

Home

TV Shows

Movies

News & Popular

My list

Watchlist



INFORMATION RETRIEVAL MILESTONE 2



PAULO RIBEIRO, UP201806505
PEDRO FERREIRA, UP201806506
PEDRO PONTE, UP201809694



| TABLE OF CONTENTS

01

Collection

- Definition
- Document

02

Indexing

- Preparing Solr

03

Schema

- Definition
- Tokenizers & Filters

04

Retrieval

- Queries results

05

Evaluation

- Metrics
- Results relevance

06

Future Work & Improvements

- Milestone 3 goals



Home

TV Shows

Movies

News & Popular

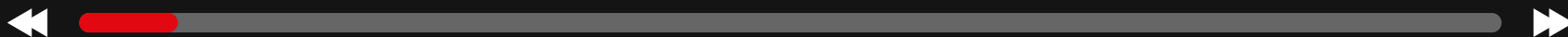
My list

Watchlist



01

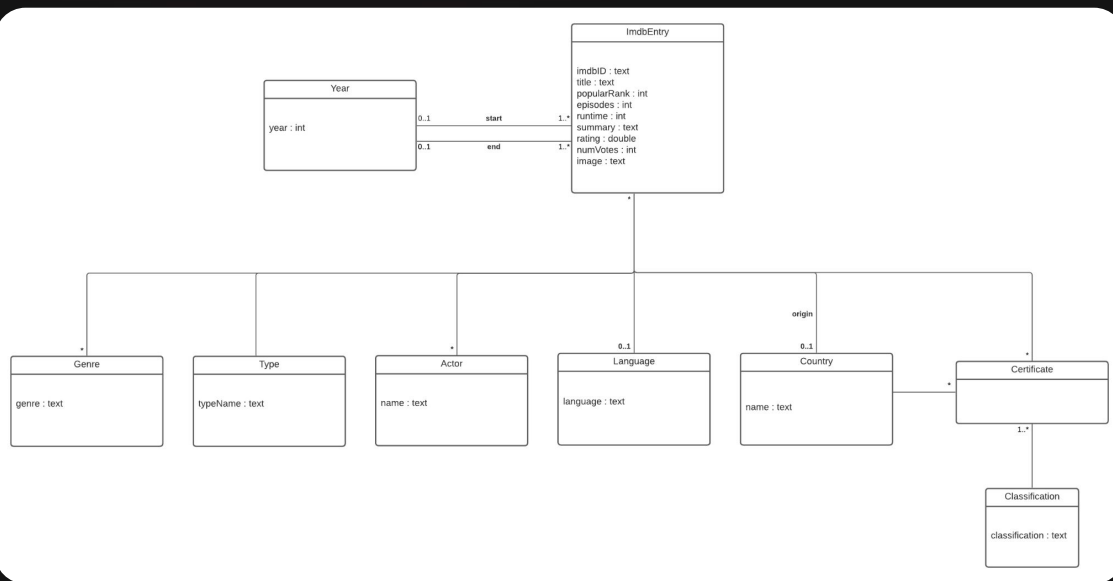
Collection





| Collection

- **Collection:** All shows in the dataset obtained from Milestone 1
- **Document:** A show
- The dataset is a single CSV file which, for convenience, we transform into a JSON file



Home

TV Shows

Movies

News & Popular

My list

Watchlist



02

Indexing





| Indexing

- Run Docker image based on Solr's container
- Dockerfile to define a custom image and perform all the basic setup operations
- Script to automatically create a core (shows), define our schema and populate the database

```
PRI-G34 > src > solr > Dockerfile
1 FROM solr:8.10
2
3 COPY netflix_list.json /data/netflix_list.json
4
5 COPY query1.json /data/query1.json
6
7 COPY query2.json /data/query2.json
8
9 COPY query3.json /data/query3.json
10
11 COPY query4.json /data/query4.json
12
13 COPY schema.json /data/schema.json
14
15 COPY startup.sh /scripts/startup.sh
16
17 ENTRYPOINT ["/scripts/startup.sh"]
18
```

```
PRI-G34 > src > solr > startup.sh
1 #!/bin/bash
2
3 precreate-core shows
4
5 # Start Solr in background mode so we can use the API to upload the schema
6 solr start
7
8 sleep 2
9
10 # Schema definition via API
11 curl -X POST -H 'Content-type:application/json' \
12 --data-binary @/data/schema.json \
13 http://localhost:8983/solr/shows/schema
14
15 sleep 2
16
17 # Populate collection
18 bin/post -c shows /data/query4.json
19
20 # Restart in foreground mode so we can access the interface
21 solr restart -f
```



