A.Y. 2024/2025
Master's Degree
Informatica

Medical Search Engine

Information Retrieval Project – January 2025

Group Partecipants:

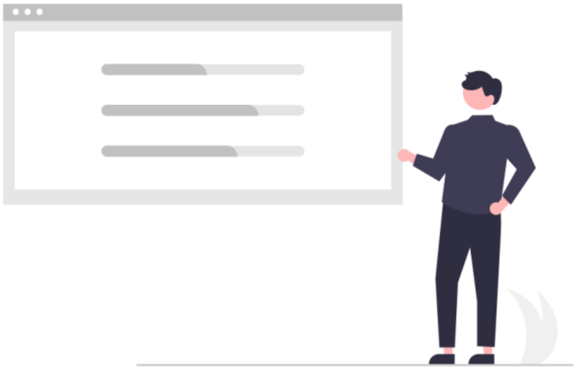
Gargiulo Elio - 869184

Piacente Cristian - 866020

Introduction

The project focus is to develop a Search Engine for Medical Information Retrieval.
Given the dataset and the initial code found on the course page, the project structure follows these parts:

- Analysis of the Dataset
- Retrieval Pipelines and Experiments
- Improvements to the baseline using Query Expansion
- Final Considerations and Future Works



Analysis of the Dataset

Analysis of the Dataset - Preprocessing

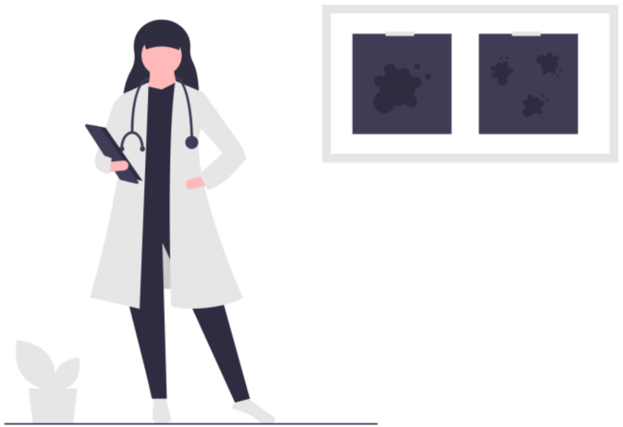
UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Before starting to analize the dataset we need to **preprocess** the data:

- Tokenization
- Normalization and Removal of Stopwords
- Stemming
- Lemmatization

Then the analysis focuses on:

- Documents Text (Abstract) and Title
- Query Text
- Relevance Judgements



Analysis of the Dataset - Summary

For both **Documents** and **Queries** we have analyzed:

- Token Count and Structure
- Vocabulary Size
- Distribution of number of terms using Histograms
- Distribution of most frequent terms using Word Clouds
- Upper and Lower tail analysis (with Outliers)

For the **Relevance Judgments**:

- Scores Distribution
- Distribution of number of relevant documents per Query
- Min/Max number of documents retrieved

Analysis of the Dataset - Tokens

For Document Text as Example: we can see the effect of the preprocessing

	text	title	doc_id	original_text_tokens	stemmed_text_tokens	lemmatized_text_tokens	original_term_count	stemmed_term_count	lemmatized_term_count
1954	INTRODUCTION: Although penile blood flow (PBF) has been recommended as an additional diagnostic test in identifying erectile dysfunction (ED) patients at risk for latent cardiovascular disease, no study has ever assessed the possible association of PBF and the relational component of sexual func...	Male sexuality and cardiovascular risk. A cohort study in patients with erectile dysfunction.	MED-3421	[introduction, although, penile, blood, flow, pbf, recommended, additional, diagnostic, test, identifying, erectile, dysfunction, ed, patients, risk, latent, cardiovascular, disease, study, ever, assessed, possible, association, pbf, relational, component, sexual, function, incident, major, card...	[introduct, although, penil, blood, flow, pbf, recommend, addit, diagnost, test, identifi, erectil, dysfunct, ed, patient, risk, latent, cardiovascular, diseas, studi, ever, assess, possibl, associ, pbf, relat, compon, sexual, function, incid, major, cardiovascular, event, mace, aim, aim, studi,...	[introduction, although, penile, blood, flow, pbf, recommended, additional, diagnostic, test, identifying, erectile, dysfunction, ed, patient, risk, latent, cardiovascular, disease, study, ever, assessed, possible, association, pbf, relational, component, sexual, function, incident, major, cardi...	123	114	122

