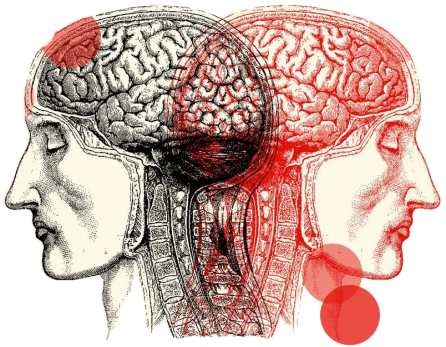
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Francisco Ribeiro | up202104797
Marisa Azevedo | up202108624
Ricardo Ferreira | up201907835
Toni Grgurevic | up202408625

Mental Disorders

Information Retrieval

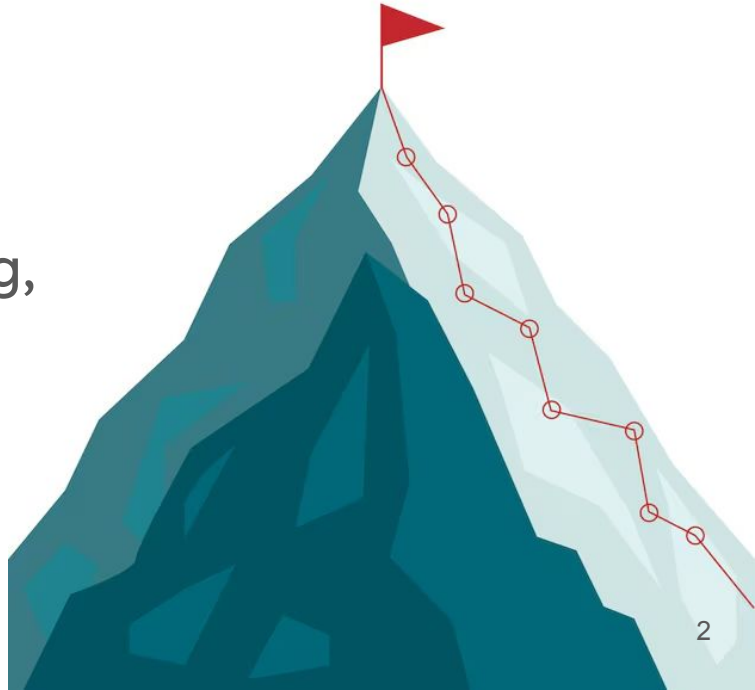


Objectives

The project consists in design and implementation of an information processing and retrieval system of Mental Diseases.

Milestone focus:

- Build an IR system for mental health disorders using Apache Solr.
- Focus on document definition, indexing, and schema creation.
- Search system configuration (queries and respective parameters).
- Evaluate and compare queries for simple schema and enhanced schema.



Document Definition

Overview:

- Each document represents a mental health disorder;
- Includes structured (e.g. page_views) and unstructured data (e.g. symptoms, content).

This is the foundation for our indexing and retrieval system.

Indexing Process

Why Indexing Matters:

- Optimizes search efficiency.
- Essential for handling large datasets.

Key Components in Solr:

- Tokenizers: Break text into searchable units.
- Filters: Refine and normalize tokens for consistency.

Custom Schema:

- Designed to enhance text field indexing (e.g., descriptions, symptoms).
- Metadata and unique identifiers stored but not indexed.

Custom Indexing Analyzers

Field Type: custom_text_general

Steps:

- StandardTokenizerFactory: Isolate words.
- ASCIIFoldingFilterFactory: Normalize characters.
- LowerCaseFilterFactory: Case-insensitive search.
- SynonymGraphFilterFactory: Expand synonyms.
- EnglishMinimalStemFilterFactory: Root word analysis.

Field Type: text_phonetic

Additional:

- PhoneticFilterFactory (Double Metaphone for phonetic similarity) (e.g., "schizophrenia" vs. "scizophrenia").

