Information Processing and Retrieval

GOODREADS' BOOKS AND REVIEWS

Diogo Almeida (up201806630)
Pedro Queirós (up201806329)

Milestone 1 Overview

The data used for this information system was extracted from Kaggle and Goodreads' website regarding books and their reviews. After the data preparation has been completed, the information is split across several csv files:

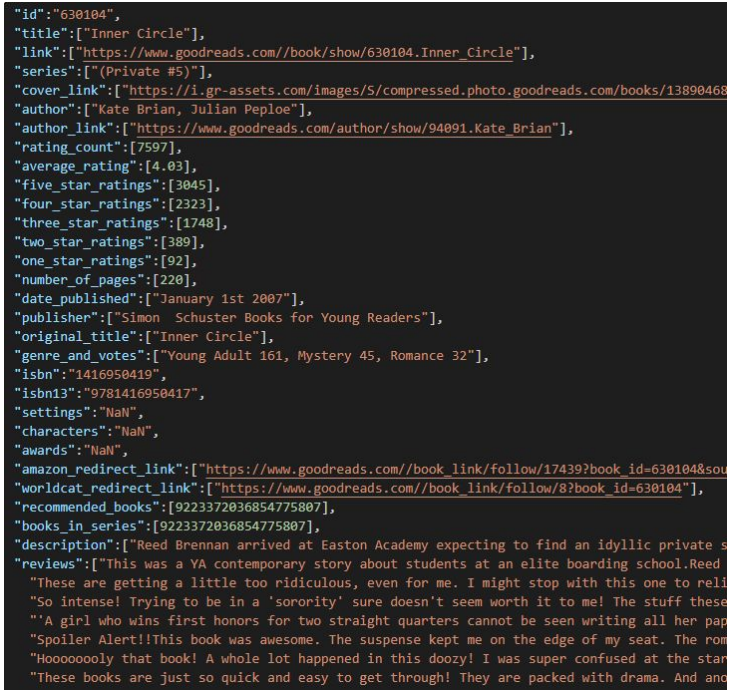
- Awards
- Characters
- Genre and votes
- Books
- Reviews

Collections and Documents

Firstly, the files regarding books, gender and votes, characters and awards were merged with the books file, each one added as an attribute of the book.

Then, the reviews were also merged in order to have for each book an array of its reviews.

The final data is gathered in a JSON file and uploaded to a previously created collection (core) in Solr



Indexing Process

The custom type standard_text was created with the following filters:

- ASCIIFoldingFilterFactory
- LowerCaseFilterFactory
- SynonymFilterFactory

Field	Type	Index
isbn, isbn13	string	No
title, series, author, date_published, publisher, original_title, genre_and_votes, settings, characters, awards, recommended_books, books_in_series, descprition, reviews	standard_text	Yes

Indexing Process

Field	Type	Index
rating_count, five_star_ratings, four_star_ratings, three_star_ratings, two_star_ratings, one_star_ratings	pint	No
average_rating	pfloat	No
link, cover_link, author_link,amazon_redirect_link, worldcat_redirect_link	string	No
isbn, isbn13	string	No
title, series, author, date_published, publisher, original_title, genre_and_votes, settings, characters, awards, recommended_books, books_in_series, descprition, reviews	standard_text	Yes

Configurations

Three system configurations were created in order to explore and obtain different results while querying the documents in the SOLR:

- Schemaless
- With the schema presented in the previous slide
- With the schema presented in the previous slide with weighting filter to boost more relevant fields

Information Need 1

Information Need: Biographies about Adolf Hitler

Relevance Judgement: “Biography” needs to be in the book’s genre, title or description. “Hitler” is also important to appear in the book’s title or author and can also appear in the description, although with less relevance.

