Goodreads Books and Reviews

Information Processing and Retrieval

Inês Silva
FEUP, Porto, Portugal
up201806385@edu.fe.up.pt

Mariana Truta
FEUP, Porto, Portugal
up201806543@edu.fe.up.pt

Rita Peixoto
FEUP, Porto, Portugal
up201806257@edu.fe.up.pt

ABSTRACT

In the current days, we come across big amounts of data and so an increasing concern to index and search efficiently appears. In this paper, one can see the process of dataset preparation and refinement with the goal of creating a book search system.

KEYWORDS

Goodreads, books, reviews, Dataset, data, preparation, data analysis, information, retrieval, processing, refinement, search engine

1 INTRODUCTION

This paper is developed within the course of Information Processing and Retrieval of the first year of the Master in Informatics and Computing Engineering.

A choice was made to work with a dataset about a very well-known book community website, Goodreads.

The motivation for choosing this dataset is the big appreciation for books and frequent use of the Goodreads website, causing the necessity of a better search system.

This paper starts by describing the dataset used and what data preparation and enrichment was applied to it, followed by a data source and quality assessment. Adding to that, it also presents the data processing pipeline and the domain conceptual model, ending with the conclusions and future work.

2 DATASETS

2.1 Books

The main dataset chosen contains the general information needed to describe a book, gathered from Goodreads website. It was retrieved from **Goodreads 100k books**, where the author retrieved the data by scraping Goodreads website.

This dataset has both numerical data, such as the number of pages, publish data, and textual data, such as the book description, genres, etc.

2.1.1 Data Preparation. The initial dataset contained a 100 000 books. After extensive analysis, however, it was noticed that the data was not as good as expected, as the original dataset had not been encoded correctly.

To initiate this process, it was first necessary to assess the consistency of the rows and the relevance of each column. With OpenRefine, it was observed that there were several cases of missing values and, after analyzing some of these books on Goodreads, it was found that the website did not contain all their information and therefore it did not make sense to include them in the dataset. In other words, all the lines with missing values were discarded.

After that, using python scripts with the pandas library, the data consistency was evaluated. The first task was to explain why there were 3 003 with 0 pages, and since this information was available in Goodreads for some books, a web scraping was performed to replace the missing pages. Nevertheless, there were still 2943 books in this situation. The importance of these situations was discussed, and an agreement was reached that the absence of these values was not that problematic for the final goal since these books may not have pages either because they are audiobooks or because there is no information on Goodreads.

Afterwards, the book format values were analyzed and cleaned, having translated some of them into English and regrouped others to normalize this column. It was noticed that the isbn13 column had poorly standardized values and therefore it was removed from the dataset since it was not relevant for the context and its preparation would become unnecessary. Posteriorly, it was also found that some rows had repeated genres which made it necessary to remove the duplicates.

Following this initial preparation, the special characters of the textual fields (description, title and authors) were cleaned. The same process was used in these three columns, which consisted of web scraping the books containing special characters. The books that were still in this situation were eliminated after this processing, since these books now included characters from non-European alphabets.

At last, all the extra whitespaces that were in the dataset were trimmed so that the final dataset would be as clean and ready as it possibly could be for the next milestones.

Table 1: Number of missing values on each column of the original dataset

author	0	**isbn13**	11435
bookformat	3228	**pages**	7752
desc	6772	**rating**	1562
genre	10467	**reviews**	0
img	3045	**title**	1
isbn	14482	**totalratings**	0

2.1.2 Properties characterization. Throughout the analysis of the dataset, was gathered information regarding the mean value, standard deviation, minimum and maximum values for each of its numerical properties (*pages*, *rating*, *reviews* and *totalratings*). These statistics are shown in Table 2.

PRI 2021/2022, FEUP,
2021.

Table 2: Statistics for the numerical values of the dataset

	pages	rating	reviews	totalratings
mean	276	3.89	182	2991
std	375.35	0.39	1449.45	36353.38
min	1	1.00	0	0
max	70000	5.00	158776	3819326

To characterize and better understand the categorical properties, it was found the most common values for each and analysed how these related to the numerical properties.

