



# Information Retrieval

## Information Processing and Retrieval

Diogo Monteiro – up202108800

João Longras – up202108780

José Santos – up202108729

Group 92



# Introduction

The objective of the project is to do an information search system, including work on data collection and preparation, information querying and retrieval, and retrieval evaluation. In this part, the focus is the information retrieval and all its process. From the creation of schemas to the evaluation of the system.

# Documents

## Attributes

- **doc\_i:** Document identifier
- **title:** Article title
- **date:** Publishing date
- **doi:** digital document identifier
- **abstract:** Article abstract




**16000**  
**documents**



# Configuration

To accommodate the diverse requirements of our document indexing and retrieval objectives, we implemented three distinct schemas: the default schema, the lexical schema, and the semantic schema. Each schema was designed to explore different facets of information retrieval, allowing us to evaluate their effectiveness in handling various types of queries.



# Default Schema

- This schema uses **basic tokenization** and **lower case filtering**, with the aim of achieving general purpose retrieval without enhancements. This way, we have a good comparison between a default and a enhanced schema.

Fields	Type	Indexed
doc_id	string	true
title	text	true
date	string	true
doi	string	true
abstract	text	true

# Lemma Schema

- The lemma schema uses the configuration of the default schema as basis, adding a **Stop Words** filter as well as **Synonym Graph Filter Factory** based on a text file obtained by using nltk functions.

Fields	Type	Indexed
doc_id	string	true
title	text	true
date	string	false
doi	string	false
abstract	text	true

# Semantic Schema

- The semantic schema implements dense vector search allowing the search system to locate documents that share semantic relations with the queries. Solr handles these vectors using **DenseVectorField**.
- Due to the scientific aspect of our dataset, we decided to use the model **Specter2** from **allenAI**. This way, we expect better results from such embeddings.

Fields	Type	Indexed
doc_id	string	true
title	text	true
date	string	false
doi	string	false
abstract	text	true
vector	covidVector	true

# Default Query

Parameters	Value
q	query_text
f1	doc_id, title, abstract, score
defType	edismax
qf	title abstract
rows	30
wf	json



# Lemma Query

Parameters	Value
q	query_text
q.op	AND
fl	doc_id, title, abstract, score
defType	edismax
qf	title^3 abstract^2
rows	30
wt	json


# Semantic Query

Parameters	Value
q	<code>{!knn f=vector topK=30}}{embedding}</code>
f1	<code>doc_id, title, abstract, score</code>
rows	30
wf	json



# Evaluation

For our evaluation process, we carried out relevance assessments for each search task. Specifically, we analysed the results presented by the system for each query associated with a given search task by analyzing each result individually to determine its relevance to the respective search task. This process allowed us to create a set of reliable data, which serves as a basis for evaluating the performance of the search system.



# Information Needs

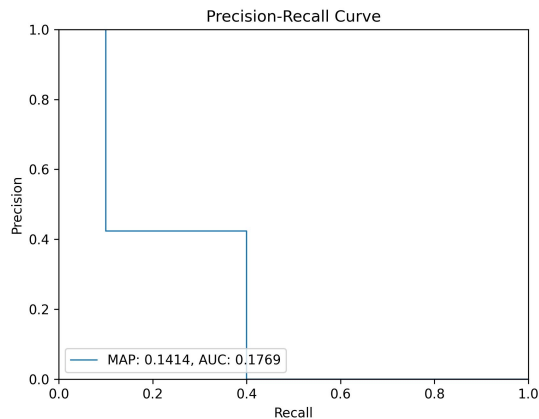
1. Articles that talk about the relationship between fever and fatigue symptoms and influenza

**Relevance:** Someone that might feel a tired and with fever in winter, might want to know if they have some variation of influenza since it is highly contagious at that time of the year.

