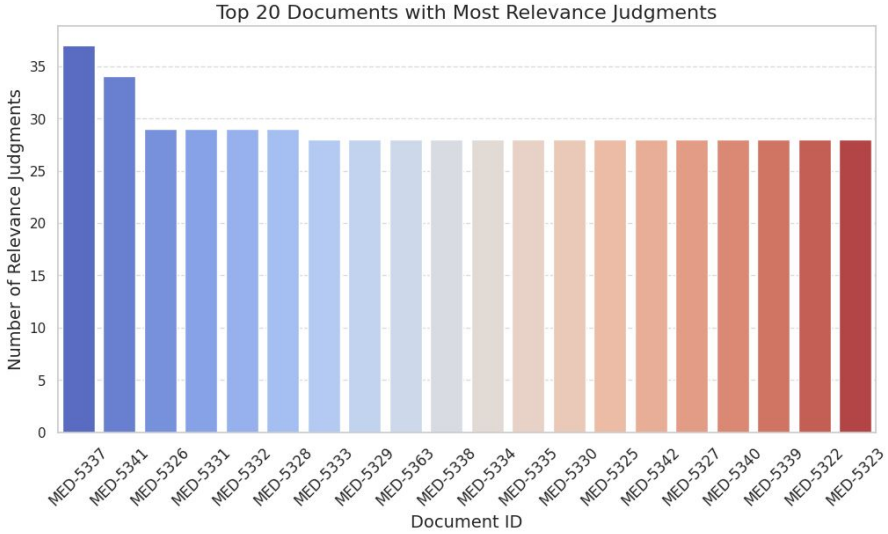
Results - queries

Results - qrels



Results - qrels





2. Retrieval pipelines

For each pipeline:

1. **Pre-processing** of the text
2. **Indexing**
3. Implementation of the **retrieval model**
4. Computation of the **evaluation metrics**:
   - MAP (Mean Average Precision)
   - NDCG (Normalized Discounted Cumulative Gain)
   - P@10
   - Recall@10

Overview

Indexing:

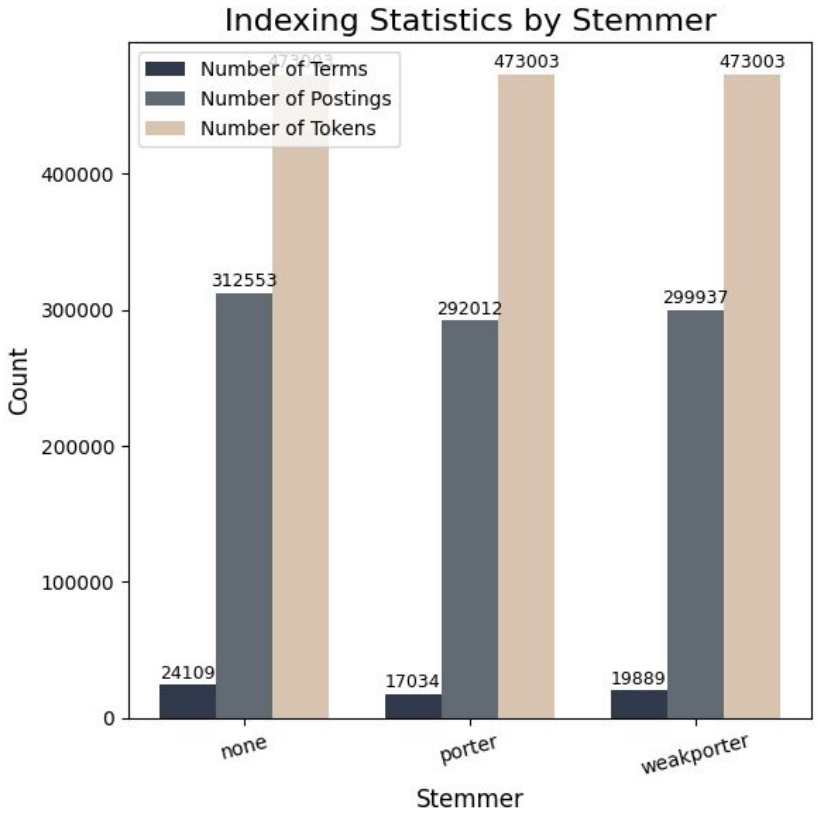
1. Full-text Corpus: Indexed titles and abstracts.
2. Title-only Corpus: Indexed document titles only.

