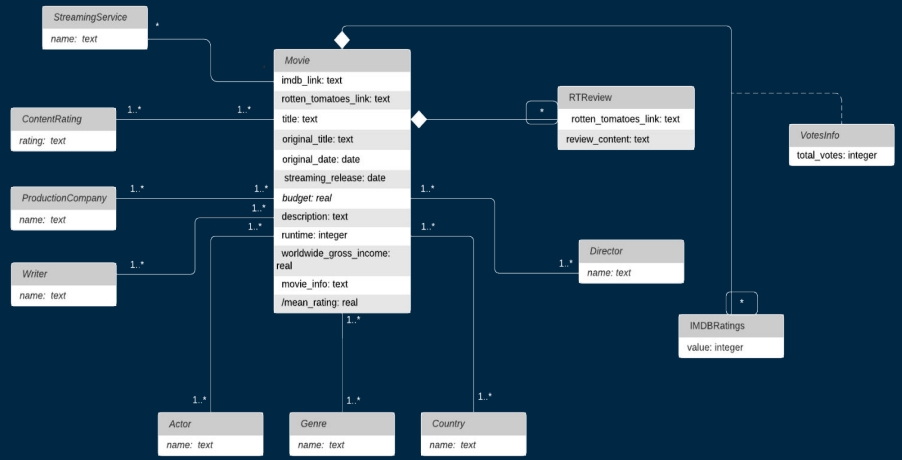
# MOVIE RATINGS AND REVIEWS

PRI 2021/2022

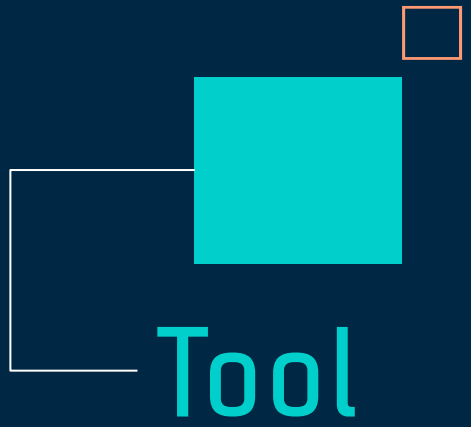
Caio Nogueira
Carlos Lousada
Tomás Mendes

# MILESTONE 1 OVERVIEW

- Data collected in Kaggle scrapped from IMDB and RottenTomatoes;
- Dataset contains over 9,000 movies with up to 20 reviews per movie (after preprocessing);
- Statistical analysis was performed to better understand the dataset



# Collection




## Solr x Elasticsearch

We decided to stick with Solr which has the most adequate use cases for our problem


## “Movies” document

One single document type with all the necessary information for information retrieval phase

# Indexing Process - Custom Field types

| Field Type | Filter | Index | Query |
|---|---|---|---|
| standard_text | ASCIIFoldingFilterFactory | X | X |
| | LowerCaseFilterFactory | X | X |
| | SynonymGraphFilterFactory | X | X |
| daterange | DateRangeField | | |

# Indexing Process -Fields

| Field | Field Type | Index |
|---|---|---|
| original_title | standard_text | true |
| original_release_date | daterange | true |
| runtime | pint | true |
| mean_vote_imdb | pfloat | true |

...

# Information Retrieval

To compare results, we used 3 different systems:

- **Schemaless;**
- **With a Schema** which was defined in the indexing process;
- Using a schema and applying **weights.**

# Information Needs

1. Space movies available on Netflix
2. Movies about slavery
3. Emotive movies about World War II
4. Movies about true crime stories
5. Christmas movies for the family

# Information Needs

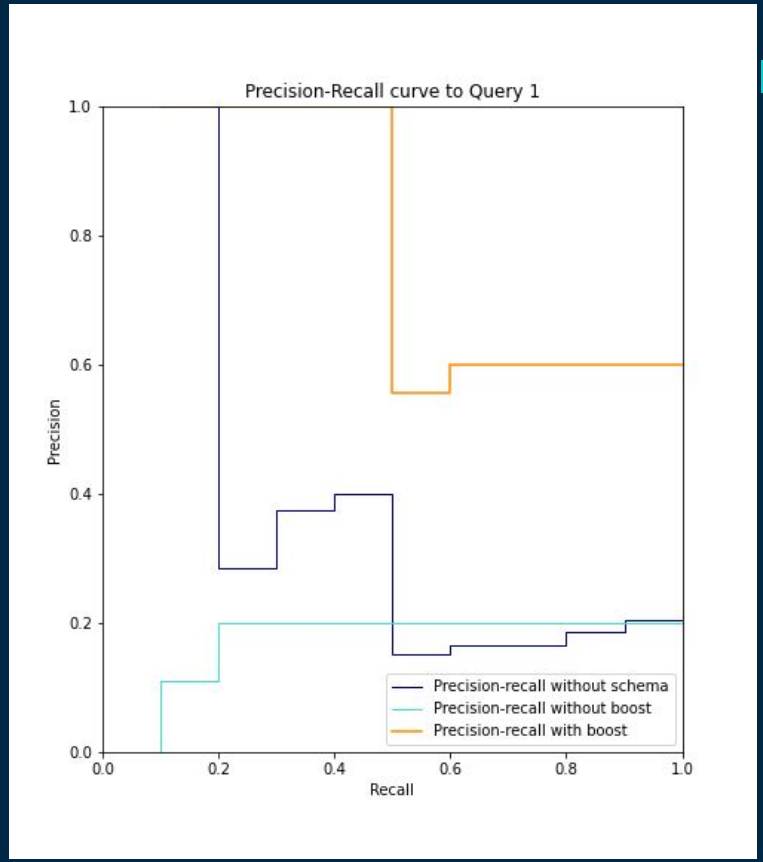
**IN1 - Space movies available on Netflix**