Home

TV Shows

Movies

News & Popular

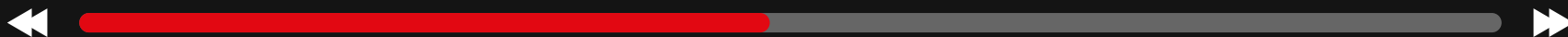
My list

Watchlist



03

Schema





| Schema

- Create some field-types:

- id
- title
- certificate
- type_plural
- lower_clean
- summary
- cast

```
{
  "name": "title",
  "class": "solr.TextField",
  "indexAnalyzer": {
    "tokenizer": {
      "class": "solr.StandardTokenizerFactory"
    },
    "filters": [
      { "class": "solr.EdgeGramFilterFactory", "minGramSize": "4", "maxGramSize": "10", "preserveOriginal": true },
      { "class": "solr.ASCIIFoldingFilterFactory", "preserveOriginal": true },
      { "class": "solr.EnglishMinimalStemFilterFactory" },
      { "class": "solr.LowerCaseFilterFactory" },
      { "class": "solr.KStemFilterFactory" }
    ]
  },
  "queryAnalyzer": {
    "tokenizer": {
      "class": "solr.StandardTokenizerFactory"
    },
    "filters": [
      { "class": "solr.ASCIIFoldingFilterFactory", "preserveOriginal": true },
      { "class": "solr.EnglishMinimalStemFilterFactory" },
      { "class": "solr.LowerCaseFilterFactory" },
      { "class": "solr.KStemFilterFactory" }
    ]
  }
}
```

```
{
  "name": "summary",
  "class": "solr.TextField",
  "indexAnalyzer": {
    "tokenizer": {
      "class": "solr.StandardTokenizerFactory"
    },
    "filters": [
      { "class": "solr.EdgeGramFilterFactory", "minGramSize": "4", "maxGramSize": "10", "preserveOriginal": true },
      { "class": "solr.StopFilterFactory", "words": "stopwords.txt", "ignoreCase": true },
      { "class": "solr.ASCIIFoldingFilterFactory", "preserveOriginal": true },
      { "class": "solr.LowerCaseFilterFactory" }
    ]
  },
  "queryAnalyzer": {
    "tokenizer": {
      "class": "solr.StandardTokenizerFactory"
    },
    "filters": [
      { "class": "solr.EdgeGramFilterFactory", "minGramSize": "4", "maxGramSize": "10", "preserveOriginal": true },
      { "class": "solr.StopFilterFactory", "words": "stopwords.txt", "ignoreCase": true },
      { "class": "solr.ASCIIFoldingFilterFactory", "preserveOriginal": true },
      { "class": "solr.LowerCaseFilterFactory" }
    ]
  }
}
```



Home

TV Shows

Movies

News & Popular

My list

Watchlist



04

Retrieval





| Query 1: English-related shows

- **Information need:** Shows with the word 'English'
- **Relevant judgement:** The language of the show must be 'English' or the summary should contain the word 'English', but if the language of 2 shows were both English, the one who also had the word in the summary would appear first.
- **Query(q):** summary : "English" OR language : "English"
- **defType:** edismax
- **Query Fields (qf):** language^2 summary





| Query 2: Shows containing occurrence of verb Do

- **Information need:** Shows from United States released between 2017 and 2021 with 'Do' in title
- **Relevant judgement:** The origin country of the show must be 'United States' and the start year should be between 2017 and 2021, and, finally, the title must have 'do' related words ('doing', 'done', 'did', ...)
- **Query(q):** originCountry : "United States" AND startYear : [2017 TO *] and title : "Doing"
- **defType:** lucene
- **tie:** 1





