Goodreads Books and Reviews

Ana Cruz[1], Ines Quarteu[2], and Filipe Recharte[3]

[1]FEUP, up201806460@up.pt
[2]FEUP, up201806279@up.pt
[3]FEUP, up201806743@up.pt

November 15, 2021

Abstract

Given the growing amount of available data and the possibility of processing a large amount of information, indexing and searching effectively is a growing concern in today's information systems. In the present paper, the object of study was the Books and Reviews from Goodreads, as well as additional data on Reviews, Genres, and Languages gathered by web scraping the site. After a process of refinement, where the data was cleaned and normalized, the dataset was characterized, which made clear that its study may be an interesting exercise.

Keywords Books, Dataset refining, Data retrieval, Data processing pipelines, Python

1 Introduction

In this report, the process of characterizing, processing, and, finally, querying the data regarding the books and their reviews from Goodreads will be presented. Initially, the datasets being used are detailed, as the origin and collection method, as well as the refinement process of each of the collections are explained thoroughly. Additionally, the Domain Conceptual Model is presented, as are some possible queries. Finally, a reflection on this paper's conclusions is featured in order to clarify the purpose and the results of this process.

2 The Datasets and its Preparations

The retrieval of the data was accomplished by scripts in Python using the Pandas tool [1]. For the refinement, it was used Python and the CSV library [2].

2.1 Books

1. *Origin and Collection Method:* The Books dataset was retrieved from the **Goodreads-Book-Datasets-With-User-Rating-2M**. This dataset is a subset of the existing books in the Goodreads [3] website and was fetched from kaggle [4].

2. *Description:* This dataset was composed of 30 files, 23 with books and some of its attributes, ISBN and author for example, and the other 7 with reviews of some of these books. The files were in CSV format and had an average of 10000 entries and an average of 20 columns each.

3. *Refinement:* Since the files of the reviews did not contain actual reviews but were instead random sentences, these files were discarded. Some of the attributes files were incomplete, therefore were also discarded. In the end, 4 files were chosen to be more refined. These files were chosen because they appeared to be the most complete ones. The columns *RatingDist5*, *RatingDist4*, *RatingDist3*, *RatingDist2*, *RatingDist1*, *RatingDistTotal* were eliminated, since there was no use to them; *language*, since most entries were null; and *CountOfTextReviews*, since it was not relevant. Every entry that had the *Description* null was eliminated, as for the remaining descriptions, the HTML tags and the new lines were removed. After this, the 4 files were merged to compose a single CSV file. After doing the scraping that is described in the following subsections, the final refinement consisted of removing the book's entries that had no genre or language. The resulting dataset was **books.csv**, with 100269 entries and 12 columns.

2.2 Reviews

1. *Origin and Collection Method:* The reviews dataset was built by scraping the Goodreads website with the book's Ids from the dataset **books.csv**.

2. *Description:* This dataset has 559874 entries and 2 columns.

3. *Refinement:* Some entries contained Arabic characters, which were removed. The resulting dataset was **reviews.csv**.

2.3 Genres

1. *Origin and Collection Method:* The genres dataset was built by scraping the Goodreads website with the book's Ids from the dataset **books.csv**.

2. *Description:* This dataset has 299782 entries and 2 columns.

3. *Refinement:* Since Goodreads does not have actual genres but instead shelves created by users, some filters were applied to this dataset. Entries with numbers and starting with down-case letters were removed. The resulting dataset was **genres.csv**.
   After removing the books that had no language associated, we used **books.csv** to remove rows from this dataset that were not used anymore.

2.4 Language

1. *Origin and Collection Method:* The language dataset was built by scraping the Goodreads website with the book's Ids from the dataset **books.csv**.

2. *Description:* This dataset has 100269 entries and 2 columns.

3. *Refinement:* For this dataset, it wasn't needed refinement. The resulting dataset was **languages.csv**.
   After removing the books that had no genre associated, we used **books.csv** to remove rows from this dataset that were not used anymore.

2.5 Authors

1. *Origin and Collection Method:* The authors' dataset was built by removing the authors' column from the dataset **books.csv**.

2. *Description:* This dataset has 100269 entries and 2 columns.

3. *Refinement:* For this dataset, it wasn't needed refinement. The resulting dataset was **authors.csv**.

3 The Data

After the selection and the refinement of the datasets are completed, there are 5 datasets. The conclusions of these datasets were achieved with Python scripts and the Pandas tool.