(1) *genre*: there are 1182 different genres in this dataset. It was created a *"Word Cloud"*, shown in fig. 1, to visually represent the existing genres, giving greater prominence to words that appear more frequently. It was also found the 20 most common genres, showing the number of books each appears in, average number of ratings and reviews, and average rating of these books in fig. 5
(2) *author*: there are 68767 different authors in this dataset. It was found the 20 authors that have the most books in the dataset, showing the number of books each one wrote, average number of ratings and reviews, and average rating of these books in fig. 6
(3) *bookformat*: there are 203 different book formats in this dataset. Ot was found the 5 authors that have the most books in the dataset, showing the number of books each one wrote, average number of ratings and reviews, and average rating of these books in fig. 7
(4) *title*: since there are 100000 books in this dataset, it was found the top 20 books with the highest number of ratings, showing the number of reviews and rating of these books in fig. 8

Figure 1: Word Cloud with the genres that occur in the dataset

2.2 Reviews

To enrich the main dataset, it was used web scraping to collect reviews written by Goodreads users to the books, since the dataset only has the number of reviews and the numerical rating.

2.2.1 Data Preparation. The original scraped dataset contained 510 709 entries, of which 130 were empty. These 130 entries were removed and more 819 were removed because they contained special characters.

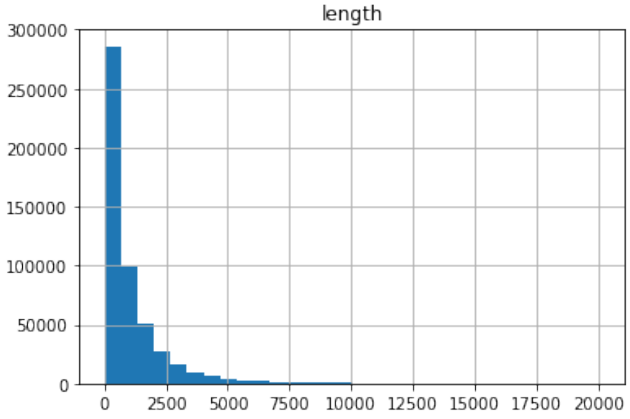
The remaining entries went through a process of data cleaning that removed unnecessary characters like '[]' and '"' that surrounded the reviews. It also removed the escape characters, the single quotes in the start and end of the review.

In the end, the dataset ready to used contained 509 760 entries.

2.2.2 Properties characterization. For the reviews' dataset, it was found relevant to know the average length of a review, and it is 1050 characters.In fig. 2, one can see the distribution of the length of the reviews.

As for most common reviews, it is the word "good" as there are 411 reviews with this value.

Figure 2: Distribution of the length of the reviews



2.3 Data Source

As for the authority of the data source, one can tell the author has already been working on datasets for a few years (has participated in a scholarly competition 8 years ago), has at least 5 datasets and has published some of his work in the last year around this theme. There are also good comments on his discussion.

This dataset was a personal project of his to learn to scrape and was published in a very well-known dataset website, Kaggle and he also gave credits to the website from where it was retrieved and shared the code of the program.

Therefore, it is concluded that it is a good data source.

2.4 Data Quality

To perform the data quality assessment were used five metrics of data quality: completeness, correctness, timeliness, consistency and integrity.

As for completeness, the dataset is 96% complete.

The level of correctness is 98%, as only a few values from the pages are not correct, and the rest is just the format of description,

book format and authors that do not comply with the expected format, as they have invalid characters or multiple languages.

In terms of timeliness, one can say that it is up-to-date for the intended use, since it was retrieved 5 months ago and the only differences noticed are the number of reviews and the rating, so it doesn't seem to be problematic for the final goal.

There are some issues regarding the consistency of some properties like the book format that, as said, contains data in different languages.

Finally, in terms of integrity, it is not possible to guarantee that all the counted reviews on the "review" column of the books' dataset are presented in extension on the reviews' dataset. The reviews' dataset was not made to extract all the reviews for each book, but the initial displayed reviews on the book's page that do not all correspond to all reviews of the book, if there are a lot.

3 DATA PROCESSING PIPELINE

In order to achieve a greater quality of the chosen data, various processing tasks were executed. These steps are represented in the data pipeline, shown in fig. 4.

3.1 Data Collection

Using the link to the webpage associated with each book, web scraping was performed. This was done with 3 purposes: to replace book descriptions that were inserted in the dataset with the wrong encoding and thus had special characters that made the text unreadable, to fill missing values in the *"pages"* column (books with 0 pages), and finally to get up to 10 reviews of each book in order to create a new dataset that allows the exploration of the reviews' textual data.

3.2 Data Cleaning

After the scraping phase, some cleaning tasks were needed to further improve the data quality.

Regarding the book dataset, the book formats were standardized (capitalized first letter, translated to English and aggregated similar values), the *"isbn13"* column was removed due to most values being missing and many inconsistencies in the data, and duplicated values in the *"genre"* column of each book were removed. Finally, special characters were removed from the *"title"* and *"author"* columns.

In the review dataset, the cleaning steps consisted of removing special characters, clening up the text and removing null reviews.

