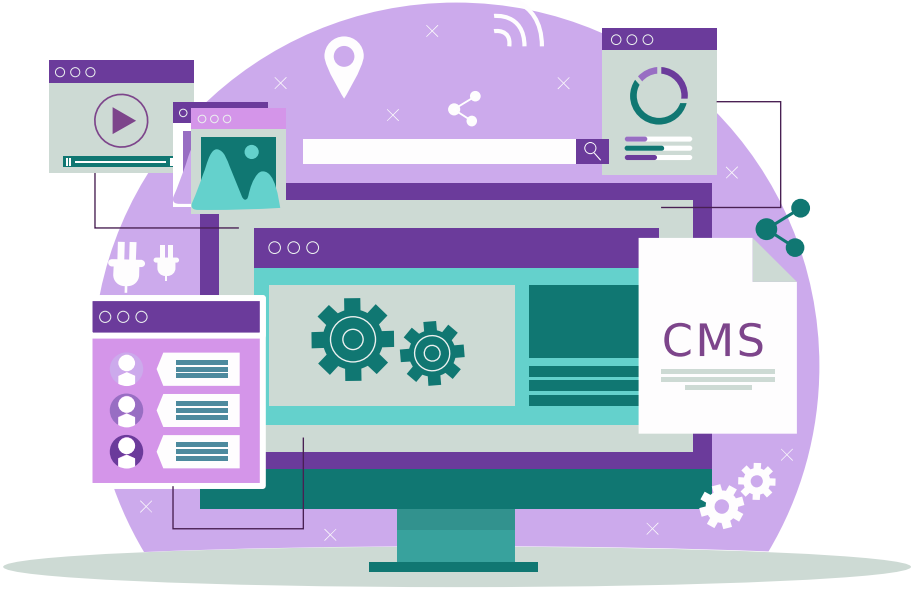
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Goodreads Books and Reviews



PRI, Group 2144
Inês Silva, up201806385
Mariana Truta, up201806543
Rita Peixoto, up 201806257

Datasets

Books

csv file

13 columns

100k
lines

from Kaggle, originally from Goodreads

Books' Reviews

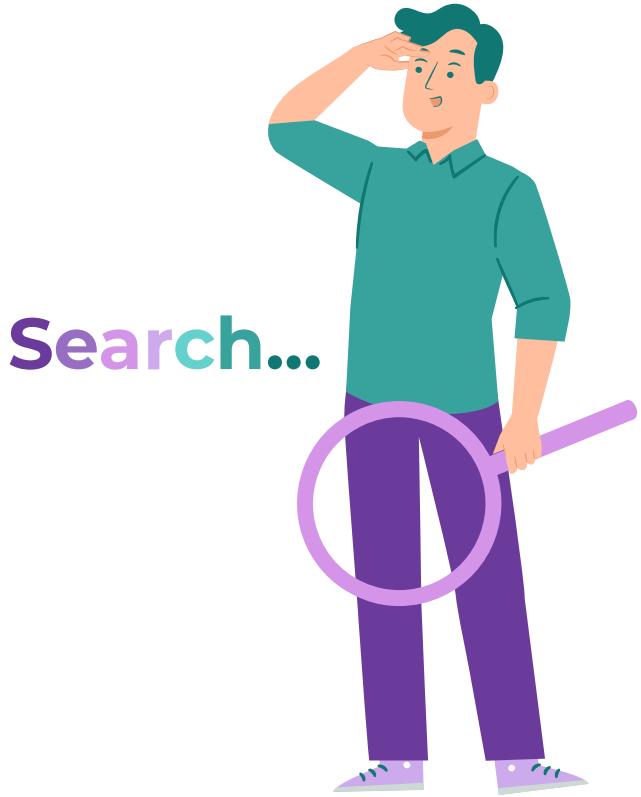
csv file

3 columns

510k
lines

scraped from Goodreads

Now with sentiment analysis ...



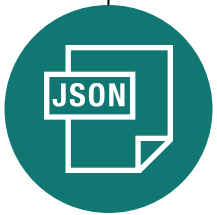
Documents

Sentiment Analysis

Using Natural Language Processing, it was added a “sentiment” to each review.

Tool Selection

Solr was the tool chosen since it is more text-oriented than ElasticSearch.



