PRIMED

A Medicine Search System

Milestone #2: Information Retrieval

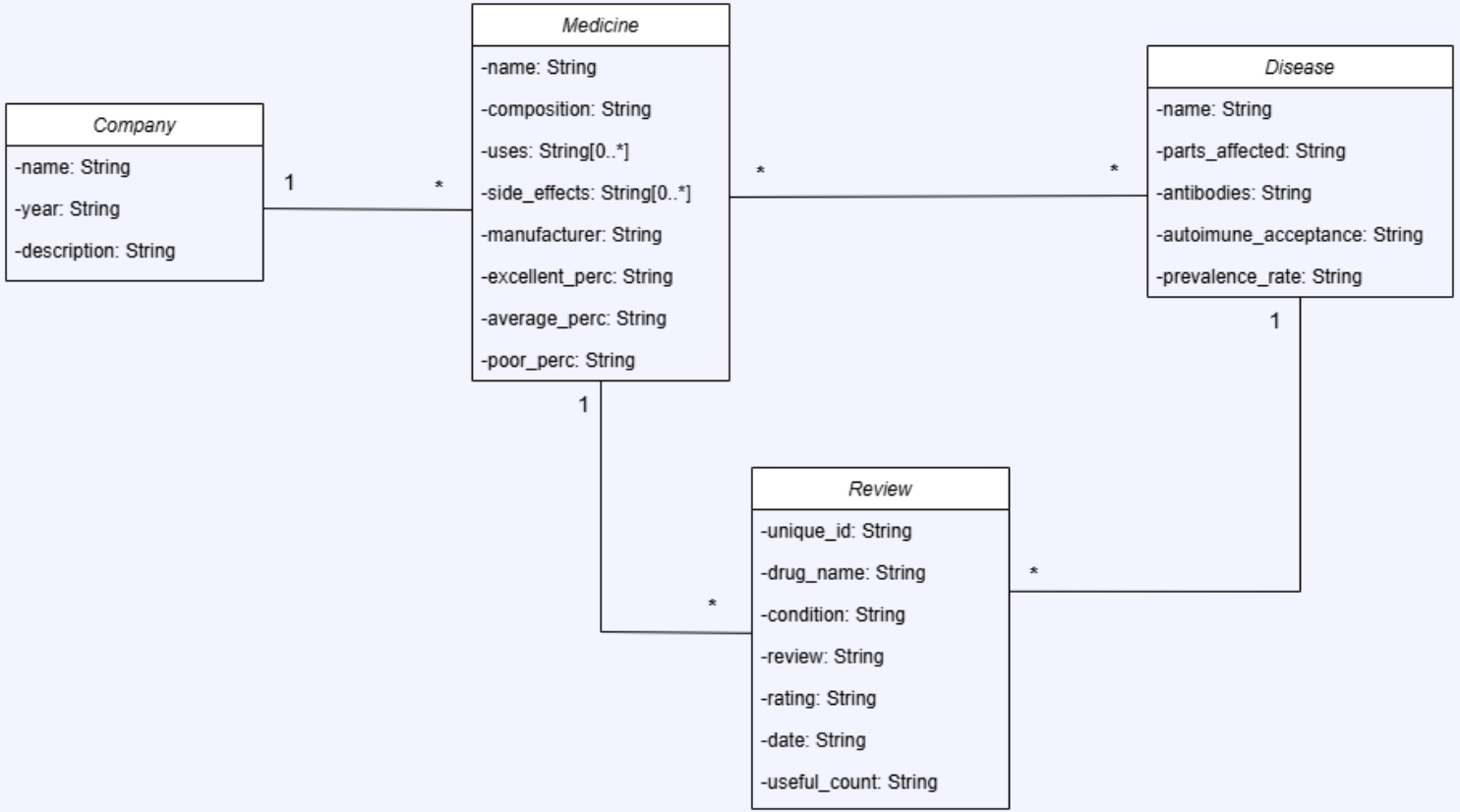
Pedro Simões, up202403063@up.pt
Miguel Garrido, up202108889@up.pt
Emanuel Maia, up202107486@up.pt
Guilherme Martins, up202403106@up.pt