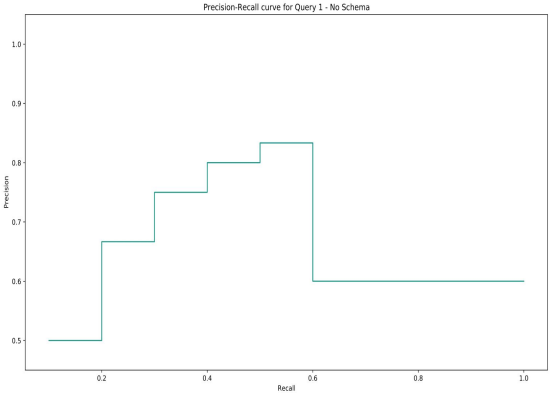
Query(q): hitler **AND** biography

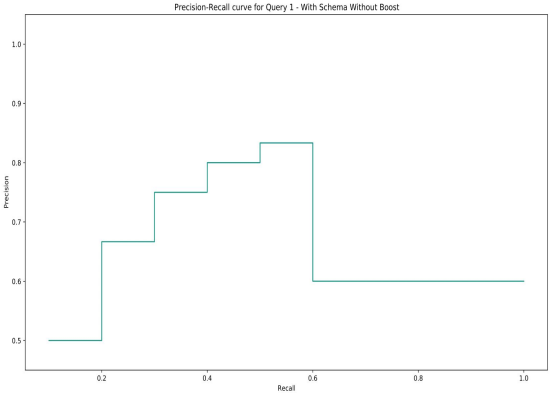
Query fields(qf): title genre_and_votes description author

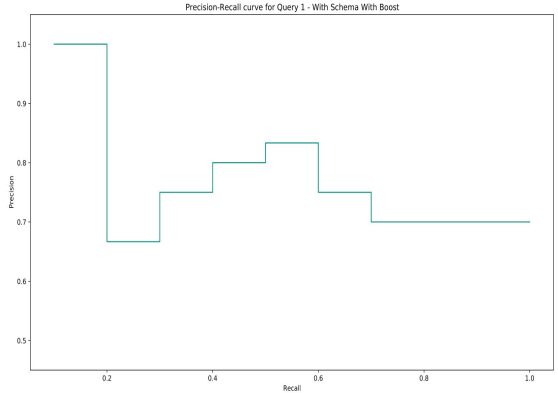
Boost Query(bq): title^3 genre_and_votes^2 description^1 author^3

Information Need 1 - Evaluation

Rank	System 1	System 2	System 3
AVP	0.781429	0.781429	0.836376
P@10	0.6	0.6	0.7







Information Need 2

Information Need: Comic books whose protagonist is Spider-Man

Relevance Judgement: The super hero name should most importantly appear in the book's title but can also appear with less relevance in the description or characters. We are also looking for comic books so this should the the book's genre.

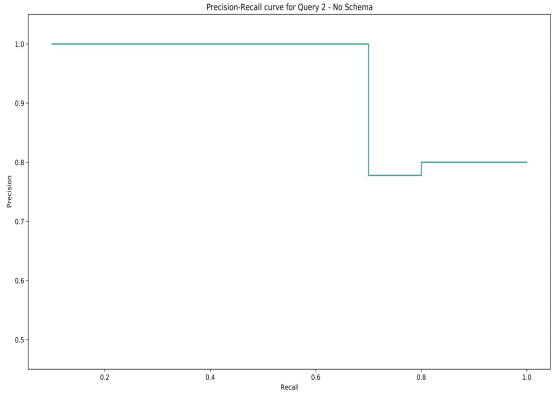
Query(q): comics **AND** ("spider-man" **OR** "Peter Parker")