Stemming strategies:

1. None.
2. Porter.
3. Weak Porter.

Retrieval Models:

1. TF-IDF
2. BM25



Analysis

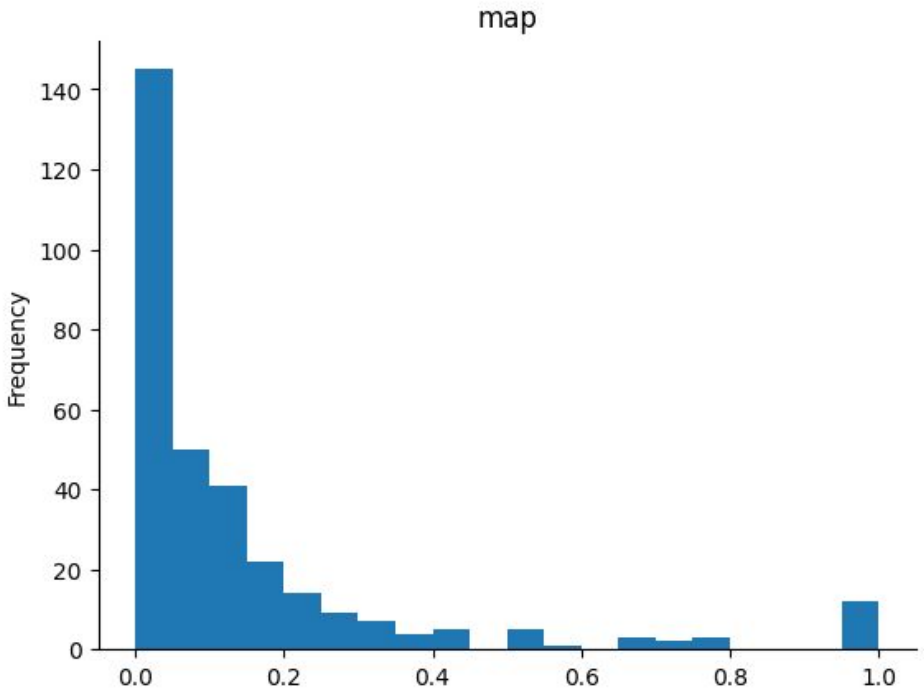
Full-text corpus:

Model Name	MAP	NDCG	P@10	Recall@10
TF-IDF	0.143	0.289	0.226	0.145
BM25	0.143	0.289	0.225	0.144

Title-only corpus:

Model name	MAP	NDCG	P@10	Recall@10
TF-IDF	0.096	0.198	0.164	0.106
BM25	0.096	0.197	0.164	0.106

Poor-performing Queries



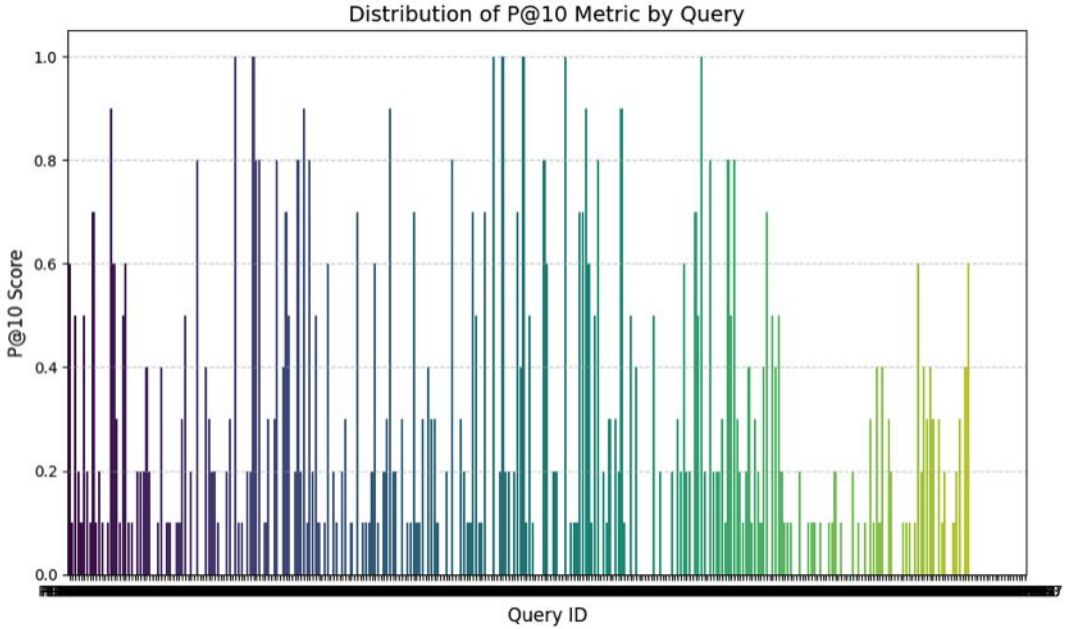
results for the pipeline ‘Full Corpus’ using BM25 as a model