query\_text: fever fatigue influenza

# Evaluation – Information Need #1

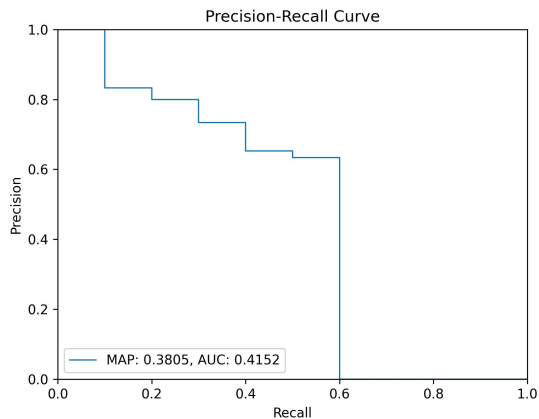
## Default



MAP: 0.14

P@20: 0.40

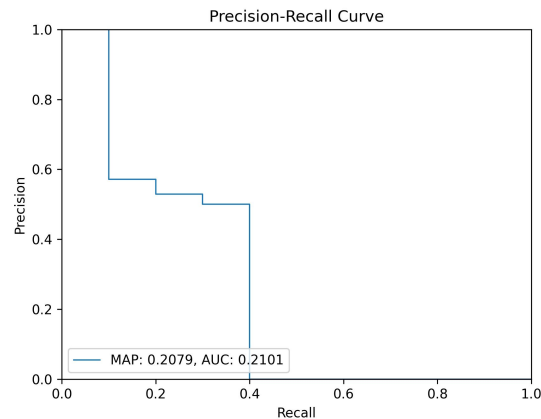
## Lemma



MAP: 0.38

P@20: 0.65

## Semantic



MAP: 0.21

P@20: 0.50

# Evaluation – Information Need #1

As expected, the semantic and lemma schemas gave much better results than the default schema, and it should be noted that the lemma schema gave the best results. This is due to the fact that the results of the semantic schema talk a lot about COVID-19 instead of influenza, since the articles are mostly about COVID-19, and also because of the similarity between the two viruses, which also have common symptoms, there is a proximity of their embeddings in the vector space generated by our scientific model.

# Information Needs

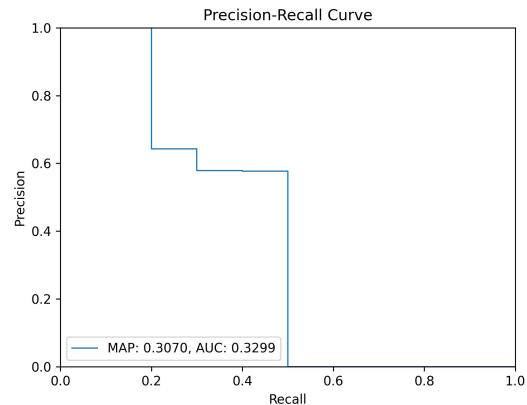
## 2. Find articles that cover the possible effects of Sars-Cov-2 on children

**Relevance:** Even after the pandemic, some cases of Sars-Cov-2(COVID) are still around. If a child catches the virus, a parent might want to know what can happen to the child.

query\_text: effect sars-cov-2 children

# Evaluation – Information Need #2

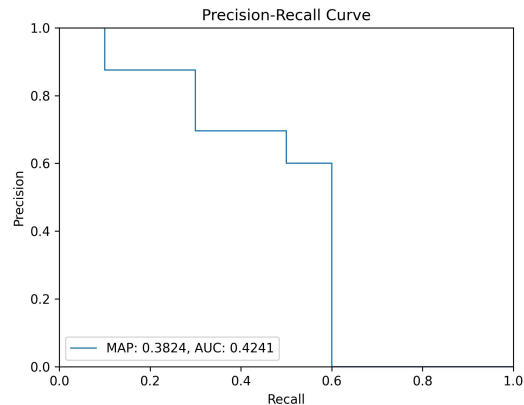
## Default



MAP: 0.31

P@20: 0.55

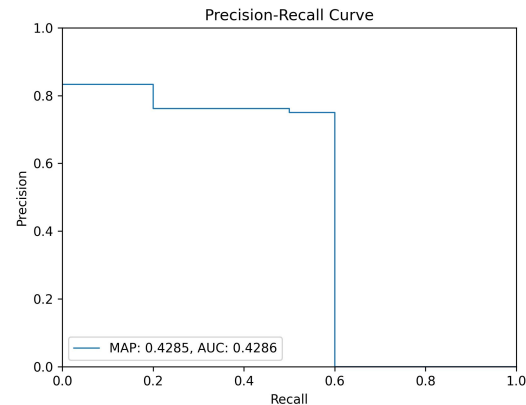
## Lemma



MAP: 0.38

P@20: 0.65

## Semantic



MAP: 0.43

P@20: 0.75



# Evaluation – Information Need #2

The results of this information need are in line with what would be expected for all of them, with the semantic schema having better results than the lemma schema, which in turn has better results than the default schema.

