Goodreads Books and Reviews

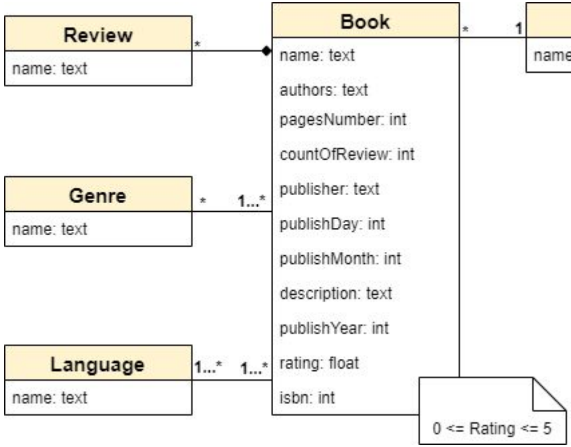
Information Processing and Retrieval
(2021-2022)

Group 22:
Ana Teresa Feliciano da Cruz
Inês Alves Quarteu
Luís Filipe Sousa Teixeira Recharte

Milestone 1: overview

Data on books and reviews from Goodreads was initially retrieved from datasets found in Kaggle and by web scraping the Goodreads site.

The collected data was refined and, finally, after the completion of the Data Preparation phase, the data was distributed by 5 CSV files: **book**, **languages**, **genres**, **reviews**, and **authors**.



Review
name: text

Genre
name: text

Language
name: text

Book
name: text
authors: text
pagesNumber: int
countOfReview: int
publisher: text
publishDay: int
publishMonth: int
description: text
publishYear: int
rating: float
isbn: int

Author
name: text

0 <= Rating <= 5

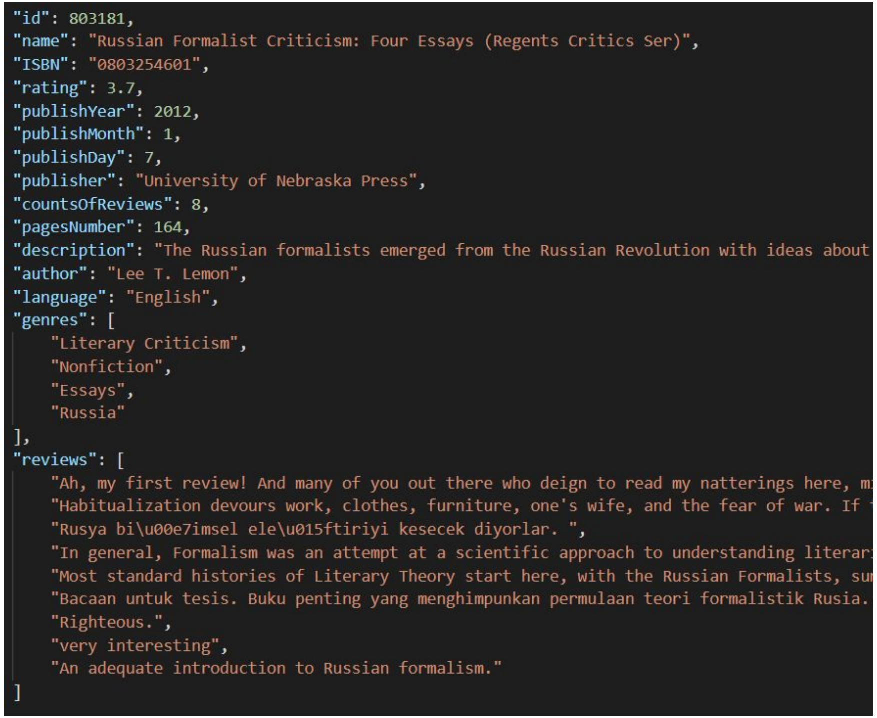
Collections and Documents

All the files were merged into a single JSON file:

- The **authors** and **languages** files were merged into the **books** file. For each of the files, a new attribute, that represents their information, was added to the books.

- To merge **reviews** and **genres** an attribute, that stores the data in an array, was added for each to the books.