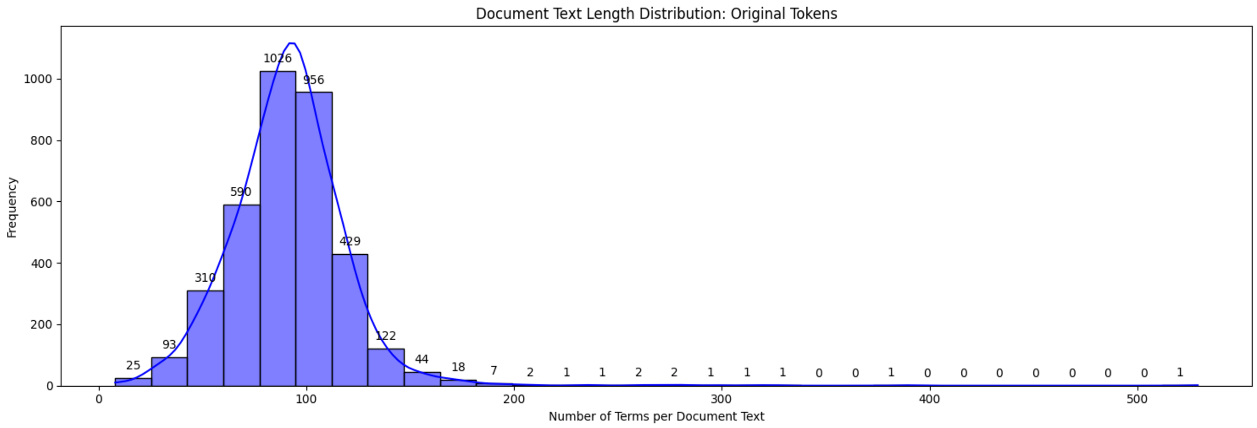
Token Information and Vocabulary Sizes for Original, Stemmed and Lemmatized Tokens

Type (Abstract)	Vocabulary Size	vs Original	vs Stemmed	vs Lemmatization
Original	24283	0	7018	2266
Stemmed	17265	-7018	0	-4752
Lemmatized	22017	-2266	4752	0

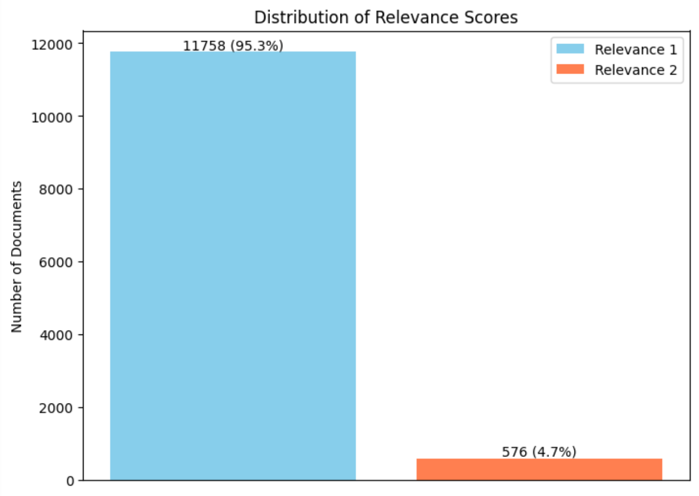
Analysis of the Dataset - Distributions

For Document Text as Example:

- We can see the most frequent terms in the word cloud.
- The distributions (which are very similar between Documents and Queries) show a normal distribution.



