U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Information Processing and Retrieval

Master in Informatics Engineering and Computation at FEUP, U.Porto

Diogo Fonte
Rodrigo Figueiredo
Sofia Rodrigo
Vitor Cavaleiro

Information Retrieval Tools

- Near Real-Time Indexing
- Flexible and Adaptable with easy configuration
- Comprehensive Administration Interfaces
- Standards Based Open Interfaces - XML, JSON and HTTP
- Advanced Full-Text Search Capabilities

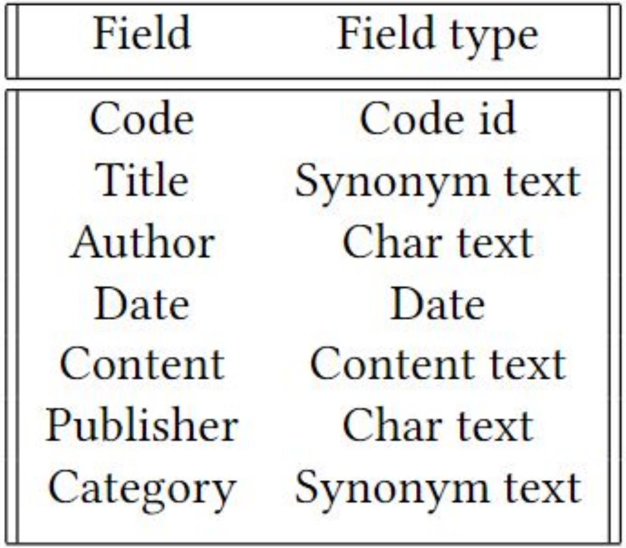
Collection

- startup.sh builds docker image with 1 core
- schema.sh creates the schema for the news collection via API
- populate.sh script populates the core with the json file content

Indexing - Schema

Filters:

- ASCII Folding Filter
- Lower Case Filter
- Mapping Char Filter
- Synonym Graph Filter
- Porter Stem Filter
- English Possessive Filter
- Hyphenated Words Filter
- Stop filter



Field	Field type
Code	Code id
Title	Synonym text
Author	Char text
Date	Date
Content	Content text
Publisher	Char text
Category	Synonym text

Retrieval

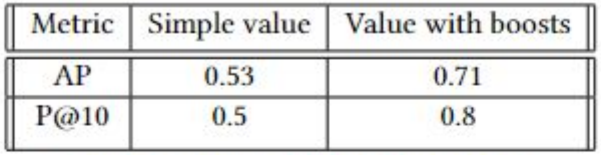
- **Fields boosts:** enhance the relevance of specific fields.
- **Term boosts**: elevate the importance of certain words.
- **Proximity searches**: emphasizes the closeness of terms.
- **Wildcards/Fuzziness**: allows for flexibility in matching variations of a term

Evaluation

- Metrics:
  - Average Precision (AP)
  - Precision at 10 (P@10)
- Manual construction of qrels.txt files.