Table of Contents

01

Milestone #1 Review

Conceptual Data Model



Information Needs

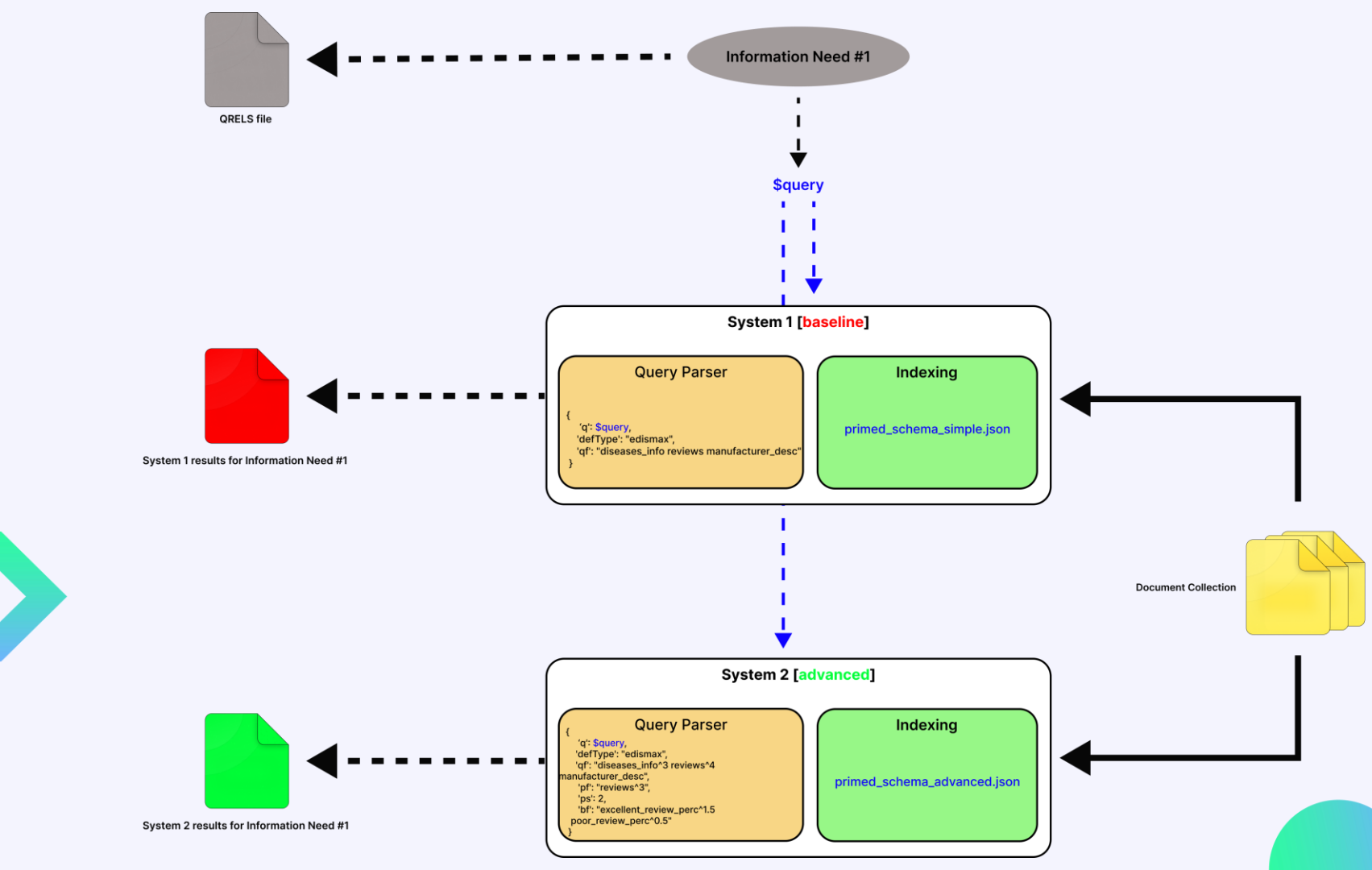
- Which medicines are more commonly used to treat the common cold?
- Trustworthy companies that provide medicines for Alzheimer's disease.
- How does the composition of medicine for diabetes vary between manufacturers like Novo Nordisk, Eli Lilly and Sanofi?
- What are the most effective treatments for managing rheumatoid vasculitis pain and inflammation?
- What is the best treatment for persistent migraine symptoms, including nausea and light sensitivity?
- Is weight gain a common side effect of antidepressants?
- What side effects have other patients experienced with medications like rituximab or methotrexate for treating vasculitis?