Custom schema

Field	Type	Indexed
name	text_phonetic	yes
type	string	yes
link	string	no
description	custom_text_general	yes
content	custom_text_general	yes
causes	custom_text_general	yes
symptoms	custom_text_general	yes
treatment	custom_text_general	yes
diagnosis	custom_text_general	yes
epidemiology	custom_text_general	yes
wikidata_id	string	no
wikidata_url	string	no
number_of_revisions	pint	yes
page_views	pint	yes
infobox	custom_text_general	yes

Table 0 - *Schema Field Types*

Retrieval System Configuration

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Query Parser:

- edismax for advanced query handling.

Parameter	Value
q	Cognitive speed
qf	description^3 symptoms^2 causes^2 treatment^1.7 diagnosis^1.5 prevention^1.0 epidemiology^1.5 content^0.5 description^4
pf	symptoms^2 causes^2
ps	2
ps2	1
wt	json
rows	25
fl	name, link, description, symptoms, epidemiology

Table 1 - *Query Parameters*

Evaluation Metrics

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Goals:

- Measure system effectiveness using precision and recall.

Key Metrics:

- Precision at K (P@K): Relevance of top results.
- Average Precision (AvP): Overall precision across ranks.
- Mean Average Precision (MAP): Aggregated AvP across queries.
- Precision-Recall Curves: Stability and performance visualization.

Evaluation Results

Query 1: “Cognitive speed”

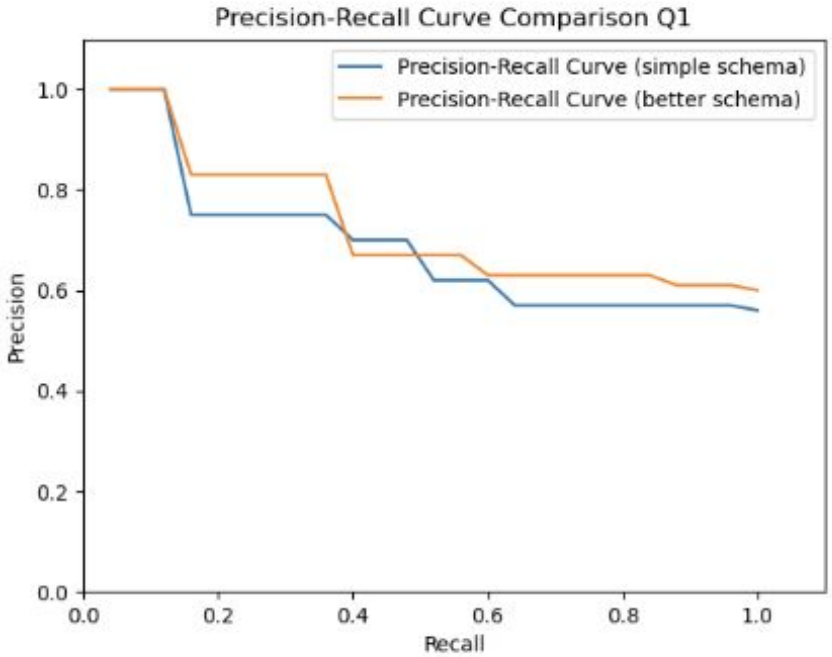


Figure 1 - *Query 1 Plot*

Rank	Syst. Simple	Syst. Complex
AvP	0.64	0.67
P@20	0.56	0.6

Table 2 - *Query 1 Results*

Evaluation Results

Query 2: “ childhood trauma”

