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

Datasets

Books' dataset:

- retrieved from the a subset of the existing books in the website; it was fetched from kaggle.
- initially 30 CSV files, 23 with books' information and the remaining 7 with reviews of some of the books; had an average of 10000 entries and an average of 20 columns each.
- final dataset with 100269 entries and 12 columns.

Reviews' dataset:

- built by scraping the Goodreads website with the books' IDs from the dataset **books.csv**.
- final dataset has 559874 entries and 2 columns.

Genres' dataset:

- built by scraping the Goodreads website with the books' IDs from the dataset **books.csv**.
- final dataset has 299782 entries and 2 columns.

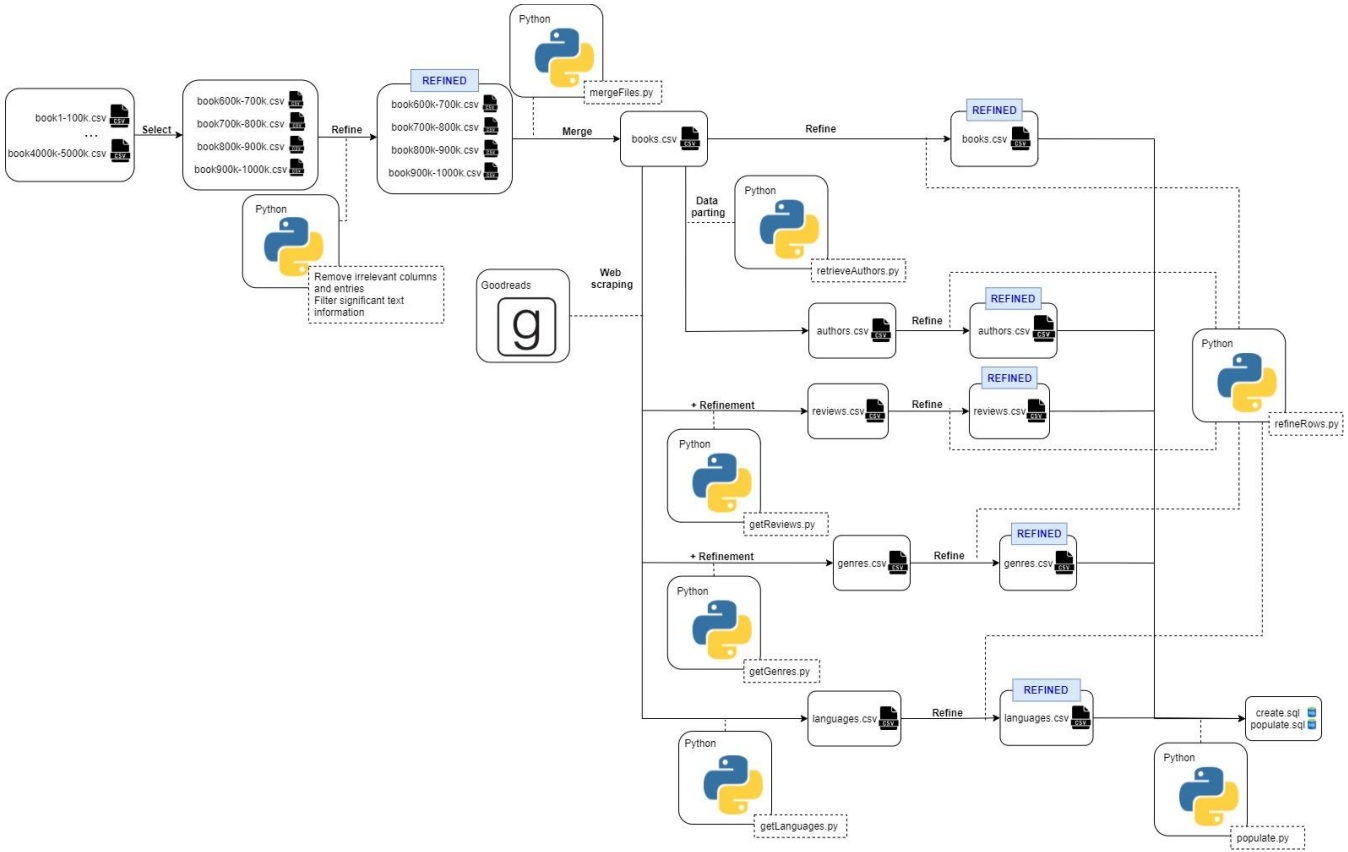
Languages' dataset:

- built by scraping the Goodreads website with the books' IDs from the dataset **books.csv**.
- final dataset has 100269 entries and 2 columns.