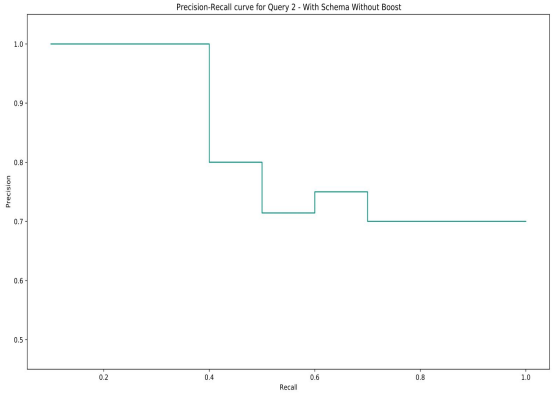
Query fields(qf): title genre_and_votes description characters

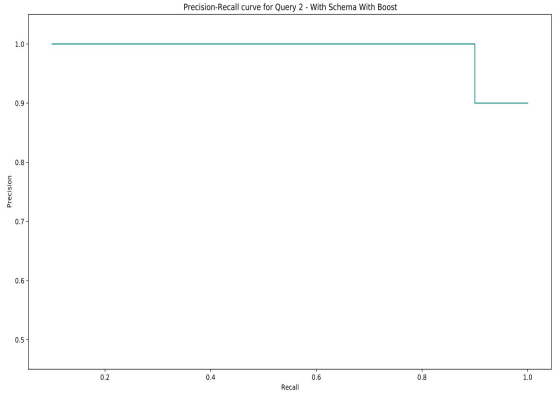
Boost Query(bq): title^3 genre_and_votes^2 description^1 characters^2

Information Need 2 - Evaluation

Rank	System 1	System 2	System 3
AVP	0.971429	0.893519	1
P@10	0.8	0.7	0.9







Information Need 3

Information Need: Non-fictional books about the life's work of Albert Einstein

Relevance Judgement: The scientist must be either the book's author or have his name in the title. He can also be part of the book's characters. His name can also appear in the description although it is less relevant. Science or Philosophy must be in the book's genre and can also appear in the description or the title.

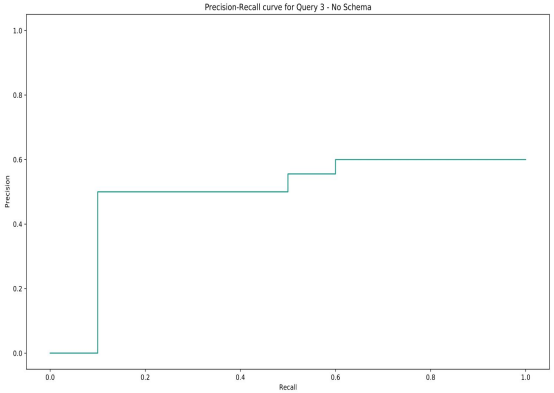
Query(q): einstein **AND -**fiction **AND** (philosophy **OR** science)

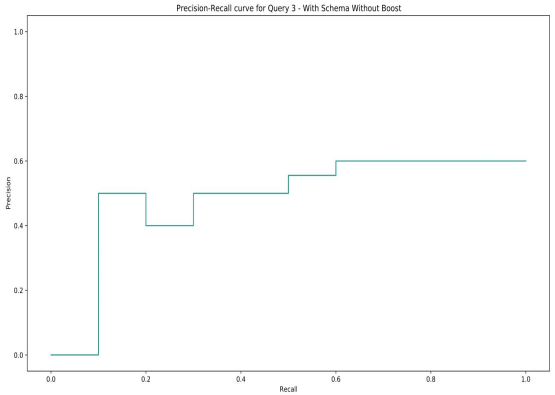
Query fields(qf): title genre_and_votes description characters

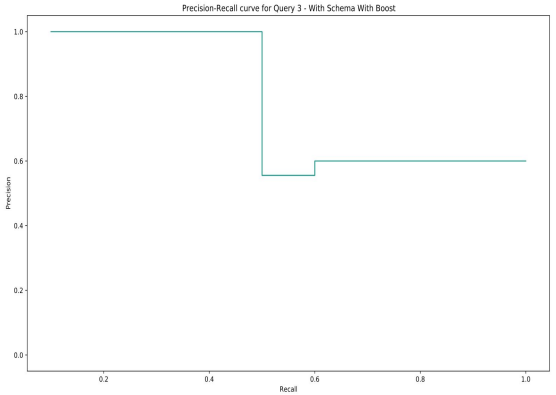
Boost Query(bq): title^3 genre_and_votes^2 description^1 characters^3

