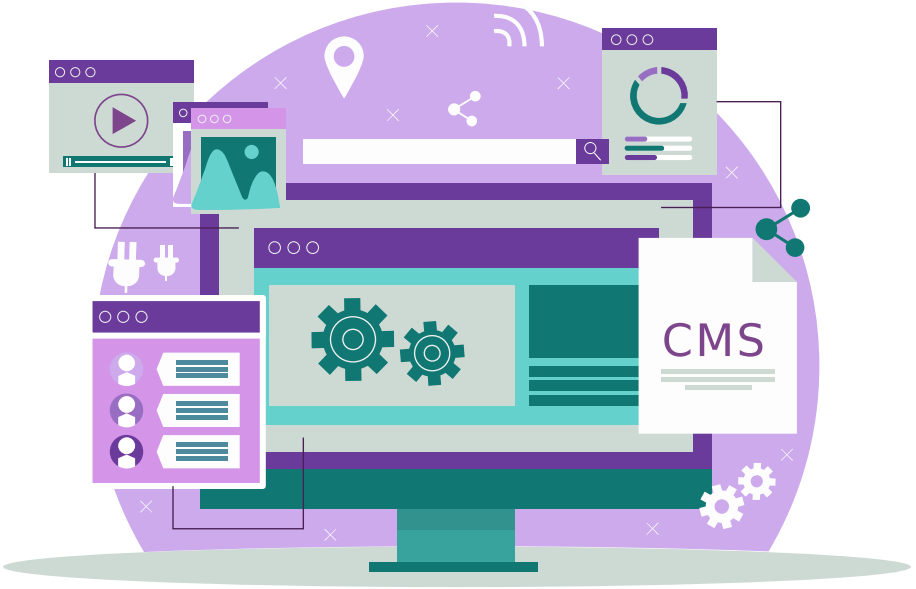
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