02

System Overview

Our system for querying and data analysis, based in Apache Solr, is organized into three main components:

- **Document Indexing**: Indexing the JSON document collection using preconfigured schemas
- **Querying**: Scripts for sending queries to Solr with different parameters
- **Evaluation**: Generation and evaluation of results in TREC format.


QRELS file
Information Need #1
$query
System 1 [baseline]
Query Parser
{
'q': $query,
'defType': "edismax",
'qf': "diseases_info reviews manufacturer_desc"
}
Indexing
primed_schema_simple.json
System 1 results for Information Need #1
Document Collection
System 2 [advanced]
Query Parser
{
'q': $query,
'defType': "edismax",
'qf': "diseases_info^3 reviews^4
manufacturer_desc",
'pf': "reviews^3",
'ps': 2,
'bf': "excellent_review_perc^1.5
poor_review_perc^0.5"
}
Indexing
primed_schema_advanced.json
System 2 results for Information Need #1

03

Document Analysis

Each document represents a specific medicine:

Field	Description
drug	Name of the medicine
composition	Active substance(s) present
applicable_diseases	Associated diseases
diseases_info	Information on the associated diseases
possible_side_effects	Possible side effects provoked by the medication
excellent_review_perc	% of excellent reviews (score > 7)
average_review_perc	% of average reviews (4 ≤ score ≤ 7)

Field	Description
poor_review_perc	% of poor reviews (score < 4)
reviews_average_rating	Average review score of the medicine
reviews	User reviews for medicines
manufacturer	Name of the company that produces the medicine
manufacturer_desc	Short description of the manufacturer
manufacturer_start	Year the company was founded
manufacturer_end	Year the company was shut down (if applicable)

Field	Type	Indexed	Multi-Valued
drug	shortText	✓	✗
composition	shortText	✓	✗
applicable_diseases	shortText	✓	✓
diseases_info	diseasesBoosted	✓	✓
possible_side_effects	shortText	✓	✓
excellent_review_perc	pdouble	✓	✗
average_review_perc	pdouble	✓	✗
poor_review_perc	pdouble	✓	✗
reviews_average_rating	pdouble	✓	✗
reviews	textBoosted	✓	✓
manufacturer_desc	textBoosted	✓	✗
manufacturer	text_general	✓	✗
manufacturer_start	text_general	✗	✗
manufacturer_end	text_general	✗	✗

Table 1: Advanced Schema

Baseline System

Parameter	Value
q	$query
q.op	AND
sort	reviews_average_rating desc
start	0
rows	30
qf	diseases_info reviews manufacturer_desc