In addition, it was the information need that gave the best results compared to the others, possibly because there were more articles on the same topic but less matching.

# Information Needs

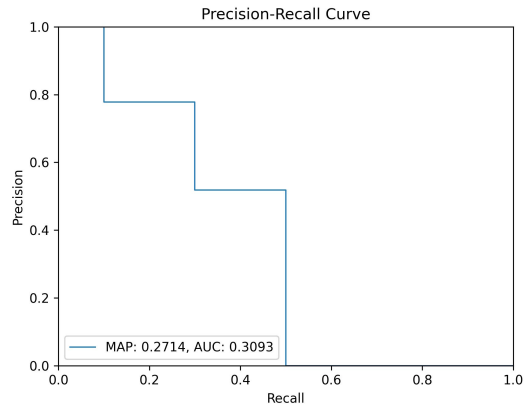
3. Find articles that analyze the effectiveness of masks and social distancing in the mitigation of the spread of a virus.

**Relevance:** Even though the pandemic in 2021 showed that social distancing and the use of masks helped reduce the spread of the coronavirus, some people might want to see the results of scientific research to better understand the motive.

query\_text: masks social distancing spread virus

# Evaluation – Information Need #3

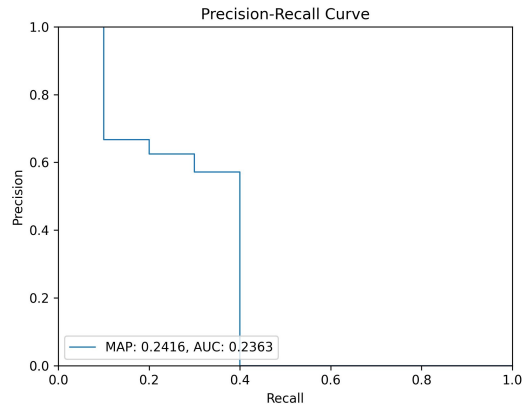
## Default



MAP: 0.27

P@20: 0.50

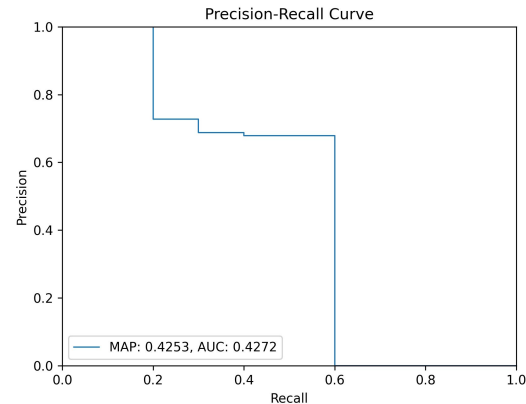
## Lemma



MAP: 0.24

P@20: 0.55

## Semantic



MAP: 0.43

P@20: 0.60

# Evaluation – Information Need #3

As expected, the semantic schema obtained better results than the default and the lemma schemas.

However, although the lemma has a higher P@20 than the default, it has a lower MAP. The reason for this may be that the search for the third information need with the lemma schema is the only one that retrieved a number of documents lower than the requested 30. Thus, this schema found 12 relevant articles out of 21, whereas the default found 14 out of 30.

As for why only 21 articles were retrieved, it's because the third information need was the longest and, together with the 'q.op': 'AND' constraint, it limited the number of articles that could be retrieved.

# Discussion

In conclusion, although in the first information need the semantic schema did not give the expected results because the objective was only influenza, in general it is still the most appropriate schema, and after that the lemma schema gave the best results.

The standard schema, while simple, easy to implement and good for exact matching, doesn't take into account stop words, which can clutter results with irrelevant matches, and doesn't recognise synonyms or contextually similar terms. The lemma schema, on the other hand, solves these problems but still lacks the semantic understanding of content that only the semantic schema provides.