queries with low MAP (<0.001)

Key Factors

1. **Ambiguity:** Queries that are too generic or vague, like "accidents," "arkansas," or "amnesia”.
2. **Rare topics:** Queries with niche or infrequent terms in the corpus (e.g., "tongue worm").
3. **Complex phrasing:** Lengthy or non-standard syntax .

High-performing Queries



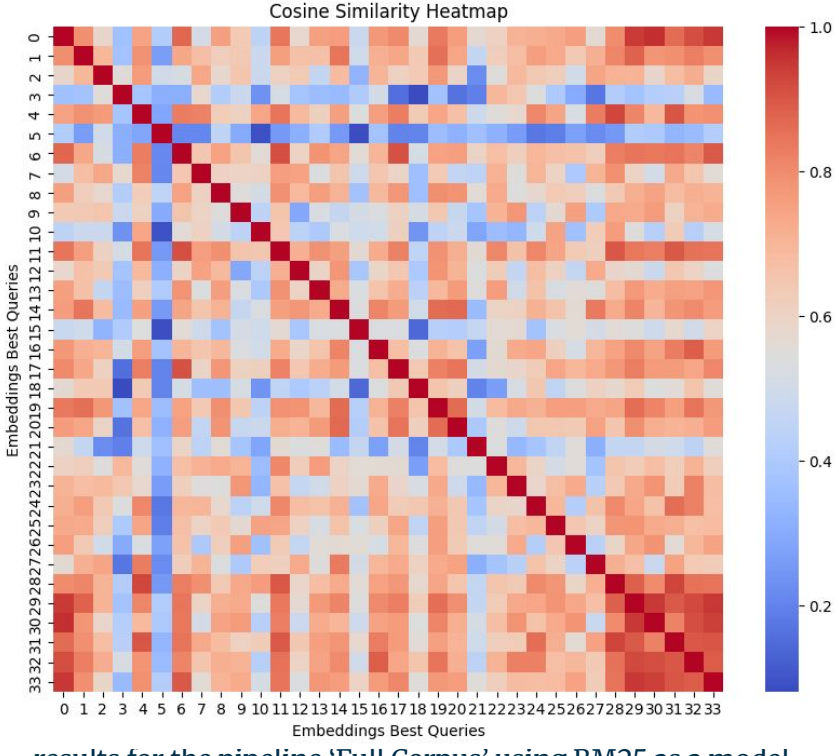
queries with high Precision@10 (>0.6)

Common Characteristics

1. **Specificity:** Use of precise and domain-relevant terms.
2. **Context relevance:** Strong alignment with the corpus vocabulary.
3. **Clear structure:** Simple and direct phrasing.

results for the pipeline ‘Full Corpus’ using BM25 as a model

High-performing Queries



results for the pipeline ‘Full Corpus’ using BM25 as a model

Word2Vec embeddings were used to compute cosine similarities to cluster semantically related queries.
These clusters revealed overlapping topics, such as dietary factors (e.g., "vegetarian diets" and "effect of carrageenan") or disease prevention strategies (e.g., "probiotics for immunity" and "caloric restriction effects").

```
Most Similar Query Pairs (qid):
Query PLAIN-153 and Query PLAIN-806
Query PLAIN-153 and Query PLAIN-1151
Query PLAIN-153 and Query PLAIN-1710
Query PLAIN-153 and Query PLAIN-1805
Query PLAIN-153 and Query PLAIN-2530
Query PLAIN-153 and Query PLAIN-2560
Query PLAIN-153 and Query PLAIN-2620
Query PLAIN-153 and Query PLAIN-2640
Query PLAIN-153 and Query PLAIN-2750
Query PLAIN-488 and Query PLAIN-1527
Query PLAIN-488 and Query PLAIN-1805
Query PLAIN-488 and Query PLAIN-2510
Query PLAIN-488 and Query PLAIN-2530
```

3. Improving retrieval pipelines

Query Expansion with Qwen

Dataset-Specific Optimization

The Qwen model was used to expand queries tailored to the NFCorpus dataset, addressing challenges with ambiguous medical queries.