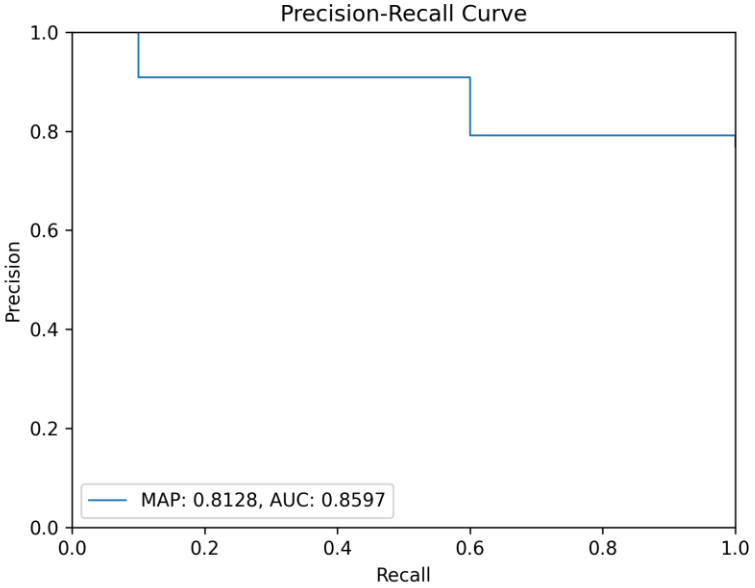
Advanced System

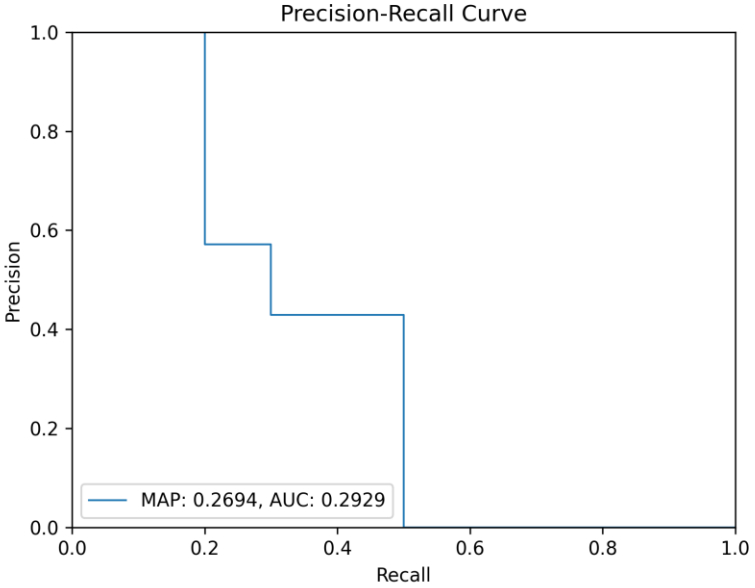
Parameter	Value
q	$query
q.op	AND
start	0
rows	30
qf	diseases_info^3 reviews^4 manufacturer_desc
pf	reviews^3
ps	2
bf	excellent_review_perc^1.5 poor_review_perc^0.5

04

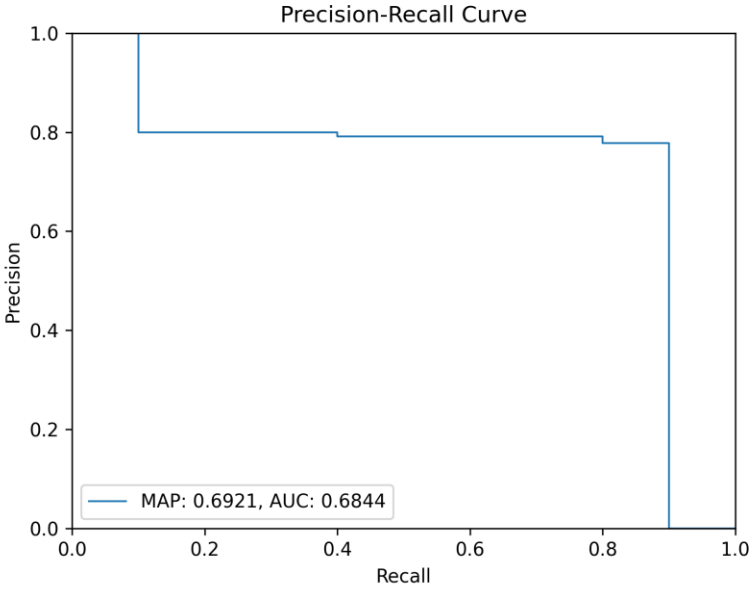
System Evaluation

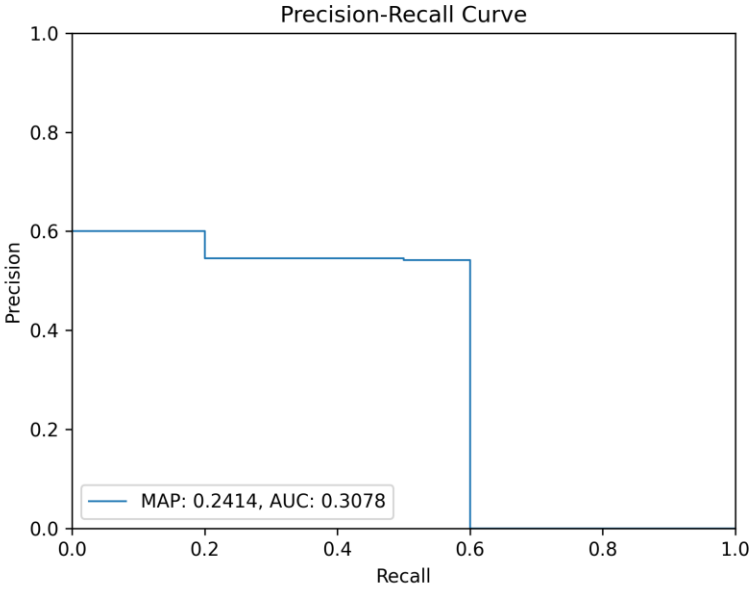
Query #1


Precision-Recall Curve
Precision
Recall
MAP: 0.8128, AUC: 0.8597
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