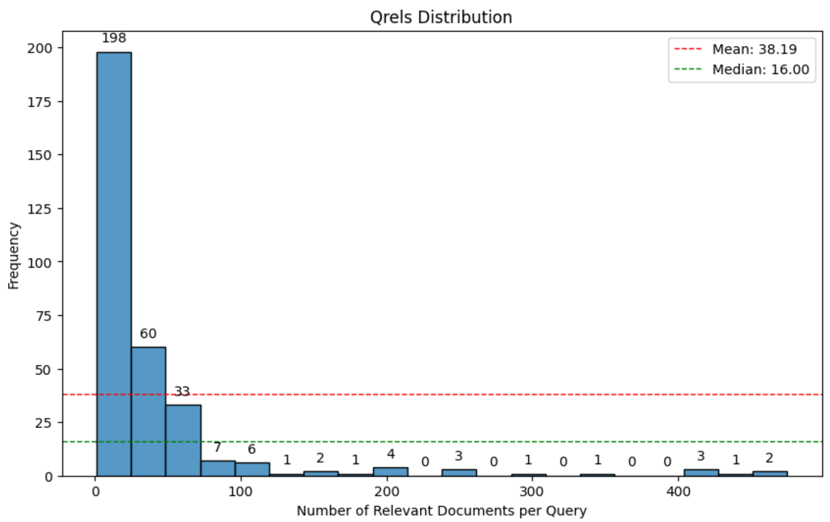
Analysis of the Dataset - QRels



Query ID	Number of Relevant Documents	Doc ID (Example)
PLAIN-681	1	MED-5017
PLAIN-660	475	MED-2509, MED-1374...

For QRels Example:

- We can see the two levels of relevance (scores).
- The distribution shows the number of relevant documents per query.
- The table is an example of Min/Max relevant documents per Query.



UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Pipelines and Experiments

Pipelines and Experiments - Summary

The process of Retrieval Pipelines and Experiments follows these steps :

- Indexing
- Building the Pipeline
- Evaluation
- In-Depth Analysis of Queries
    - Top 10 Queries in Precision@10
    - Inconsistencies in QRels and Relevant Pairs (Not shown here because of length)

Pipelines and Experiments - Indexing

The indexing has been done using two distinct Stemmers:

- **PorterStemmer**
- **SnowballStemmer**, which is an improved version of PorterStemmer with multiple language supported.

And specifically on:

- **Only Document Text**
- **Only Document Titles**
- **Both Document Titles and Text**

```
Number of documents: 3633
Number of terms: 18596
Number of postings: 336960
Number of fields: 2
Number of tokens: 567901
Field names: [title, text]
Positions:   false
```

PorterStemmer

```
Number of documents: 3633
Number of terms: 18596
Number of postings: 336960
Number of fields: 2
Number of tokens: 567901
Field names: [title, text]
Positions:   false
```

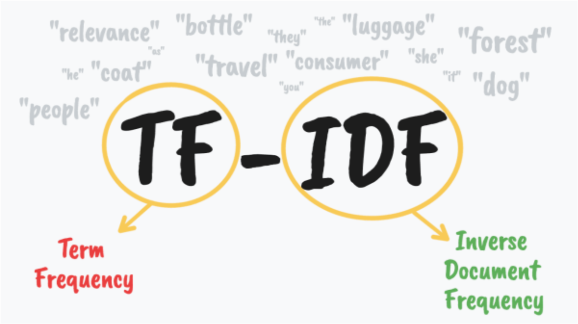
SnowballStemmer

- The two Stemmers behave the same, probably because the dataset contains specifical terms for medical purposes.

Pipelines and Experiments - Models

Two different Models have been used for the Retrieval Pipelines and Evaluation:

- BM25
- TF-IDF



The metrics considered for evaluation purposes are:

- Precision@10
- Recall@10
- Mean Average Precision
- Normalized Discounted Cumulative Gain

Pipelines and Experiments - Results

Indexing Type	Model	P@10	R@10	MAP	NDCG
Only Documents Titles	BM25	0.169	0.109	0.099	0.203
	TF-IDF	0.168	0.109	0.098	0.203
Only Documents Text	BM25	0.226	0.145	0.146	0.295
	TF-IDF	0.227	0.146	0.146	0.295
Both Documents Titles and Text	BM25	0.232	0.150	0.149	0.298
	TF-IDF	0.231	0.148	0.148	0.298
The results shown consider only the SnowballStemmer as the performances are the same as PorterStemmer					

The performances are not particularly high, but we can try to improve them using **Query Expansion**.