Document Structure

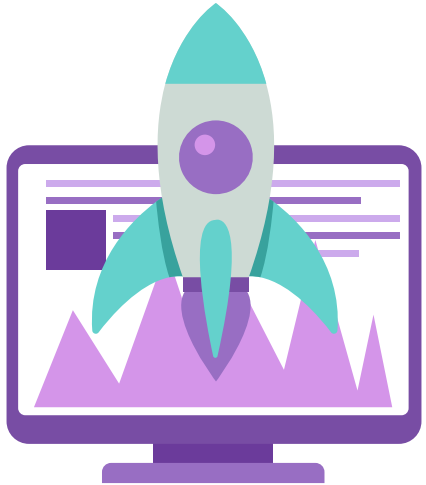
Nested documents were the first option, but those were quite limiting. So instead, two lists were created for the reviews in each book.

Information Retrieval

01 Search System

02 Query Parsers & Parameters

03 Information Retrieval 1



Information Retrieval 2 04

Information Retrieval 3 05

Information Retrieval 5 06

Information Retrieval 5 06

Indexing

commaText
author
genre

text_general
desc
positive_reviews
negative_reviews

gramText
bookformat
title

string
img
isbn
link

pint
pages
reviews
totalratings
sentiment

pfloat
rating

All fields are indexed except img and link fields.

Search Systems

System 1

- Schemaless search system

System 2

- Schema presented in previous slide
- Default weights for each field

System 3

- Schema presented in previous slide
- Defined weights for each field

Query parsers

q

defines the main query

q.op

defines the default operator (AND, in this case) for query expressions

qf

list of fields, each of which is assigned a boost factor to increase or decrease that particular field's importance in the query

fq

Jupiter is a gas giant and the biggest planet of them all

Standard Query Parser

DisMax Query Parser

Extended DisMax Query Parser

Extended DisMax

- ★ Improved Proximity;
- ★ Includes advanced stop words handling;
- ★ Allows the specification of the fields the user is allowed to query;
- ★ Disallows the direct search on the fields and supports the specification of fields' weight.

Information retrieval 1

Information Retrieval

One intends to find a great cooking book to offer their mom, who's vegan and doesn't have a lot of cooking skills

Query

easy and delicious vegan recipes

Relevance judgement

The intention was to retrieve books where the reviews mentioned easy and delicious, with the mention of vegan in the gender and/or in the description.

1.5 genre default desc 2.0 positive reviews

Information retrieval 2

Information Retrieval

One is looking for an interesting fiction book or a romance, with a good plot that will surely get them hooked on the story.

Query

interesting AND (fiction OR romance)

Relevance judgement

The intention was to retrieve books whose positive reviews mentioned the fact that the book was interesting and the gender had either fiction or romance

1.8 genre

default positive reviews

Information retrieval 3

Information Retrieval

One is feeling lonely and nostalgic about their childhood household and is looking for a book that talks about family.

Query

family

Relevance judgement

The intention was to retrieve books about families, giving priority to the book who have in the gender and/or in the description family and that actually involves families.

2.5 genre

default title

1.5 desc

Information retrieval 4

Information Retrieval

After a break-up, one feels the need to believe in love again, so books with good clichés of adult romance are always a comfort in the cold days of the holiday season.

Query

good cliche adult romance

Relevance judgement

The intention was to retrieve books whose gender has romance, and that the positive reviews said it was a good cliché.

0.8 genre

0.01 negative_reviews

1.2 positive_reviews

Information retrieval 5

Information Retrieval

As a typical history lover dad, my dad is reading a book about the Nazi holocaust. I don't know the book's name, I only remember its cover, and I want to tell my sister about it.

Query

nazi holocaust history

Relevance judgement

The intention was to retrieve books whose gender contained history and holocaust and whose description and/or title refer to the word Nazi.

1.8 genre

default title

1.2 desc

Evaluation

Query 1			
Rank	S1	S2	S3
1	Y	Y	Y
2	Y	Y	Y
3	Y	Y	Y
4	Y	Y	Y
5	Y	Y	Y
6	N	Y	Y
7	Y	Y	Y
8	Y	Y	Y
9	Y	Y	Y
10	N	Y	Y

Query 2			
Rank	S1	S2	S3
1	N	Y	Y
2	N	Y	Y
3	Y	Y	Y
4	N	Y	Y
5	-	Y	N
6	-	N	Y
7	-	Y	N
8	-	N	Y
9	-	N	Y
10	-	N	N

Query 3			
Rank	S1	S2	S3
1	Y	N	Y
2	N	N	Y
3	-	N	N
4	-	N	N
5	-	Y	N
6	-	Y	Y
7	-	Y	Y
8	-	Y	N
9	-	Y	Y
10	-	N	Y