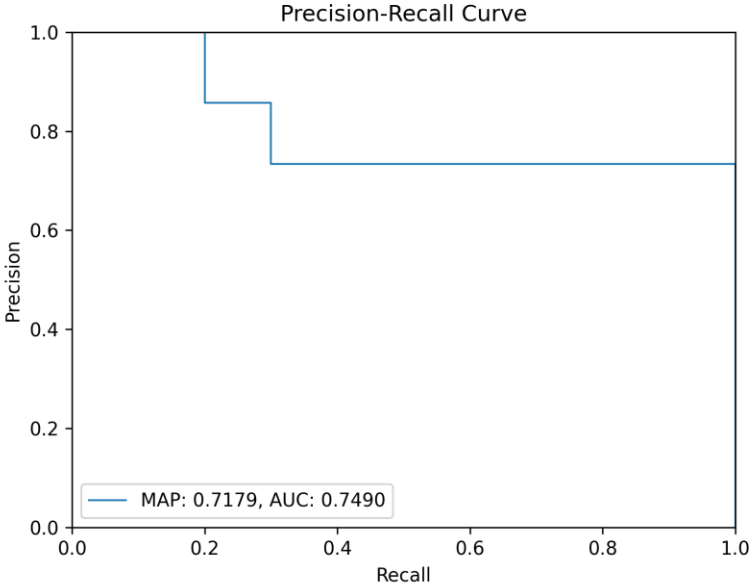

Precision-Recall Curve
Precision
Recall
MAP: 0.2694, AUC: 0.2929
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

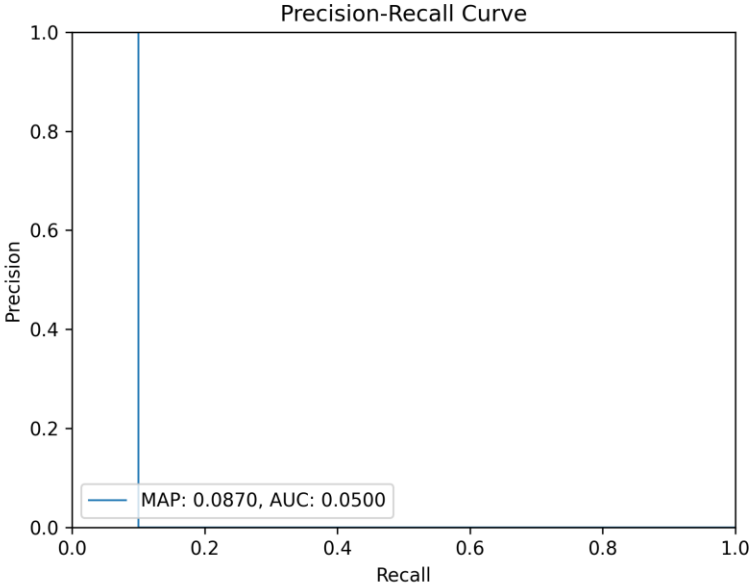
Query #2


Precision-Recall Curve
Precision
Recall
MAP: 0.6921, AUC: 0.6844
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

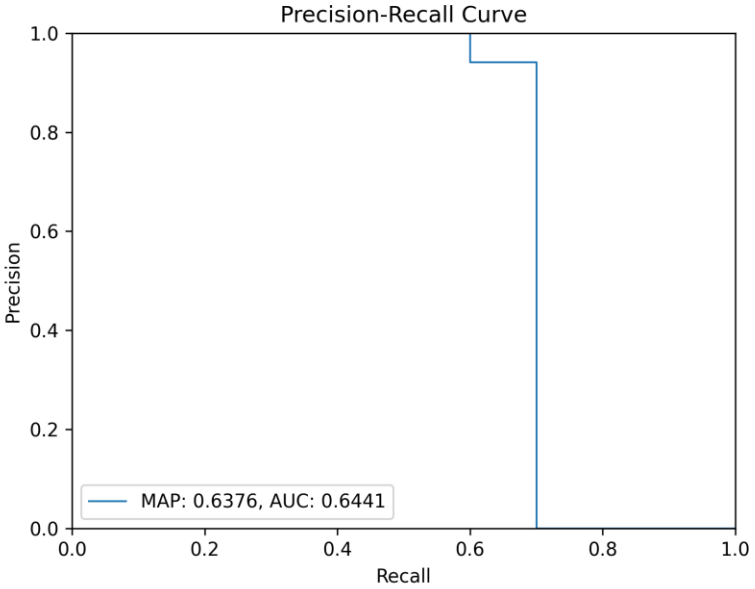

Precision-Recall Curve
Precision
Recall
MAP: 0.2414, AUC: 0.3078
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