Pipelines and Experiments – Top 10 Queries

For every indexer and stemmer we have shown the top 10 queries for **Precision@10**, for example considering **SnowballStemmer and Indexing on both Text and Titles:**

Top 10 Queries by P@10 in BM25 results:

	qid	P@10	query
155	PLAIN-1837	1.0	pesticides
153	PLAIN-1805	1.0	Parkinsons disease
169	PLAIN-1983	1.0	rapamycin
215	PLAIN-2530	1.0	Infectobesity Adenovirus 36 and Childhood Obesity
145	PLAIN-1710	1.0	neurocysticercosis
148	PLAIN-1741	1.0	nuts
63	PLAIN-721	1.0	BMAA
14	PLAIN-153	1.0	How Should I Take Probiotics
57	PLAIN-660	1.0	beans
43	PLAIN-488	0.9	adenovirus 36

BM25 Model

Top 10 Queries by P@10 in TF-IDF results:

	qid	P@10	query
169	PLAIN-1983	1.0	rapamycin
63	PLAIN-721	1.0	BMAA
215	PLAIN-2530	1.0	Infectobesity Adenovirus 36 and Childhood Obesity
155	PLAIN-1837	1.0	pesticides
14	PLAIN-153	1.0	How Should I Take Probiotics
145	PLAIN-1710	1.0	neurocysticercosis
148	PLAIN-1741	1.0	nuts
153	PLAIN-1805	1.0	Parkinsons disease
57	PLAIN-660	1.0	beans
176	PLAIN-2061	0.9	seafood

TF-IDF Model

Pipelines and Experiments – Worst 10 Queries

For every indexer and stemmer we have also shown the worst 10 queries for **Precision@10**, for example considering **SnowballStemmer and Indexing on both Text and Titles:**

Worst 10 Queries by P@10 in BM25 results:

	qid	P@10	query
206	PLAIN-2440	0.0	More Than an Apple a Day Combating Common Diseases
212	PLAIN-2500	0.0	The Saturated Fat Studies Buttering Up the Public
127	PLAIN-1485	0.0	lard
128	PLAIN-1496	0.0	leeks
130	PLAIN-1516	0.0	Lindane
132	PLAIN-1537	0.0	lowcarb diets
133	PLAIN-1547	0.0	lyme disease
115	PLAIN-1353	0.0	hernia
117	PLAIN-1374	0.0	hormonal dysfunction
298	PLAIN-3392	0.0	Healthiest Airplane Beverage

BM25 Model

Worst 10 Queries by P@10 in TF-IDF results:

	qid	P@10	query
130	PLAIN-1516	0.0	Lindane
161	PLAIN-1897	0.0	polypropylene plastic
160	PLAIN-1887	0.0	poisonous plants
159	PLAIN-1877	0.0	plantbased diet
151	PLAIN-1784	0.0	oxen meat
146	PLAIN-1721	0.0	NIHAARP study
144	PLAIN-1700	0.0	Native Americans
143	PLAIN-1690	0.0	National Academy of Sciences
221	PLAIN-2590	0.0	Do Vegetarians Get Enough Protein
217	PLAIN-2550	0.0	Barriers to Heart Disease Prevention

TF-IDF Model

Query Expansion

Query Expansion - Techniques

Provided by **PyTerrier**, we have used the following techniques:

- RM3 Relevance Model
- Bo1 Divergence
- Kullback Leibler Divergence

As an extra experiment we have also tried the **spaCy** Library. SpaCy uses **Word Embeddings** for Query Expansion.

The baseline used for comparation is the best one found in the previous part.