3.1 Books

The main dataset is the book's dataset which contains 12 columns:

- *Id:* this id is the Goodreads id for the book
- *Name:* this is the name of the book; every book has a name
- *ISBN:* this stands for International Standard Book Number which is an id for the book
- *Rating:* this is the rating of the book given by the users, this rating is a float number between 0 and 5
- *Publish Year, Publish Month and Publish Day:* these attributes represent the publishing date of the book
- *Publisher:* which publisher published the book
- *Count of Reviews:* a number consisting in how many reviews were given to the book
- *Pages Number:* the number of pages of the book
- *Description:* the plot of the book; every book has a description

Conclusions

After analysis, it was concluded that about 63.6% of books have a rating greater than 3 and less than 4, 34.2% are rated between 4 and 5, and a small portion is rated less than 3.
It was also concluded that the vast majority of books were published between 2000 and 2007, and the dataset has much less data on publications before 1980 and after 2010.
Finally, regarding the length of the description, it was concluded that about 47% of the descriptions have a maximum of 100 words, 37% have between 101 to 200 words and only 16% of the descriptions have more than 200 words.
(further information may be found in charts 3, 4, 5 and 6)

3.2 Reviews

- *Id:* this id is the Goodreads id for that book
- *Review:* a text user review of the book; books can have one review, multiple reviews or none

Conclusions

After analysis, it was concluded that only about 18% of the books did not have any reviews. The maximum number of reviews per book is 30, and only 4 books have 30 reviews. The majority of the books have a maximum of 10 reviews, and only a very small part has more than 10.
Besides that it was also concluded that about 69% of the reviews have a maximum of 100 words, 17% between 100 and 200, and only about 14% have more than 200 words.
(further information may be found in charts 7 and 8)

3.3 Genres

- *Id:* this id is the Goodreads id for that book
- *Genre:* the genre of the book; for each book there can be more than one genre

Conclusions

After analysis, it was concluded that are an enormous amount of different genres. The most common one is NonFiction followed by Fiction, History, Childrens, and Classics, etc.
(further information may be found in chart 9)

3.4 Languages

- *Id:* this id is the Goodreads id for that book
- *Language:* the language the book is written; books can have more than one language

Conclusions

After analysis, it was concluded that about 97% of the books are written in English. The other 3% are divided by other languages such as Spanish, French, etc.
(further information may be found in chart 10)

3.5 Authors

The authors dataset contains 2 columns:

- *Id:* this id is the Goodreads id for the book
- *Author:* this is the author of the book; for each book, there is only one author, but there is more than one book that was written by the same author

Conclusions

After analysis, it was concluded that the author that wrote more books is William Shakespeare, who wrote 180 books. Following him is Agatha Christie with 165 books, and so on.
(further information may be found in chart 11)

4 Data processing Pipeline

To process the data, this pipeline was used. The process starts by downloading the **Goodreads-Book-Datasets-With-User-Rating-2M** dataset and selecting 4 CSV files. With Python scripts the refinement such as removing irrelevant columns and entries, and separating the data was possible. In addiction to that, merging these files at the end.
By the use of other Python scripts, by web scraping, other information was obtained, such as languages, reviews, and genres.
After that, all files were refined again.
Last but not least, from the data of the CSV files, the SQL files were built.