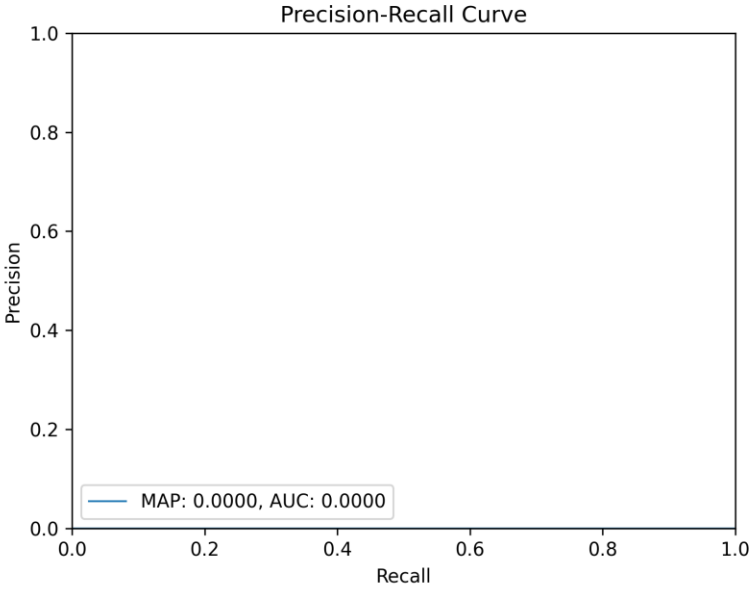
Query #3


Precision-Recall Curve
Precision
Recall
MAP: 0.7179, AUC: 0.7490
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve
Precision
Recall
MAP: 0.0870, AUC: 0.0500
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Query #4


Precision-Recall Curve
Precision
Recall
MAP: 0.6376, AUC: 0.6441
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve
Precision
Recall
MAP: 0.0000, AUC: 0.0000
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

References

1. **Sematext**. Getting started with apache solr. https://sematext.com/guides/solr/.solr introduction and explanation
2. **Apache Software Foundation**. Solr's schema file. https://solr.apache.org/guide/6_6/overview-of-documents-fields-and-schema-design.html#solr-s-schema-file,2017.
3. **MDN**. Http request methods. https://developer.mozilla.org/en-US/docs/Web/HTTP/Methods, 2024.
4. **Apache Software Foundation**. Tokenizers. https://solr.apache.org/guide/solr/latest/indexing-guide/tokenizers.html.
5. **Apache Software Foundation**. Filters. https://solr.apache.org/guide/solr/latest/indexing-guide/filters.html.
6. **Apache Software Foundation**. The extended dismax query parser. https://solr/.apache.org/guide/6_6/the-extended-dismax-query-parser.html, 2017.
7. **Apache Software Foundation**. The dismax query parser. https://solr.apache.org/guide/6_6/the-dismax-query-parser.html, 2017.
8. **Apache Software Foundation**. Using 'slop'. https://solr.apache.org/guide/8_6/the-extended-dismax-query-parser.html#using-slop, 2020.
9. **Sérgio Nunes**. Understanding relevance judgements. https://gitlab.up.pt/pri/tutorials/-/blob/main/06-evaluation/README.md?ref_type=heads#2-understanding-relevance-judgements, 2024.
10. **Keylabs**. Understanding precision at k (p@k). https://keylabs.ai/blog/understanding-precision-at-k-p-k/.
11. **Deval Shah**. Mean average precision (map) explained: Everything you need toknow. https://www.v7labs.com/blog/mean-average-precision, 2022.
12. **Doug Steen**. Precision-recall curves. https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248, 2020.
13. **Amber Roberts**. What is pr auc? https://arize.com/blog/what-is-pr-auc/, 2022.