Query Expansion - Results

Query Expansion Technique	Model	P@10	R@10	MAP	NDCG
Original Queries	BM25	0.232	0.150	0.148	0.298
	TF-IDF	0.231	0.148	0.148	0.298
RM3 Expansion	BM25	0.253	0.169	0.173	0.375
	TF-IDF	0.251	0.168	0.173	0.376
Bo1 Divergence Expansion	BM25	0.251	0.165	0.174	0.373
	TF-IDF	0.251	0.165	0.173	0.373
KL Divergence Expansion	BM25	0.250	0.164	0.174	0.373
	TF-IDF	0.250	0.165	0.173	0.374
spaCy Query Expansion	BM25	0.201	0.134	0.130	0.294
	TF-IDF	0.200	0.140	0.129	0.294

Besides **spaCy**, we can see some improvements in the overall performance, with the three techniques provided by PyTerrier producing similar results.

Query Expansion - SpaCy

Query Id	Query
PLAIN-2	Do Cholesterol Statin Drugs Cause Breast Cancer co nothin somethin and that cause space sha havin nuff where
PLAIN-12	Exploiting Autophagy to Live Longer co and that cause vs havin there pm not
PLAIN-23	How to Reduce Exposure to Alkylphenols Through Your Diet dare and that cause havin these nt s
PLAIN-3432	Healthy Chocolate Milkshakes cinnamon diet raspberries sweeteners health dietary
PLAIN-44	Who Should be Careful About Curcumin should somethin that cause need you

We can see that **spaCy** doesn't put very coherent words, probably because of the highly technical language.
-> A possible solution: **SciSpaCy**

Final Considerations

Final Considerations and Future Works

In **conclusion**, the results obtained show little difference but constant improvements. The **average metrics** are not really high but there is room for improvements in possible future works:

- **Neural Re-ranking** using BERT or using **Pseudo Relevance Feedback**.
- **Query expansion** using other techniques such as **LLMs** or using **SciSpaCy**.
- Using different **Models** for the retrieval pipelines.

Extra: ASPIRE Recall Results

	Recall@50	Recall@1000
tfidf only text PorterStemmer	0.208	0.360
tfidf only title PorterStemmer	0.162	0.222
tfidf title text PorterStemmer	0.214	0.365
tfidf only text SnowballStemmer	0.208	0.360
tfidf only title SnowballStemmer	0.162	0.222
tfidf title text SnowballStemmer	0.214	0.365
bm25 only text PorterStemmer	0.210	0.359
bm25 only title PorterStemmer	0.162	0.222
bm25 title text PorterStemmer	0.215	0.363
bm25 only text SnowballStemmer	0.210	0.359
bm25 only title SnowballStemmer	0.162	0.222
bm25 title text SnowballStemmer	0.215	0.363

Higher @K –> Higher Recall

Extra: ASPIRE Precision Results

	P@5	P@10	P@25	P@50	P@100	Rprec
tfidf only text PorterStemmer	0.289	0.227	0.143	0.098	0.064	0.169
tfidf only title PorterStemmer	0.226	0.168	0.103	0.065	0.040	0.118
tfidf title text PorterStemmer	0.297	0.231	0.145	0.100	0.065	0.171
tfidf only text SnowballStemmer	0.289	0.227	0.143	0.098	0.064	0.169
tfidf only title SnowballStemmer	0.226	0.168	0.103	0.065	0.040	0.118
tfidf title text SnowballStemmer	0.297	0.231	0.145	0.100	0.065	0.171
bm25 only text PorterStemmer	0.290	0.226	0.144	0.098	0.064	0.168
bm25 only title PorterStemmer	0.225	0.169	0.104	0.065	0.040	0.118
bm25 title text PorterStemmer	0.298	0.233	0.146	0.100	0.065	0.171
bm25 only text SnowballStemmer	0.290	0.226	0.144	0.098	0.064	0.168
bm25 only title SnowballStemmer	0.225	0.169	0.104	0.065	0.040	0.118
bm25 title text SnowballStemmer	0.298	0.233	0.146	0.100	0.065	0.171

Higher @K –> Lower Precision