3.3 Data Enrichment

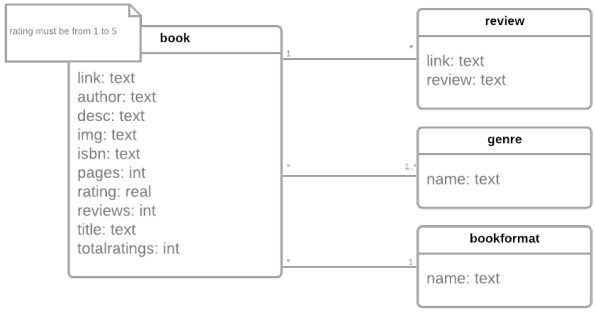
The final step of the developed pipeline consisted of combining the datasets that resulted from the cleaning stage in a sql database. The conceptual model is described in section 4.

4 CONCEPTUAL MODEL

As shown in fig.3, the conceptual model consists of four main classes:

- book: each book has a Goodreads link, a list of its authors, its description, a representing image, its isbn, the number of pages, its rating, the number of reviews, its title and the number of total ratings;

Figure 3: Domain Conceptual Diagram



- review: each review consists of the link of the book it relates to and its text;
- genre: category of book characterized by a particular style, form or content;
- bookformat: format of the book.

5 CONCLUSION

In this paper is presented the process the datasets' went through to reach its final state, ready to use.

Throughout this milestone, the datasets were well analysed and studied in order to conclude which data cleaning and preparation tasks were necessary for them to work to the project goal.

Being happy with the final result, one can say the datasets are ready for the goals of the next milestones', having gone through cautious analyse.

6 FUTURE WORK

At the end of this milestone, it was realized that the discovery of some issues of the dataset was rather late. Because of that, a lot of new scraping was made, which took out a lot of the exploration time for future implementations.

In the future work, is expected to build indexes, retrieve information and an integration of the datasets in a search system.

REFERENCES

[1] Dhamani, M., 2021. GoodReads 100k books. [online] Kaggle.com. Available at: <https://www.kaggle.com/mdhamani/goodreads-books-100k>
[2] Openrefine.org. n.d. Documentation · OpenRefine. [online] Available at: <https://openrefine.org/documentation.html>
[3] Goodreads. n.d. [online] Available at: <https://www.goodreads.com/>
[4] Pandas.pydata.org. n.d. pandas 1.3.4 documentation. [online] Available at: <https://pandas.pydata.org/docs/>
[5] Matplotlib.org. n.d. Matplotlib 3.4.3 documentation. [online] Available at: <https://matplotlib.org/stable/contents.html>

7 ANNEX

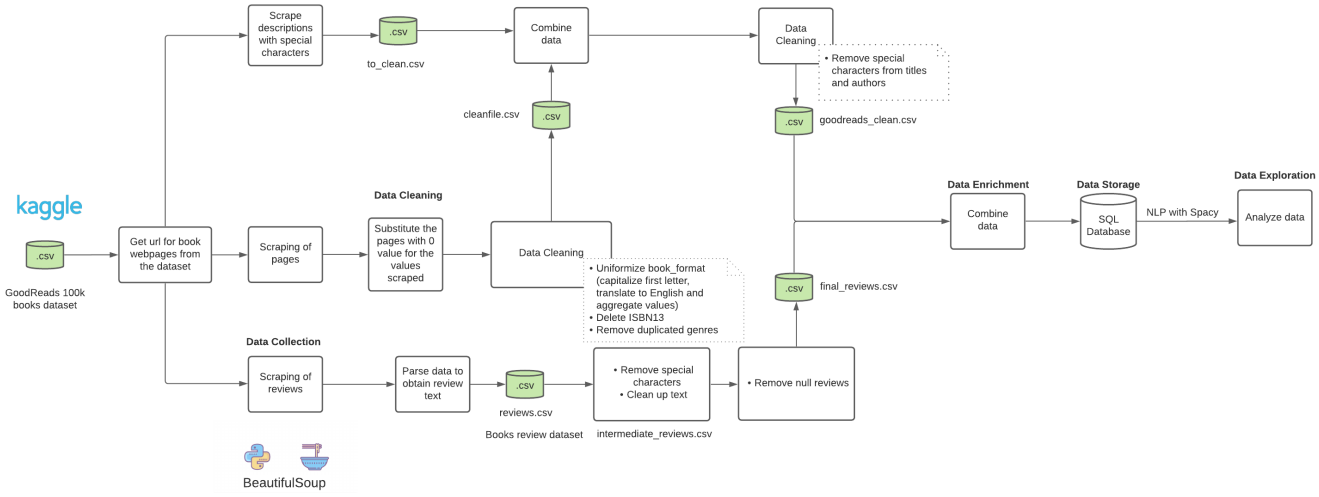


Figure 4: Data processing pipeline

Figure 5: Average number of reviews and ratings, and average rating of the books that belong in the top 20 genres