The collection was imported into Solr.



```
"id": 803181,
"name": "Russian Formalist Criticism: Four Essays (Regents Critics Ser)",
"ISBN": "0803254601",
"rating": 3.7,
"publishYear": 2012,
"publishMonth": 1,
"publishDay": 7,
"publisher": "University of Nebraska Press",
"countsOfReviews": 8,
"pagesNumber": 164,
"description": "The Russian formalists emerged from the Russian Revolution with ideas about
"author": "Lee T. Lemon",
"language": "English",
"genres": [
    "Literary Criticism",
    "Nonfiction",
    "Essays",
    "Russia"
],
"reviews": [
    "Ah, my first review! And many of you out there who deign to read my natterings here, m
    "Habitualization devours work, clothes, furniture, one's wife, and the fear of war. If
    "Rusya bi\u00e7imsel ele\u015ftiriyi kesecek diyorlar. ",
    "In general, Formalism was an attempt at a scientific approach to understanding literar
    "Most standard histories of Literary Theory start here, with the Russian Formalists, su
    "Bacaan untuk tesis. Buku penting yang menghimpunkan permulaan teori formalistik Rusia.
    "Righteous.",
    "very interesting",
    "An adequate introduction to Russian formalism."
]
```

Indexing process

Field	Type	Indexed
ISBN	*string*	False
rating	*pfloat*	True
publishYear, publishMonth, publishDay, countOfReviews, pagesNumber	*pint*	True
name, publisher, description, author, language	*text*	True
reviews, genres	*text_en*	True

Configurations

Three system configurations were thought of, to be able to achieve different results while querying the documents:

- Schemaless mode;
- With the schema described in the Indexing Process;
- With the use of the weighting filter.

Information needs

1. History books about family
2. Fiction Novels
3. Books fast to read
4. Emotive books about World War II holocaust
5. Romance books about Friends for a christmas gift

Information Need 1

Information need: History books about family

Relevant judgement: The genre of the book must be ‘history’ and the description and reviews should contain the word ‘family’, but it being in the description will give a more accurate result;

Query (*q*): history family

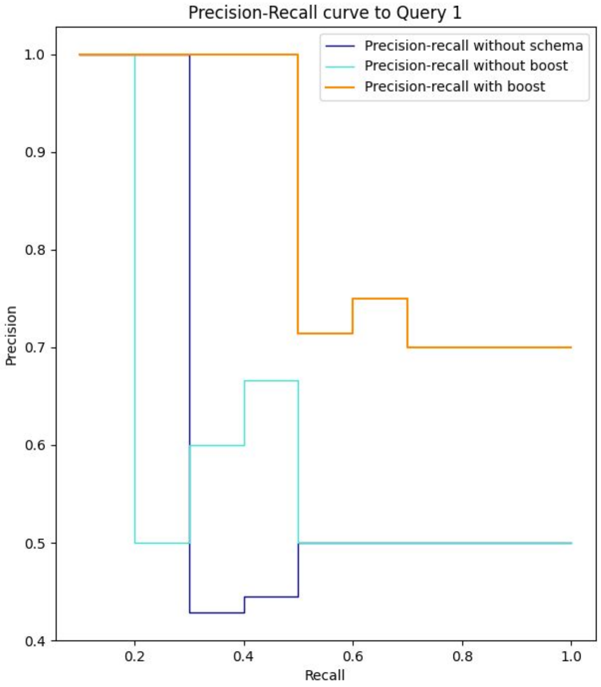
Query Fields (*qf*): genres, description, reviews

Boost Query (*bq*): (genres:history)^20 (description:family)^15

Evaluation: IN1

History books about family

	P@10	AvP	R@10
Schemaless	50%	69%	50%
Schema	50%	70%	50%
Boosted	70%	90%	70%



Information Need 4

Information need: Emotive books about World War II holocaust

Relevant judgement: The genres must match any of the words, the reviews must have ‘emotive’ related words (using anagrams with the *text_en* default type), and, finally, the description should have any of these words.