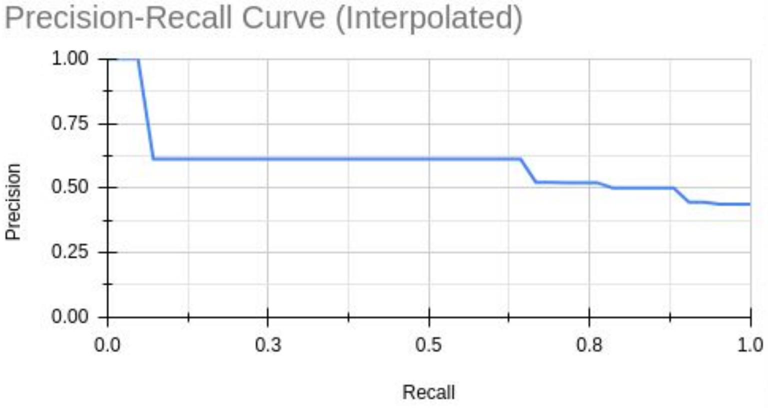
Evaluation - Query 1

Find news articles where Trump spoke on the immigration crisis

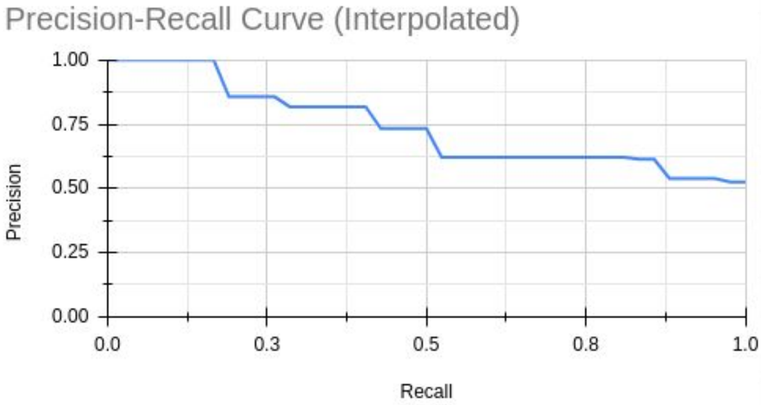


Metric	Simple value	Value with boosts
AP	0.53	0.71
P@10	0.5	0.8

Table 3: Precision metrics for Q1



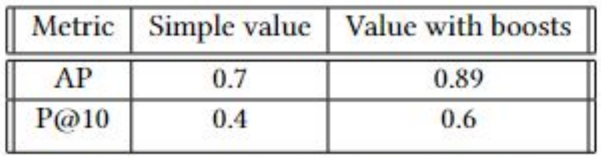
(simple schema without boosts)



(improved schema with boosts)

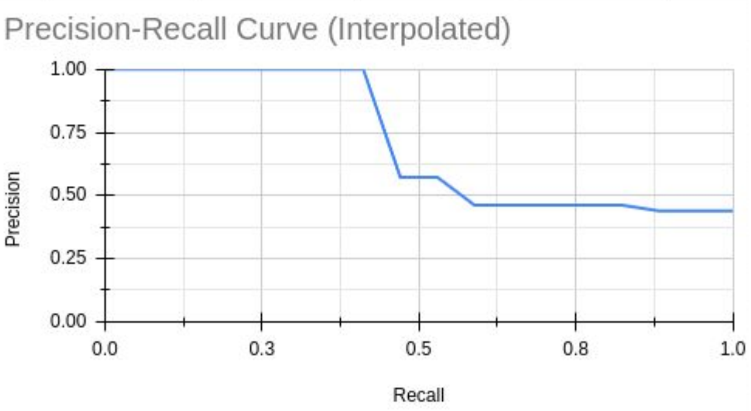
Evaluation - Query 2

Find news about LeBron's good performances in games his team won

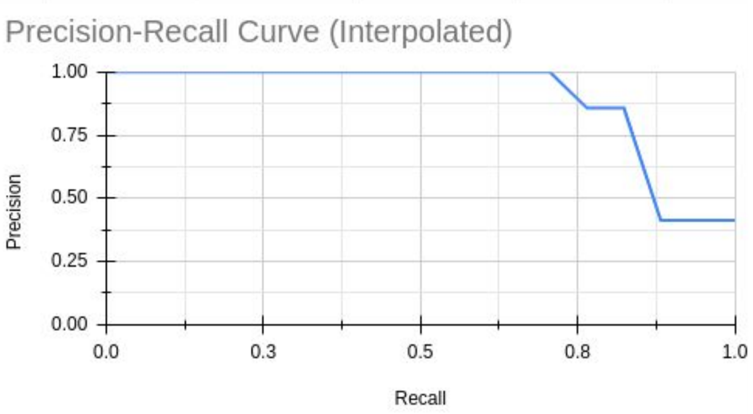


Metric	Simple value	Value with boosts
AP	0.7	0.89
P@10	0.4	0.6

Table 5: Precision metrics for Q2



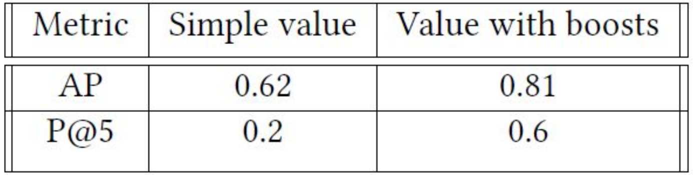
(simple schema without boosts)



(improved schema with boosts)