Prompt Design

A specific prompt was crafted to generate three query expansions, focusing on adding descriptive language, synonyms, and medical terminology while maintaining the original intent.

Evaluation of Expansions

Three expanded versions were generated for each query, with a focus on clarifying intent and ensuring relevance to medical research, and each expansion was evaluated for effectiveness.

BM25 with Relevance Model 3

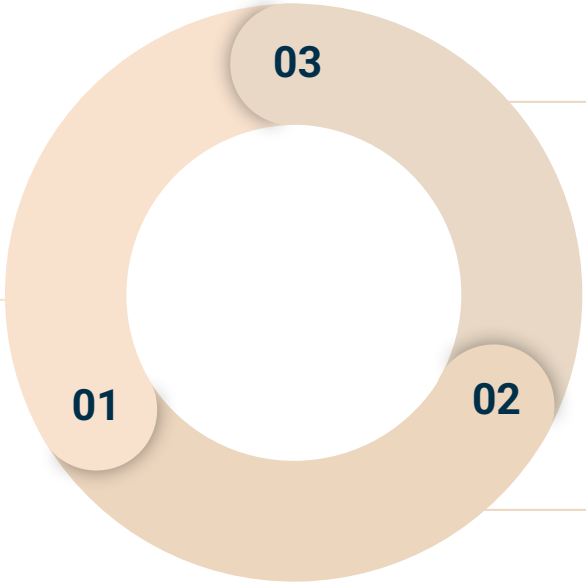
RM3 is a query expansion method based on **pseudo-relevance** feedback. The core idea of the Relevance Model is to refine the query by analyzing the distribution of terms within the original query and the initially retrieved documents assumed to be relevant. By incorporating new, representative terms into the query, RM3 enhances its semantic coverage, increasing the likelihood of retrieving relevant documents.

The pipeline follows a three-step process:

Initial Retrieval	Retrieves a preliminary set of documents using BM25
RM3 Query Expansion	Expands the original query by analyzing the retrieved documents to identify and add relevant terms, improving the ability to retrieve related documents
Second Retrieval	Performs a second round of retrieval with BM25, using the expanded query

BM25 with Neural Re-ranking using BERT

Initial Retrieval
An initial set of candidate documents is retrieved using the BM25 ranking model.



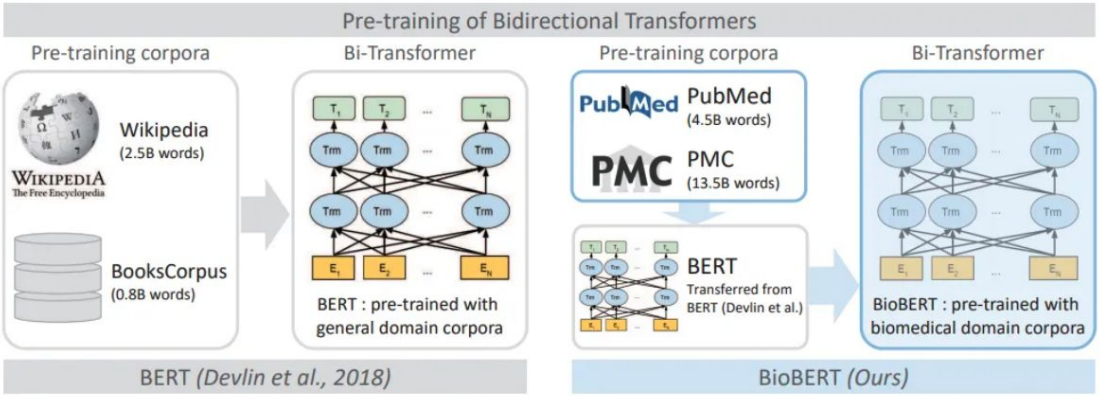
Enhanced Ranking Quality

Leveraging deep contextual embeddings, BERT improves the ranking by capturing nuanced meanings and context from the query and documents.