**Query (q):** space sci-fi

**Filter query (fq)**: available_netflix: "True"

**Query filters (qf):** original_title^10, movie_info^50, review_content^20

| | Schemaless | Schema | Boosted |
|---|---|---|---|
| P@10 | 0.36 | 0.27 | 0.92 |
| AvP | 0.4 | 0.2 | 0.6 |



# Information Needs

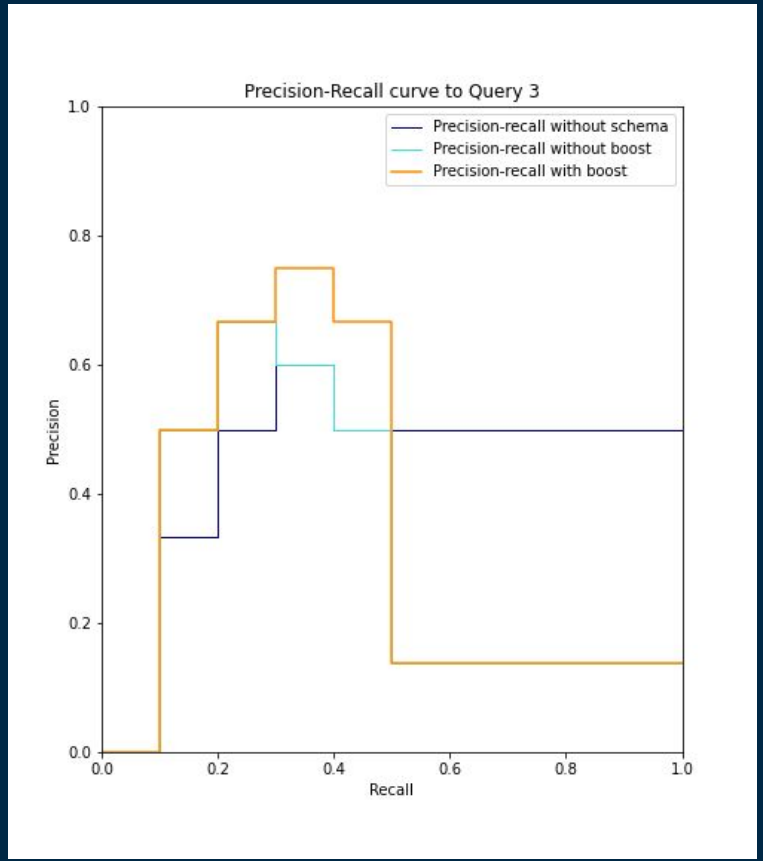
**IN3 - Emotive movies about World War II**

**Query (q):** "World war II" emotional

**Filter query (fq):** mean_vote_imdb: [8.0 TO 10.0]

**Query fields (qf):** movie_info^20, review_content^40

| | Schemaless | Schema | Boosted |
|---|---|---|---|
| P@10 | 0.5 | 0.5 | 0.5 |
| AvP | 0.564444 | 0.627778 | 0.68619 |