Query 4			
Rank	S1	S2	S3
1	N	N	Y
2	Y	N	Y
3	N	Y	Y
4	N	Y	Y
5	N	N	Y
6	N	N	Y
7	N	Y	Y
8	N	Y	Y
9	N	Y	Y
10	N	N	N

Query 5			
Rank	S1	S2	S3
1	Y	Y	Y
2	N	Y	Y
3	N	N	Y
4	N	Y	Y
5	Y	Y	N
6	N	N	Y
7	N	Y	Y
8	N	N	N
9	Y	N	N
10	N	N	N

Evaluation

	Query 1			Query 2			Query 3			Query 4			Query 5		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Avg Precision	0.9208	1.0	1.0	0.3333	0.8078	0.8386	1.0	0.4645	0.6255	0.2525	0.4941	0.9341	0.4917	0.6994	0.8130
P@10	0.8	1.0	1.0	0.1	0.6	0.7	0.1	0.5	0.5	0.1	0.5	0.8	0.3	0.5	0.6

System 1

Mean Average Precision **0.5997**

System 2

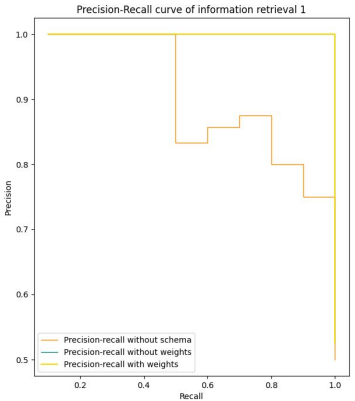
Mean Average Precision **0.6773**

System 3

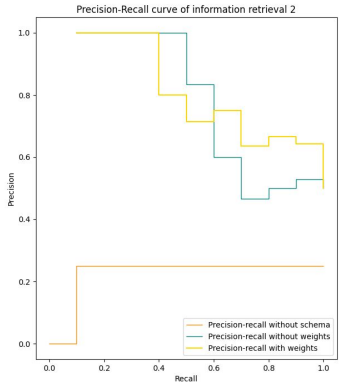
Mean Average Precision **0.8423**

Evaluation

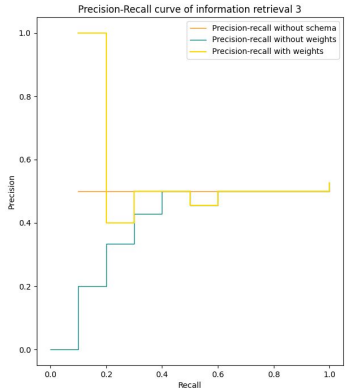
Query 1



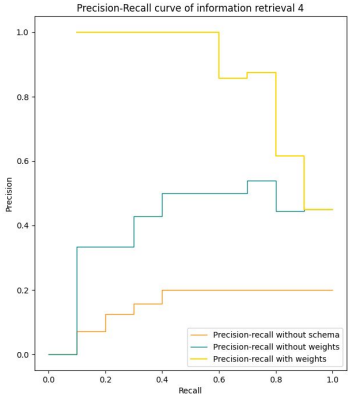
Query 2



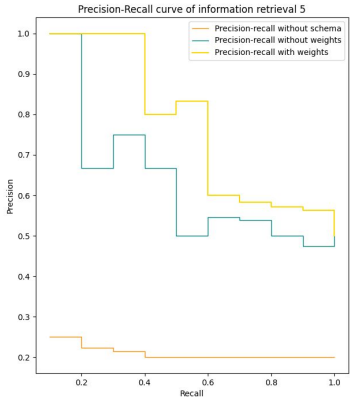
Query 3



Query 4



Query 5



Conclusion

- In a second stage of the process was developed an indexing process, sustained in the exploration and analysis of different Filters and Tokenizers provided by Solr. Several purposeful information needs were carefully conceived and used to evaluate and compare the developed retrieval systems.

- The results obtained prove the initial belief that the weights and schema would have a big impact in the quality of the search system.

- It points out that the combination of schema with weighted fields brings better results, but still has room for improvement, as the mean average precision is still approximately 84%.

Future work

Improve sentiment analysis

Improve quality of search results

Create final version of the search system