Query(*q*): “world war II” emotive holocaust

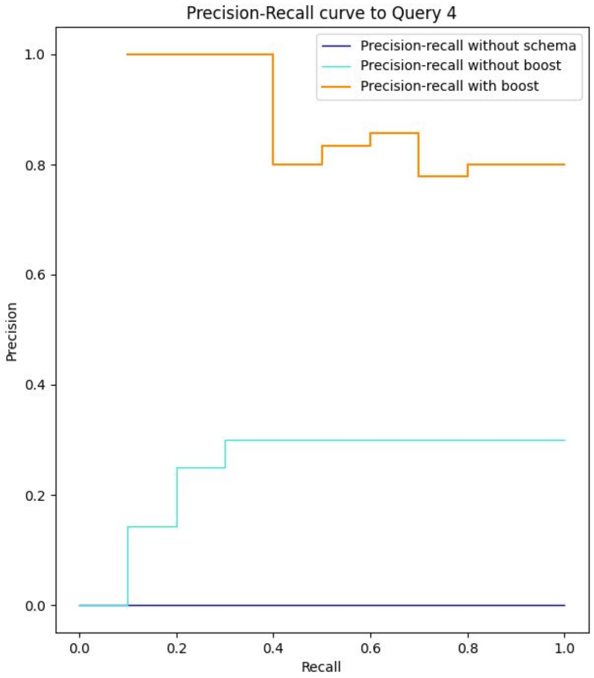
Query Fields(*qf*): genres description reviews

Boost Query(*bq*): reviews^80 genres^100 description^10

Evaluation: IN4

Emotive books about World War II holocaust

	P@10	AvP	R@10
Schemaless	0%	0%	0%
Schema	30%	10%	30%
Boosted	70%	90%	70%



Information Need 5

Information need: Romance books about Friends for a christmas gift

Relevant judgement: The genres must have ‘romance’, reviews must match the words ‘christmas’ and ‘gift’ and description should have the word ‘friends’

Without weights:

Query (*q*): christmas gift

Filter Query(*fq*): description: friends

Filter Query(*fq*): genres: romance

Query Fields (*qf*): reviews

With weights:

Query (*q*): christmas gift

Query Fields (*qf*): reviews

Phrase Fields(*pf*): reviews^30

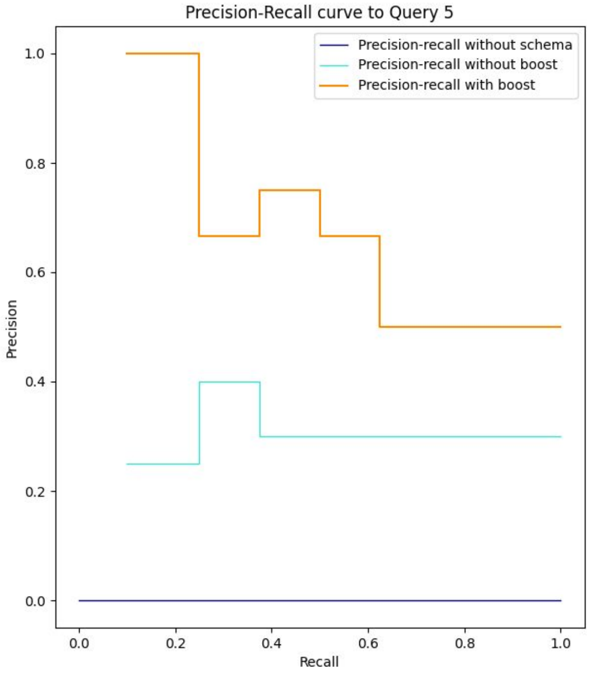
Phrase Slop(*ps*): 5

Boost Query(*bq*): (genres:romance)^60 (description:friends)^20

Evaluation: IN5

Romance books about Friends for a christmas gift

	P@10	AvP	R@10
Schemaless	0%	0%	0%
Schema	30%	46%	38%
Boosted	50%	75%	63%



Questões ?