Neural Re-Ranking with BERT

BERT (Bidirectional Encoder Representations from Transformers) refines the rankings by evaluating the semantic relationship between the query and document.

BM25 with Neural Re-ranking using BioBERT



Ref: BioBERT paper

BioBERT is a BERT variant pre-trained on biomedical datasets (e.g., PubMed), designed to understand medical context and terminology.

We use **BioBERT ReRanker in PyTerrier**: Custom re-ranker leveraging pre-trained BioBERT model (`dmis-lab/biobert-base-cased-v1.1` from Hugging Face) to refine BM25 results. It processes queries and titles to compute semantic relevance.

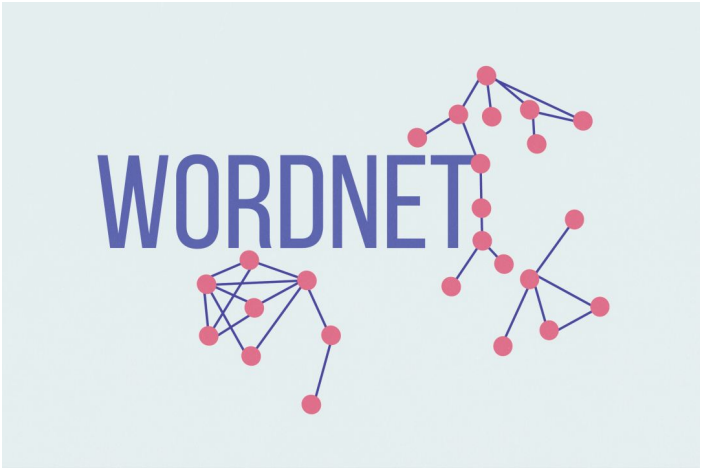
BioBERT re-ranks BM25 outputs, improving search accuracy by capturing biomedical context.

Query Expansion with Synonyms

The original queries are expanded using WordNet, a lexical thesaurus, to enrich words with synonyms via the **nltk** library.

Process:

- Each query is split into individual words.
- Synonyms for each word are retrieved from WordNet.
- If synonyms are available, the word is replaced with one; otherwise, the original word is retained.



Results with Extended Queries

Best results with query expansion using the Qwen model:

Model Name	MAP	NDCG	P@10	Recall@10
TF-IDF	0.1o7	0.282	0.168	0.113
BM25	0.109	0.284	0.172	0.116
BM25 + RM3	0.121	0.318	0.186	0.126
BM25 + BERT	0.016	0.150	0.021	0.009

Query Expansion with WordNet: Performance worsened, highlighting the difficulty of reformulating queries with specialized medical terminology.

Results with WordNet:

Model Name	MAP	NDCG	P@10	Recall@10
TF-IDF	0.069	0.142	0.104	0.077
BM25	0.068	0.142	0.103	0.076
BM25 + RM3	0.086	0.207	0.117	0.086
BM25 + BERT	0.057	0.083	0.016	0.002

Results with Original Queries

TF-IDF vs BM25: Nearly identical results, both struggling to capture semantic relationships in medical texts.

BM25 + RM3: Significant improvement in NDCG and recall due to term reweighting via pseudo-relevance feedback. Best results.

Model Name	MAP	NDCG	P@10	Recall@10
TF-IDF	0.143	0.289	0.226	0.145
BM25	0.143	0.289	0.225	0.144
BM25 + RM3	**0.164**	**0.363**	**0.241**	**0.162**
BM25 + BERT	0.063	0.181	0.084	0.062
BM25 + BioBERT	0.058	0.173	0.081	0.061

BM25 + BERT: Poor performance due to insufficient context in short abstracts and titles.