# Information Needs

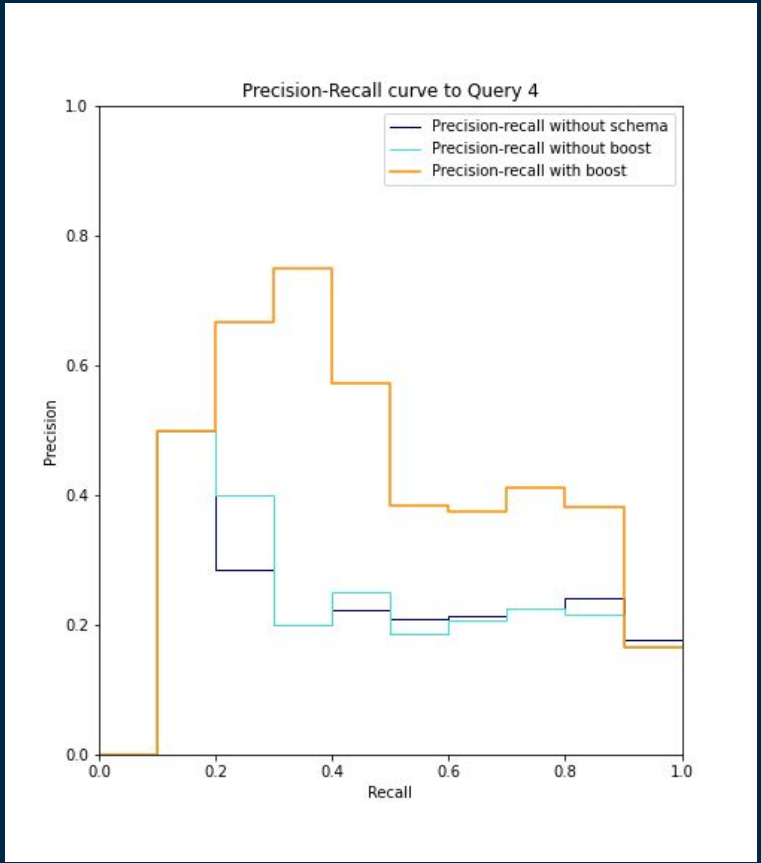
**IN4 - Movies about true crime stories**

**Query (q):** true crime story

**Query fields (qf):** movie_info^30, review_content^50

**Phrase Slop (ps):** 3

| | Schemaless | Schema | Boosted |
|---|---|---|---|
| P@10 | 0.3 | 0.3 | 0.5 |
| AvP | 0.33899 | 0.349472 | 0.559504 |

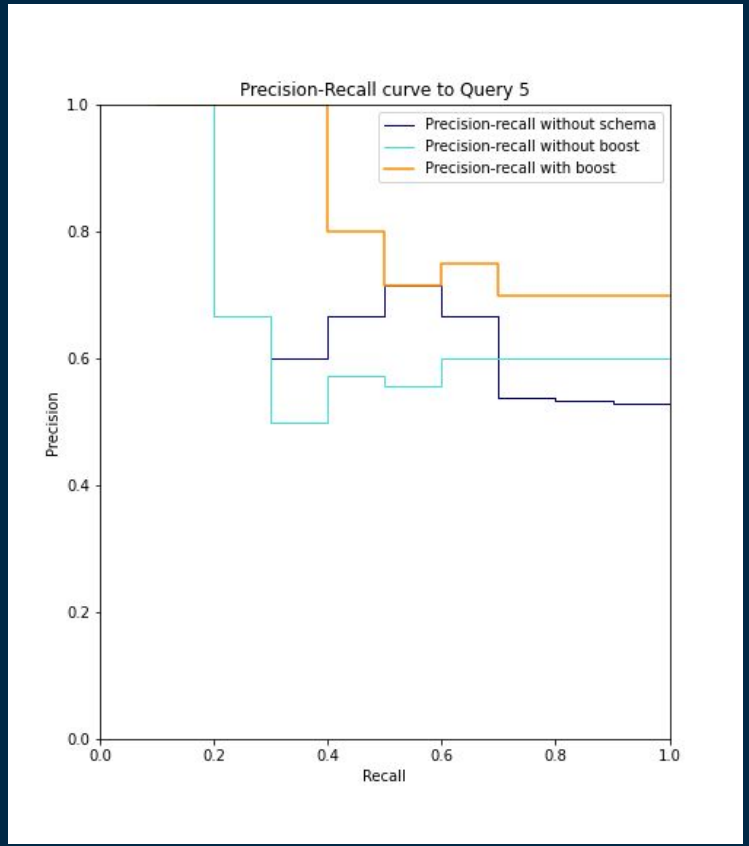


# Information Needs

**IN5 - Christmas movies for the family**

**Query (q):** Christmas AND time
**Filter query (fq):** genres: "Kids & Family"
**Query fields (qf):** original_title^40, movie_info^30, review_content^20
**Phrase Slop (ps):** 5

| | Schemaless | Schema | Boosted |
|---|---|---|---|
| P@10 | 0.7 | 0.7 | 0.7 |
| AvP | 0.696715 | 0.696715 | 0.893519 |



# Conclusions

**Mean Average Precision** (5 information needs)

| Schemaless | Schema | Boosted |
|---|---|---|
| 0.4655424 | 0.4398454 | 0.6419682 |

# Future Work

- Improve search engine - develop new features and techniques;
- Implement a graphical interface for the Information Retrieval System.