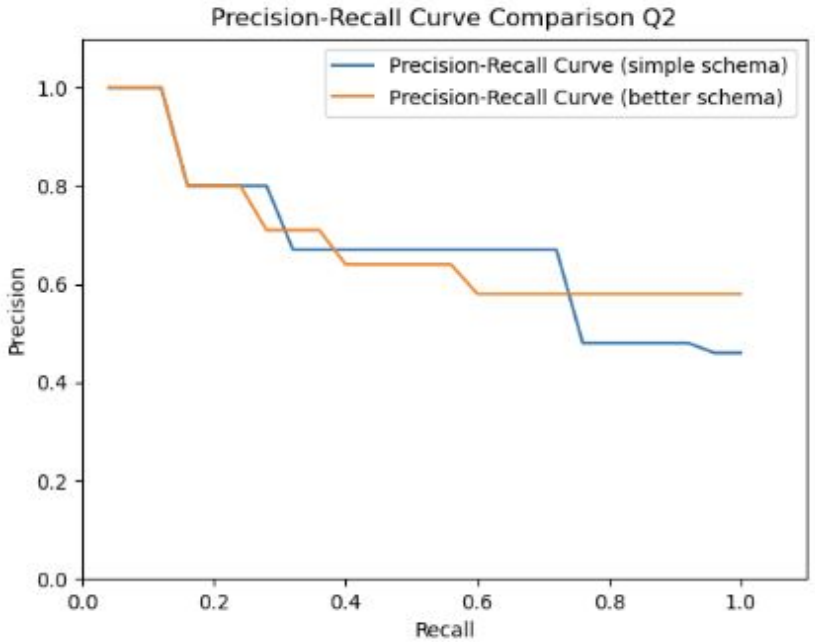


Figure 2 - *Query 2 Plot*

Rank	Syst. Simple	Syst. Complex
AvP	0.6	0.61
P@20	0.44	0.44

Table 3 - *Query 2 Results*

Evaluation Results

Query 3: “ Improvement with behavioral therapies”

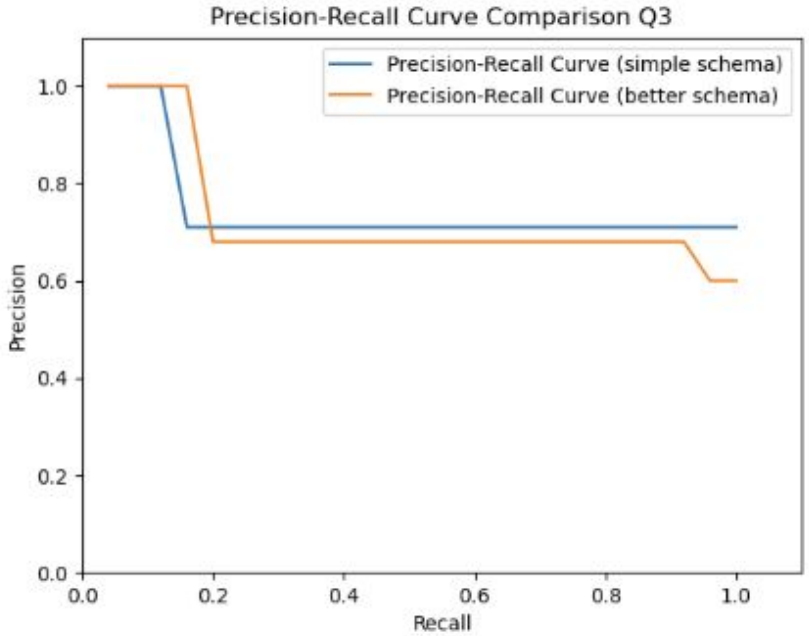


Figure 3 - *Query 3 Plot*

Rank	Syst. Simple	Syst. Complex
AvP	0.64	0.61
P@25	0.68	0.6

Table 4 - *Query 3 Results*

Evaluation Results

Query 4: “Frequent on children”