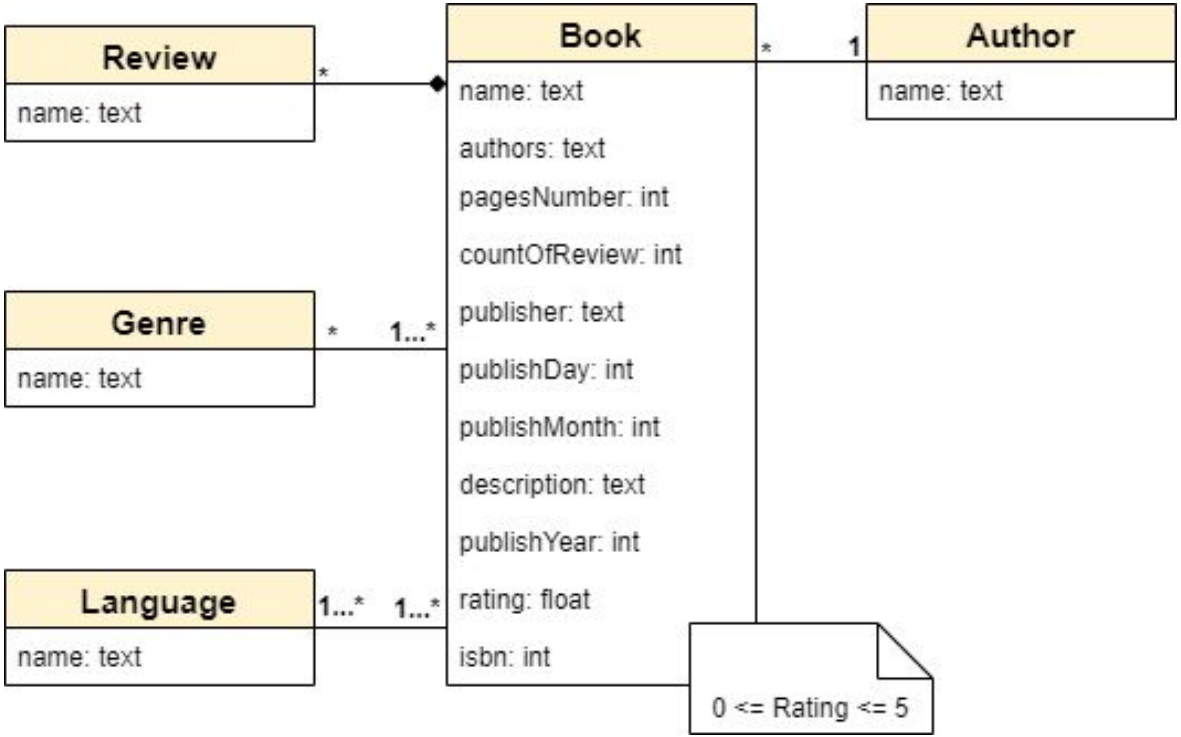
Authors' dataset:

- built by removing the authors' column from the dataset **books.csv**
- final dataset has 100269 entries and 2 columns.

Pipeline


book1-100k.csv
book4000k-5000k.csv
Select
book600k-700k.csv
book700k-800k.csv
book800k-900k.csv
book900k-1000k.csv
Refine
Python
Remove irrelevant columns and entries
Filter significant text information
REFINED
book600k-700k.csv
book700k-800k.csv
book800k-900k.csv
book900k-1000k.csv
Python
mergeFiles.py
Merge
books.csv
Refine
REFINED
books.csv
Data parting
Python
retrieveAuthors.py
Goodreads
Web scraping
authors.csv
Refine
REFINED
authors.csv
+ Refinement
reviews.csv
Refine
REFINED
reviews.csv
Python
getReviews.py
+ Refinement
genres.csv
Refine
REFINED
genres.csv
Python
getGenres.py
languages.csv
Refine
REFINED
languages.csv
Python
getLanguages.py
Python
refineRows.py
create.sql
populate.sql
Python
populate.py

Conceptual Model



Data Characterization: Books

Id	Id given by Goodreads for each book
Name	Name of the book; Every book has a name
ISBN	International Standard Book Number
Rating	Rating of the book given by the users, between 0 and 5
Publish Year, Month and Day	These attributes represent the publishing date of the book
Publisher	Which publisher published the book
Count of Reviews	Number that represents the reviews given to a book
Pages Number	The number of pages of a book
Description	The plot of the book; Every book has a description

Data Characterization: Others

All the remaining files have 2 columns and each of them has the attribute ID, defined for all as the **id given by Goodreads to a book**.

The remainder of attributes for each of the files are in the following tables.

Review	A text user review of a book; Each book may have one, multiple reviews, or none
Genre	Genre of the book; Each book may have more than one genre
Language	Language the book was written in; A book can have been written in more than one language
Author	Author of the book; Each book can only have one author, but more than one book can be written by the same author

Search Tasks

In order to extract information from the data, the following are the examples of some possible queries:

- Search books by genre
- Search books by a given text
- Search the top 5 books
- Search the reviews of the books whose rating are above 4
- Search the books published during the summer of the year 2004
- Search reviews by a given text
- Search for the books of a given author