| Query 3: High rating with Russian or Portuguese certificate

- **Information need:** Ended shows between 2010 and 2020 with Russian or Portuguese certificates limited by rating
- **Relevant judgement:** The rating of the show must be higher than 7 and must have been released after 2010 and ended until 2020. The certificates available for that show should contain an entry for Russia or for Portugal. The order of the results implies that shows with higher number of votes appear first
- **Query(q):** rating : [7 TO *] AND startYear : [2010 TO *] AND endYear : [* TO 2020] AND (certificate : "Rússia" OR certificate : "Portugal")
- **defType:** lucene
- **sort** -> numVotes : DESC





| Query 4: Action Series Actors

- **Information need:** Action series with actors D.B., Lesley-Ann, J.K. or Matt
- **Relevant judgement:** The show cast must have 'D.B' or 'Lesley-Ann' or 'Matt' and the genre must be 'Action'. The results should be ordered by the number of episodes, which sub-intends that the user is only interested in series.
- **Query(q):** (cast : "DB" OR cast : "Lesley Ann" OR cast : "jk" OR cast : "Matt") AND genres : "Action"
- **defType:** lucene
- **sort** -> episodes : ASC



Home

TV Shows

Movies

News & Popular

My list

Watchlist



05

Evaluation






| Evaluation

Query 1

Query 2

	Query 1	Query 2
	<ul style="list-style-type: none">- 20 shows- 15 relevant- 'Englishman' in the summary	<ul style="list-style-type: none">- 20 shows- 5 relevant- Variances of verb 'Do'

Schema-less:

	Metric	Value
1	Average Precision	1.0
2	Precision at 15 (P@15)	0.933333
3	Recall at 15 (R@15)	0.933333
4	F1 at 15 (R@15)	0.933333

Enhanced schema:

	Metric	Value
1	Average Precision	1.0
2	Precision at 15 (P@15)	1.0
3	Recall at 15 (R@15)	1.0
4	F1 at 15 (R@15)	1.0

Schema-less:

	Metric	Value
1	Average Precision	1.0
2	Precision at 5 (P@5)	0.2
3	Recall at 5 (R@5)	0.2
4	F1 at 5 (R@5)	0.2

Enhanced schema:

	Metric	Value
1	Average Precision	1.0
2	Precision at 5 (P@5)	1.0
3	Recall at 5 (R@5)	1.0
4	F1 at 5 (R@5)	1.0





| Evaluation

Query 3

Query 4



- 100 shows
- 40 relevant
- 'Rússia' certificates accentuation

- 30 shows
- 5 relevant
- User expects only series

Query 3

Schema-less:

	Metric	Value
1	Average Precision	1.0
2	Precision at 40 (P@40)	0.175
3	Recall at 40 (R@40)	0.175
4	F1 at 40 (R@40)	0.175

Enhanced schema:

	Metric	Value
1	Average Precision	1.0
2	Precision at 40 (P@40)	1.0
3	Recall at 40 (R@40)	1.0
4	F1 at 40 (R@40)	1.0

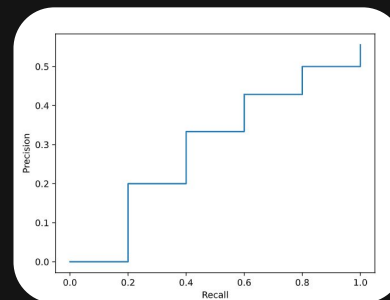
Query 4

Schema-less:

	Metric	Value
1	Average Precision	1.0
2	Precision at 10 (P@10)	0.2
3	Recall at 10 (R@10)	0.4
4	F1 at 10 (R@10)	0.266667

Enhanced schema:

	Metric	Value
1	Average Precision	0.224162
2	Precision at 10 (P@10)	0.5
3	Recall at 10 (R@10)	1.0
4	F1 at 10 (R@10)	0.666667



Home

TV Shows

Movies

News & Popular

My list

Watchlist



06

Future Work





I FUTURE WORK & IMPROVEMENTS

QUERY IDEAS

Think of new queries in order to better evaluate and test the final search system

SEARCH INTERFACE

Development of an interface to facilitate user interaction

RETRIEVAL UPGRADES

Additional schema changes in order to have a more robust and relevant search system

FINAL RESULT

Our goal is an elaborate information search and retrieval system





THANKS!

Do you have any questions?

CREDITS: This presentation template was created by Slidesgo,
including icons by Flaticon and infographics & images by Freepik