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

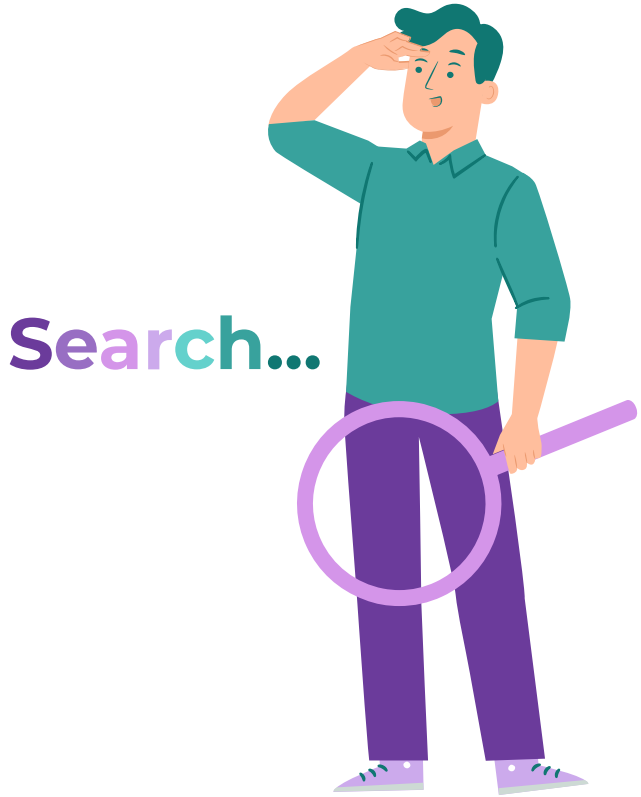
Books' Reviews

csv file

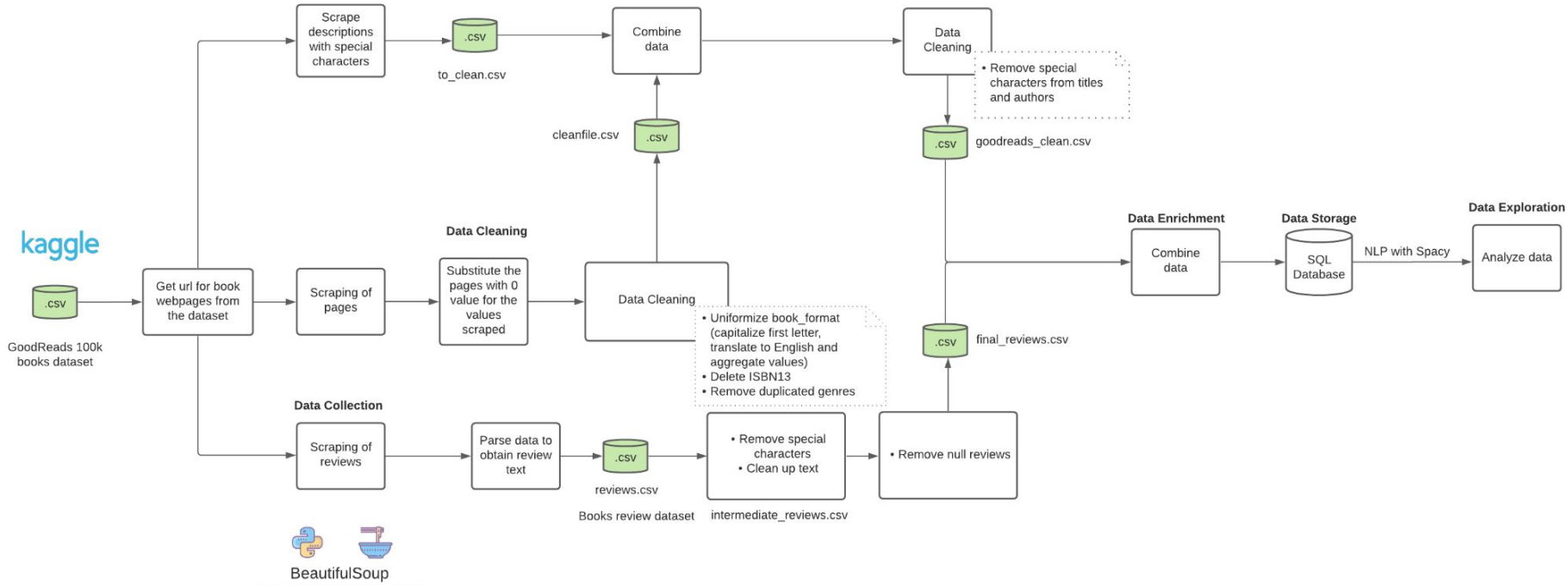
2 columns

510k
lines

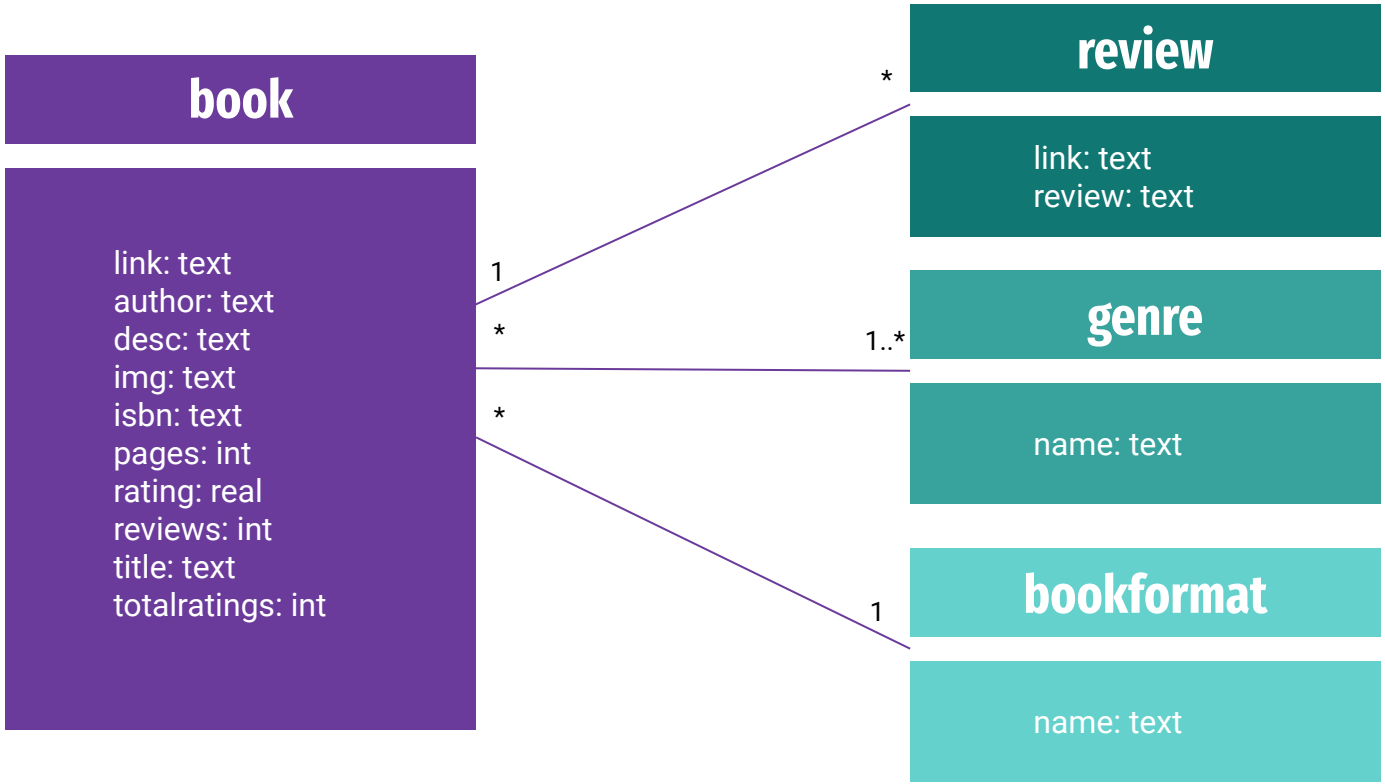
scraped from Goodreads



Pipeline


kaggle
.csv
GoodReads 100k books dataset
Get url for book webpages from the dataset
Scrape descriptions with special characters
.csv
to_clean.csv
Combine data
Data Cleaning
• Remove special characters from titles and authors
.csv
goodreads_clean.csv
cleanfile.csv
.csv
Data Cleaning
Scraping of pages
Substitute the pages with 0 value for the values scraped
Data Cleaning
• Uniformize book_format (capitalize first letter, translate to English and aggregate values)
• Delete ISBN13
• Remove duplicated genres
.csv
final_reviews.csv
Data Collection
Scraping of reviews
Parse data to obtain review text
.csv
reviews.csv
Books review dataset
• Remove special characters
• Clean up text
intermediate_reviews.csv
• Remove null reviews
BeautifulSoup
Data Enrichment
Combine data
Data Storage
SQL Database
NLP with Spacy
Data Exploration
Analyze data