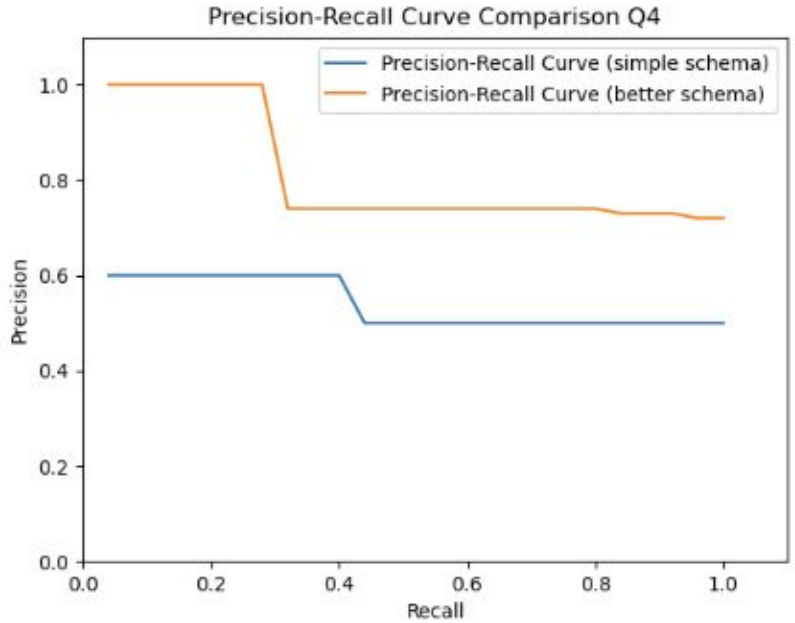


Figure 4 - *Query 4 Plot*

Rank	Syst. Simple	Syst. Complex
AvP	0.41	0.74
P@25	0.48	0.72

Table 5 - *Query 4 Results*

Evaluation Results

Query 5: “caused by genetics inherited.”

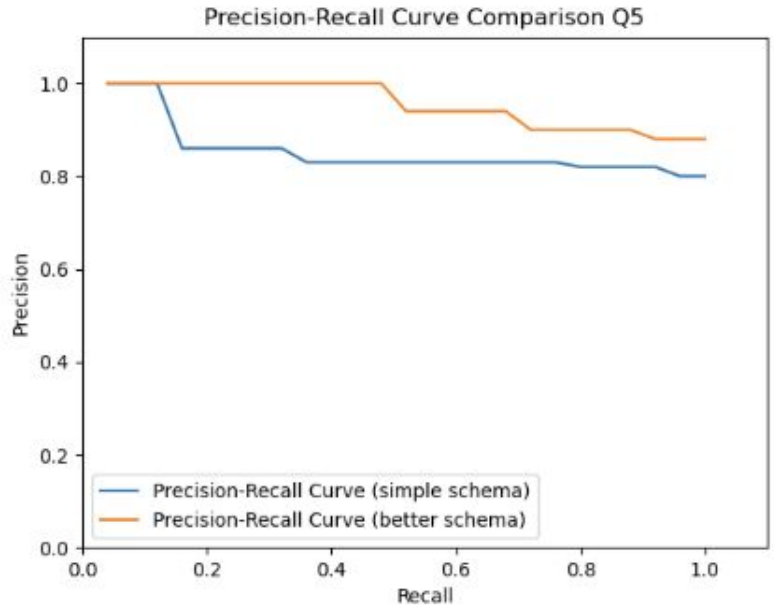


Figure 5 - *Query 5 Plot*

Rank	Syst. Simple	Syst. Complex
AvP	0.81	0.94
P@25	0.8	0.88

Table 6 - *Query 5 Results*

Comparative Evaluation

Global	Syst. Simple	Syst. Complex
MAP	0.62	0.714

Table 7 - *MAP Scores Global*

Conclusion and Future Work

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Achievements:

- Successful implementation of an information retrieval system for mental health data.
- Demonstrated value of custom schema and advanced analyzers

Key Takeaways:

- Complex schema improves relevance but requires balanced optimization.

Next Steps:

- Develop a user interface for enhanced interaction.
- Improve usability and information retrieval quality for mental health data.

The end of the Powerpoint

Thanks for the attention. Do you have any question?