Information Need 3 - Evaluation

Rank	System 1	System 2	System 3
AVP	0.582275	0.565608	0.92
P@10	0.6	0.6	0.6







Information Need 4

Information Need: Fantasy children's books set in the medieval era that are easy to read

Relevance Judgement: We are looking for books that are easy to read, something that is not in the books' informations. So it is important to search in the reviews of each book. The genres of the books must include fantasy and children's and the story should be set in a medieval era (should mention "kingdom")

Query(q): hitler **AND** biography

Filter Query(fq): genre_and_votes:fantasy

Filter Query(fq): genre_and_votes:childrens

Filter Query(fq): description:kingdom

Query fields(qf): reviews

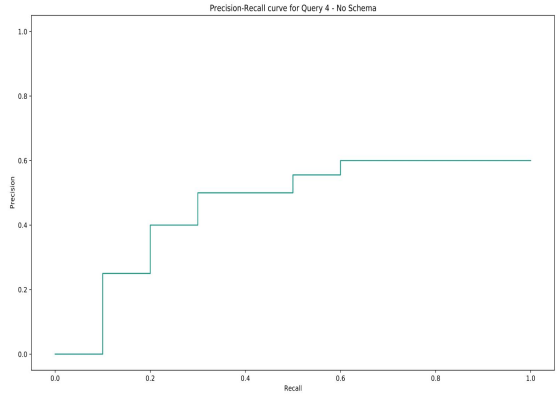
Phrase fields(pf): reviews^5

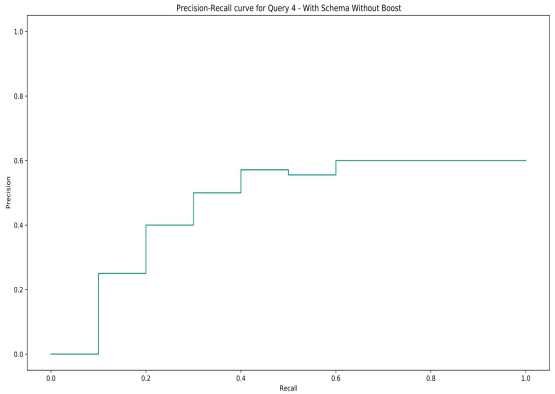
Phrase slop(ps): 5

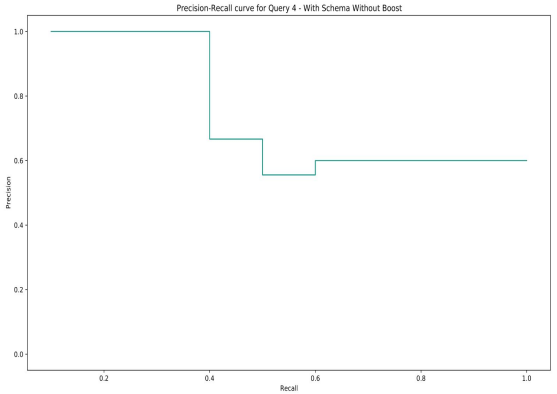
Boost Query(bq): reviews^2

Information Need 4 - Evaluation

Rank	System 1	System 2	System 3
AVP	0.493386	0.50496	0.862857
P@10	0.6	0.6	0.6







Conclusions and Future Work

For the information processing and retrieval we took the following steps:

- Merged all the csv files into one JSON file;
- Uploaded the data to the chosen tool - SOLR;
- Queried and evaluated the results for each system.

Regarding future work, some of the implemented search tasks will be:

- Improve query results by using NLP to better recognize user needs;
- Development of a graphic interface for the information system.