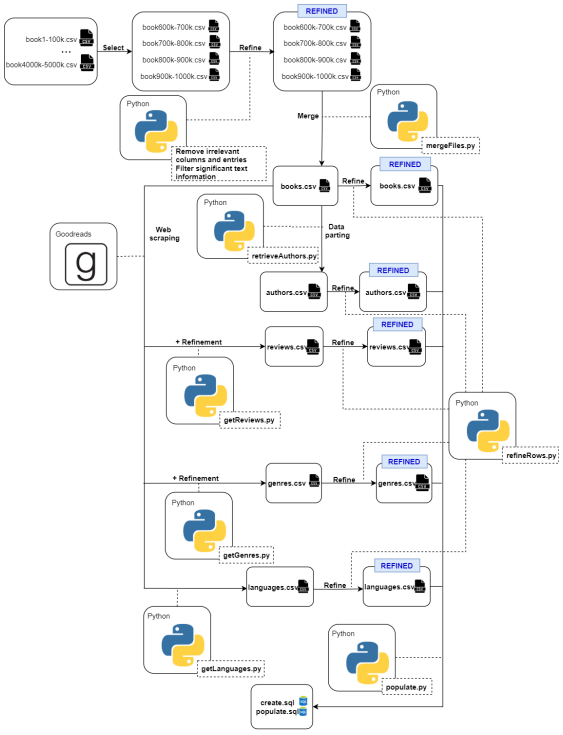

Figure 1: Data processing Pipeline

5 Domain Conceptual Model

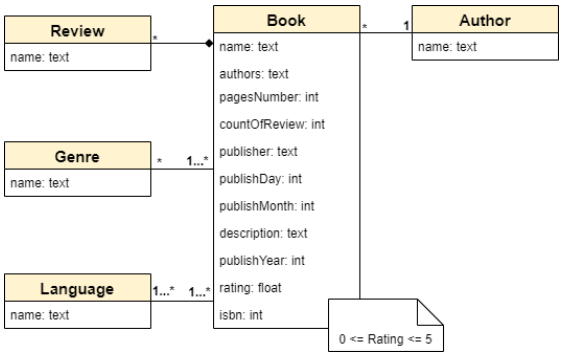

Figure 2: Domain Conceptual Diagram

6 Possible Queries

To extract information from the data, the following are examples of some possible queries.

1. Search books by genre
2. Search books from a given text
3. Search the top 5 books
4. Search the reviews of the books which rating is above 4
5. Search the books published during the summer of the year 2004
6. Search reviews by a given text
7. Search for the books of a given author

For the search tasks, from the user query are used some keywords to perform the actual query. Exemplifying:

User input: Books about climbing, love, and baking.

Performed query: Search the book's description, genre, and reviews for books that contain these words. Different fields may have different weights and it is prioritized to find as many words as possible in common with the user input.

Conclusions and Future Work

In this report, the chosen datasets and the process of rectification implemented to them are described. Initially, the datasets were obtained through actions variety of and refined. The refinements included the removal of irrelevant columns and entries, as well as the filtering of significant text information. Afterward, the information available in the datasets was analyzed, and, with that in mind, relevant charts were traced.

Additionally, the Pipeline used to process the data is identified, as is the Domain Conceptual Model which describes The Domain Conceptual model describes the relations between the different entities.

As future work, a system able to search information within these datasets will be developed.