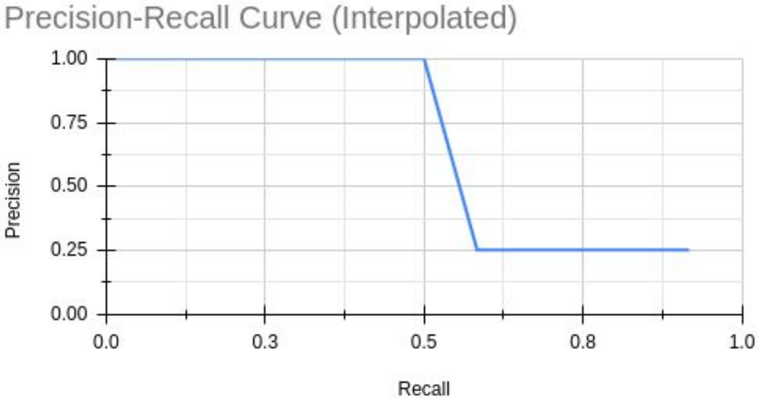
Evaluation - Query 3

Find articles related to homicides investigated by the FBI in 2017



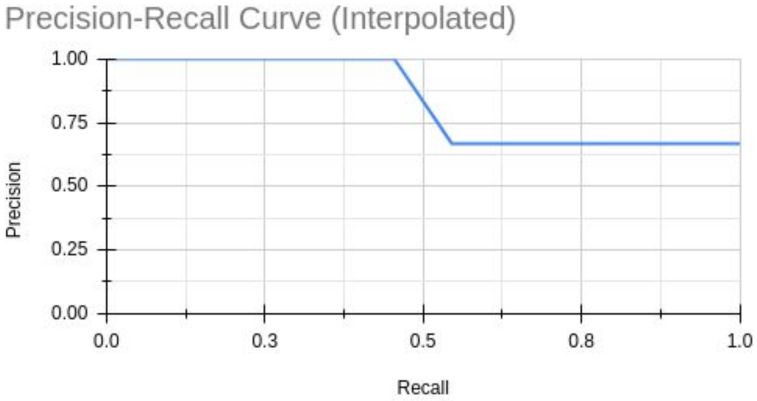
Metric	Simple value	Value with boosts
AP	0.62	0.81
P@5	0.2	0.6

Table 7: Precision metrics for Q3



Precision-Recall Curve (Interpolated)

(simple schema without boosts)

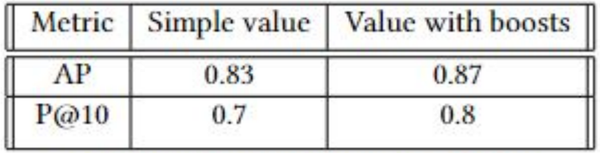


Precision-Recall Curve (Interpolated)

(improved schema with boosts)

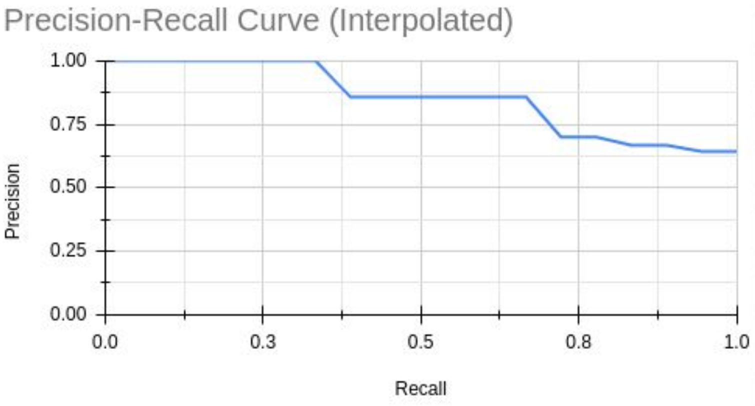
Evaluation - Query 4

Find news articles regarding the conflicts between republicans and democrats about gun ownership

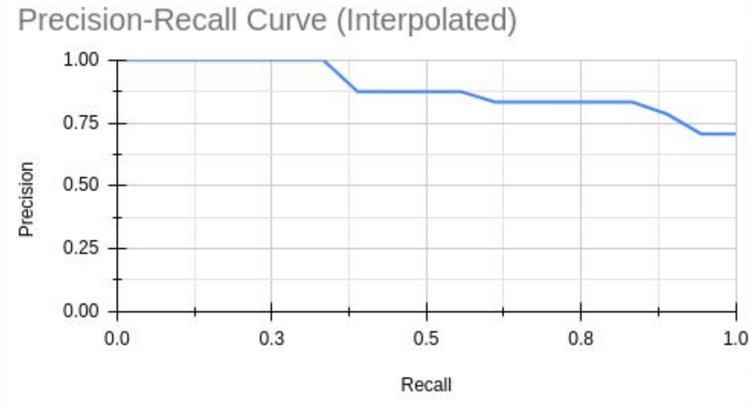


Metric	Simple value	Value with boosts
AP	0.83	0.87
P@10	0.7	0.8

Table 9: Precision metrics for Q4



(simple schema without boosts)



(improved schema with boosts)