BioBERT: No improvement, likely due to limited generalization to the specific sub-domain of the corpus

4. ASPIRE

ASPIRE Evaluation (1)

Overall Measures Combined

	BM25 + BioBERT-res	BM25 + RM3-res	BM25 + BERT-res	BM25-res	TF-IDF-res
Total Queries	305	305	305	305	305
Relevant Documents	12,159	12,159	12,159	12,159	12,159
Relevant Retrieved Documents	3,317	5,745	3,317	3,317	3,315

Precision Measures Combined

	BM25 + BioBERT-res	BM25 + RM3-res	BM25 + BERT-res	BM25-res	TF-IDF-res
P@5	0.0916	0.3003	0.1028	0.2892	0.2873
P@10	0.0808	0.2406	0.0836	0.2251	0.226
P@25	0.0612	0.159	0.0648	0.1458	0.1453
P@50	0.0482	0.1107	0.05	0.0982	0.0976
P@100	0.0353	0.0743	0.038	0.0629	0.0626
Rprec	0.0745	0.1868	0.0791	0.1686	0.1691

ASPIRE Evaluation (2)

Statistical significance is tested against the selected baseline (TF-IDF-res.txt) using a paired two-sided t-test at a significance level (0.05). Multiple testing correction is performed using the (Bonferroni) method.

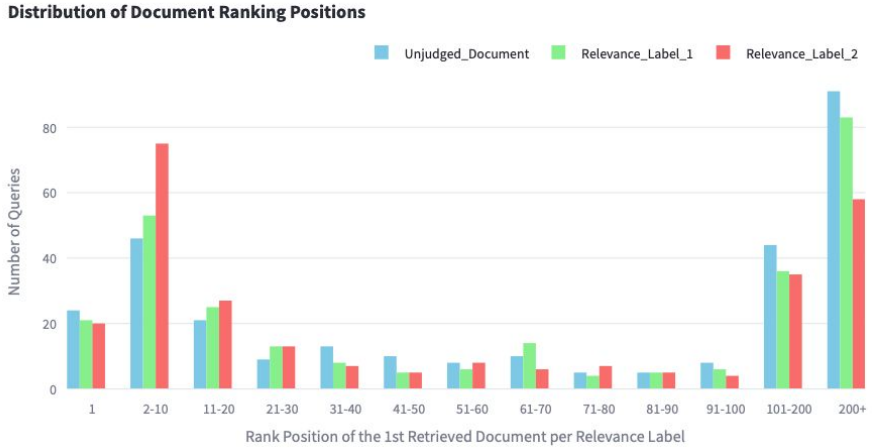
	AP@100	P@10	nDCG@10	R@50	RR@1000
TF-IDF-res (Baseline)	0.138	0.226	0.312	0.204	0.524
BM25 + BioBERT-res	0.054 | $^{0.001}_{0.001}$	0.081 | $^{0.001}_{0.001}$	0.106 | $^{0.001}_{0.001}$	0.098 | $^{0.001}_{0.001}$	0.193 | $^{0.001}_{0.001}$
BM25 + RM3-res	0.154 | $^{0.379}_{1.000}$	0.241 | $^{0.487}_{1.000}$	0.326 | $^{0.553}_{1.000}$	0.250 | $^{0.023}_{0.092}$	0.514 | $^{0.770}_{1.000}$
BM25 + BERT-res	0.058 | $^{0.001}_{0.001}$	0.084 | $^{0.001}_{0.001}$	0.113 | $^{0.001}_{0.001}$	0.105 | $^{0.001}_{0.001}$	0.204 | $^{0.001}_{0.001}$
BM25-res	0.138 | $^{0.874}_{1.000}$	0.225 | $^{0.439}_{1.000}$	0.311 | $^{0.451}_{1.000}$	0.206 | $^{0.131}_{0.525}$	0.521 | $^{0.429}_{1.000}$

Format is Measure | $^{\text{p-value}}_{\text{corrected p-value}}$. If the observed difference from the baseline is statistically significant, the background of the measure is green. The highest value per measure is underscored.

ASPIRE Evaluation (2)

Experiment: BM25 + RM3-res

Experiment: BM25 + BioBERT-res



Looking at the Distribution of Document Ranking Positions, RM3 retrieves more relevant documents overall compared to BERT and BioBERT. However, with BioBERT, more relevant documents tend to appear in the top-ranking positions.

Conclusion

1. **Neural Models (BERT, BioBERT)**

   They underperform due probably to insufficient fine-tuning and challenges in adapting to the corpus structure.

2. **BM25 + RM3**

   BM25 + RM3 is the most effective approach, leveraging term-weighting to handle specialized medical vocabulary.

3. **Original Queries**

   Original queries perform better, aligning well with document content, while expanded queries introduce noise and reduce effectiveness.