References

[1] `https://pandas.pydata.org/`

[2] `https://docs.python.org/3/library/csv.html`

[3] `https://www.goodreads.com/`

[4] `https://www.kaggle.com/bahramjannesarr/goodreads-book-datasets-10m`

7 Annex I - Dataset Analysis

Books Dataset Analysis

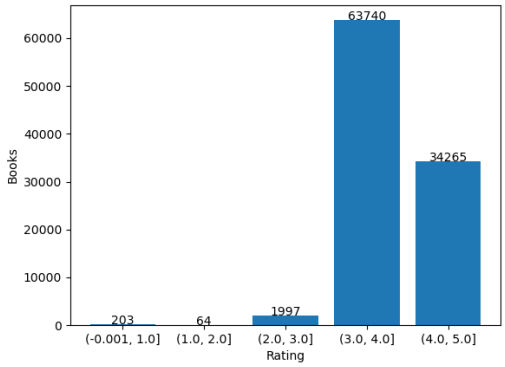

Figure 3: Books per Rating

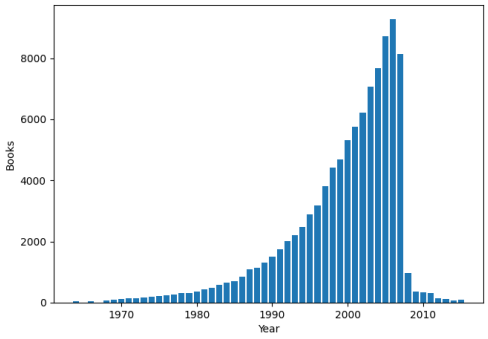

Figure 4: Books per Year

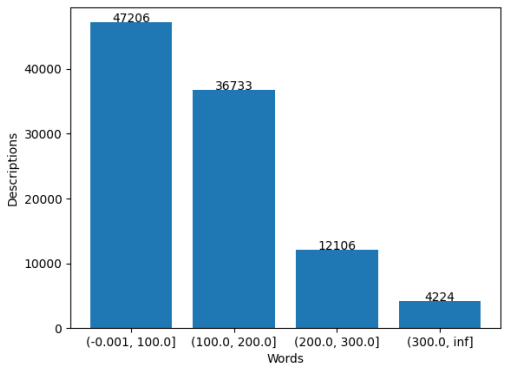

Figure 5: Descriptions per Words

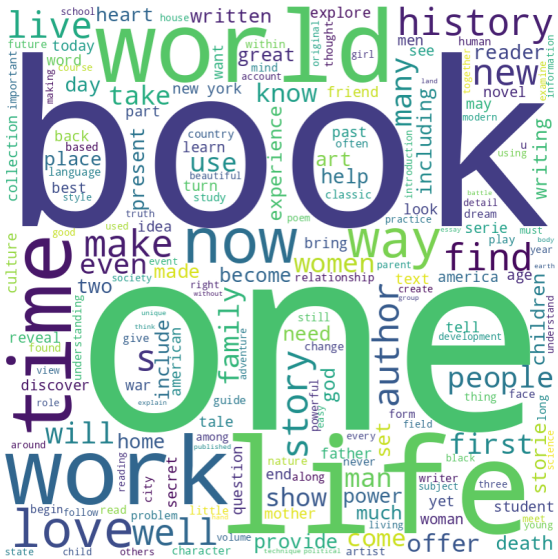

Figure 6: Most common words in all descriptions.

Reviews Dataset Analysis

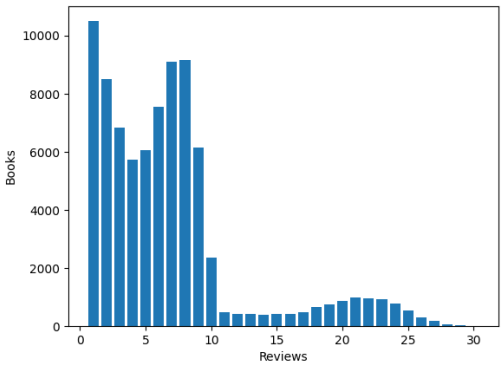

Figure 7: Books per Reviews

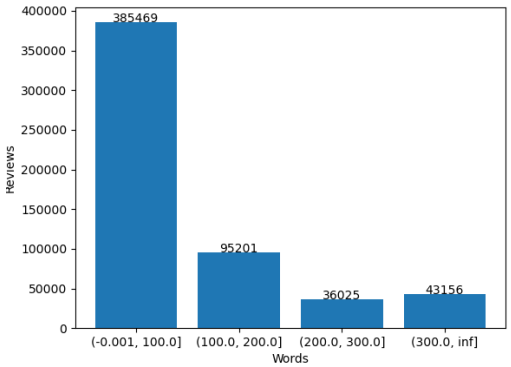

Figure 8: Reviews per Words

Genres Dataset Analysis

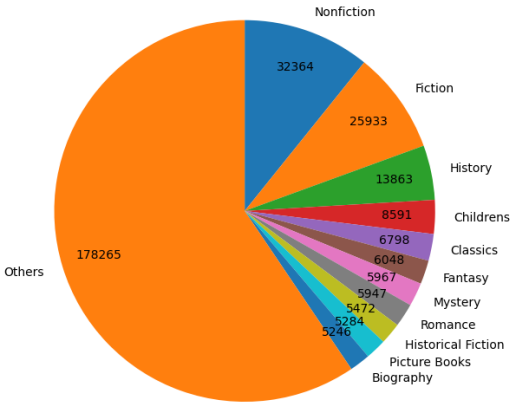

Figure 9: Books per Genre

Languages Dataset Analysis

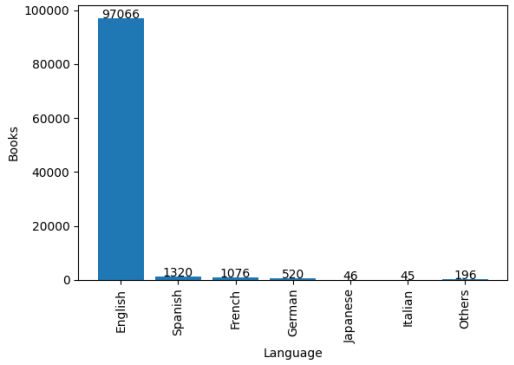

Figure 10: Books per Language

Authors Dataset Analysis

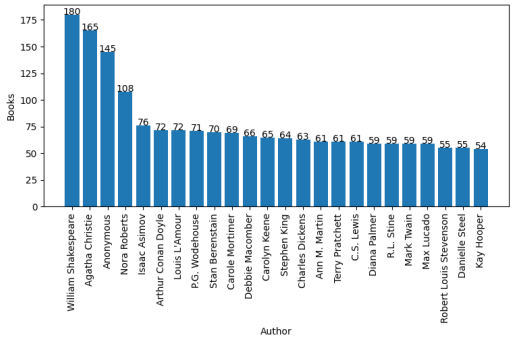

Figure 11: Books per Author