Domain Conceptual Model



Data Characterization

Book Format
Paperback ≈ 50%

Book Format
Hardcover ≈ 25%

Rating
Average 3.83

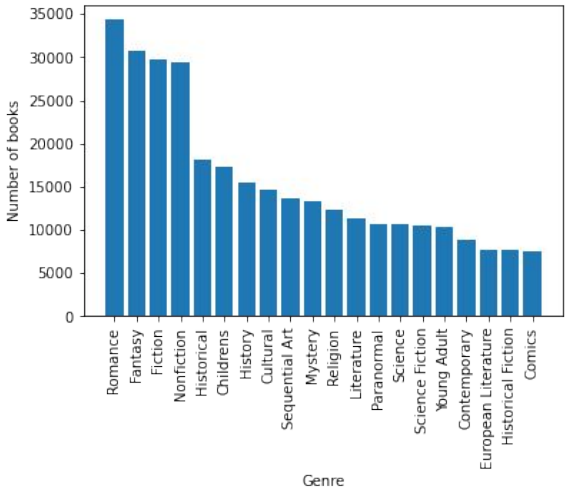
Review
Average number: 181

Review
Average length (characters): 1050

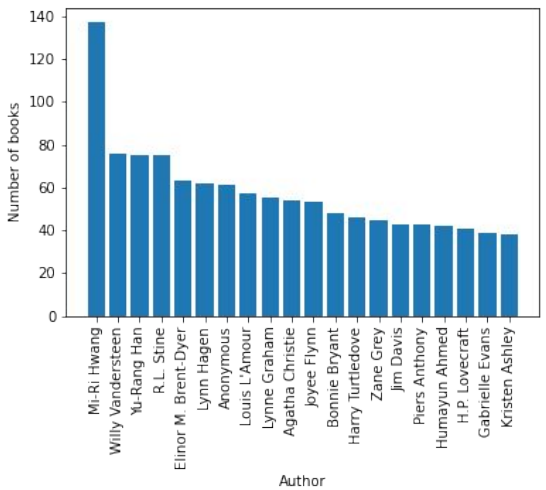
Pages
Average number: 276

Data Characterization

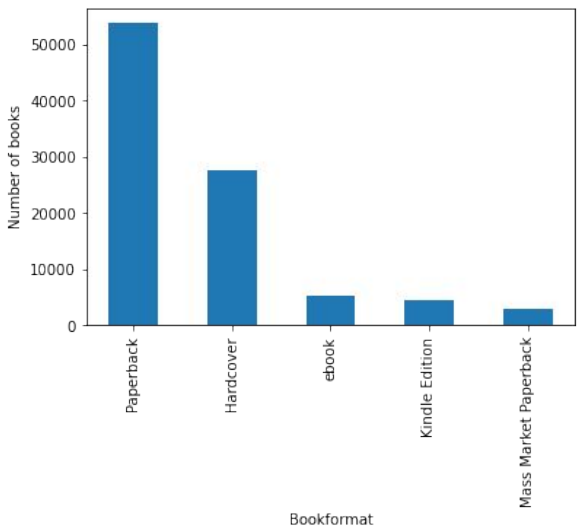
Top 20 genres



Top 20 authors

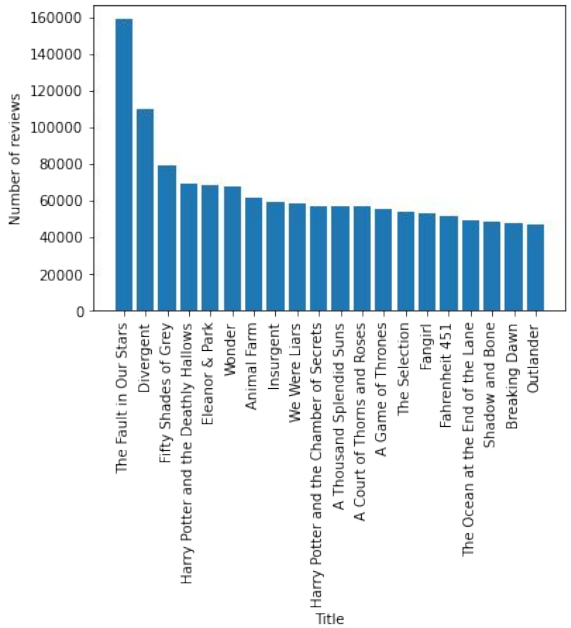


Top 5 book formats

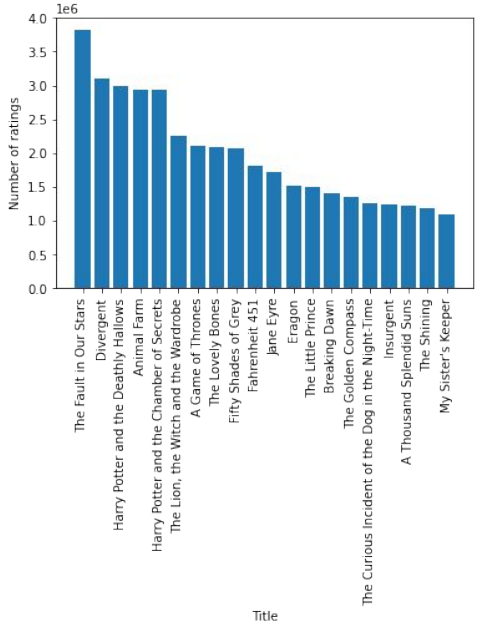


Data Characterization

Top 20 books with most reviews



Top 20 books with most ratings



Book dataset

author	Book format	desc	genre	img	isbn	isbn13	link	pages	rating	reviews	title	Total ratings
Charlotte Fiell,Emmanuelle Dirix	Paperback	Reveals that several hundred thousand Indians ...	Couture,Fashion,Historical,Art,Nonfiction	https://i.gr-assets.com/images/S/compressed.ph...	1906863482	9.78E+12	https://goodreads.com/book/show/10010552-fashi...	576	4.51	6	Fashion Sourcebook 1920s	41
Week 02	Paperback	Fashion Sourcebook - 1920s is the first book i...	Politics,History	https://i.gr-assets.com/images/S/compressed.ph...	948984147	9.78E+12	https://goodreads.com/book/show/1001077.Hungar...	124	4.15	2	Hungary 56	26
...	...	...	...	...	...	...	...	...	...	...	...	...

Reviews Dataset

id	review
0	This collection of essays is from 1994 and the ...
1	When I read this book, I did not know what to ...
2	This was good. Very extensive. Helped in plann ...
3	No atendió mi expectativa, puede que se deba p ...
...	...

Future work

Review Milestone 1

Revise if all the developed work is ready for the following stages

Building indexes

Analyse the documents and identify their indexable components and build the indexes

Retrieve Information

Demonstrate the indexes and retrieval process

Create search system

Integration of the datasets in a search system