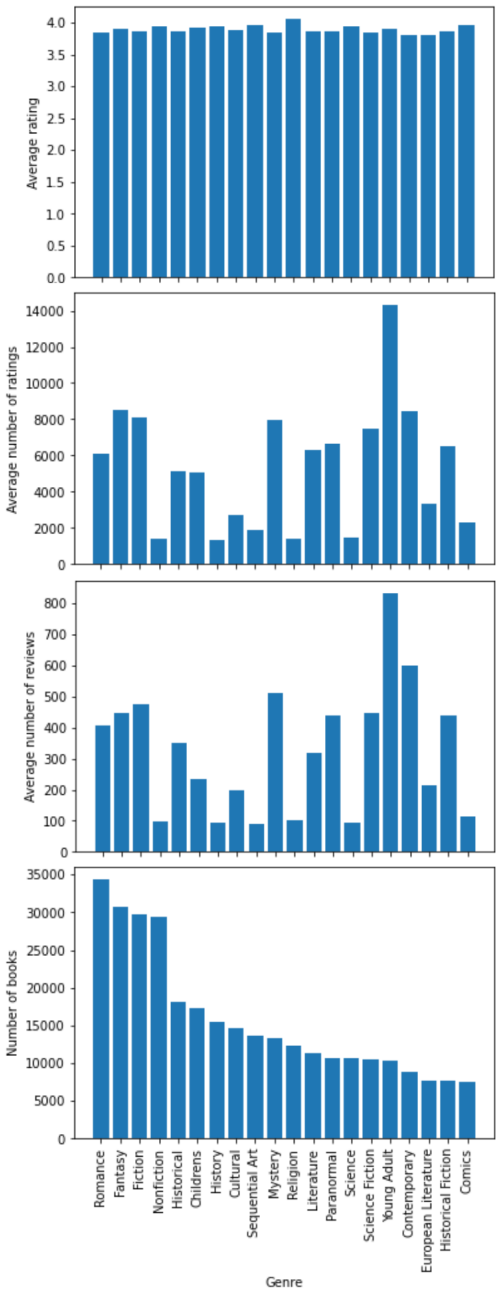


Figure 6: Average number of reviews and ratings, and average rating of the books written by the top 20 authors

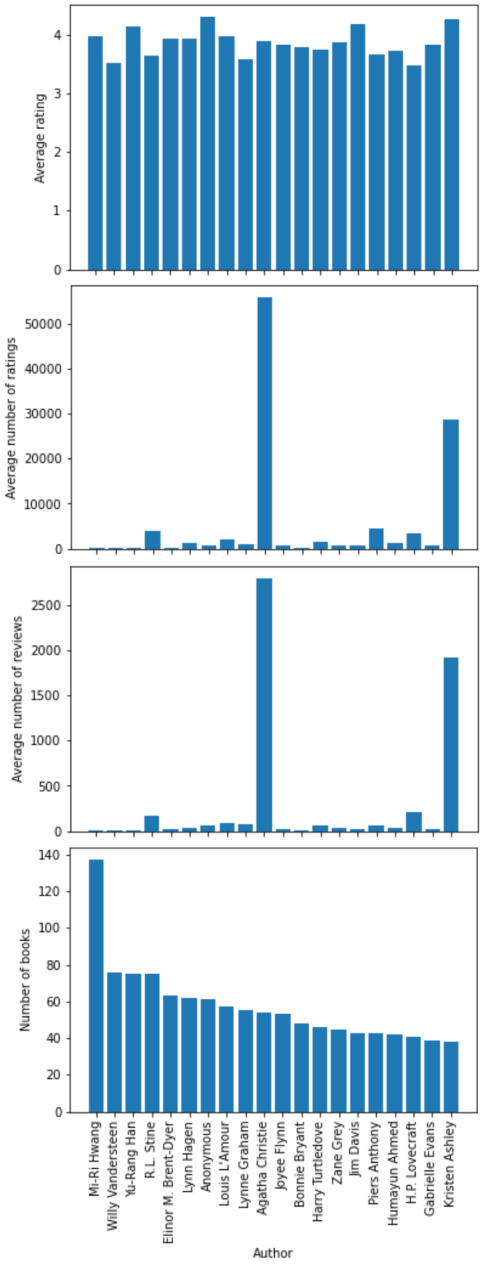


Figure 7: Average number of reviews and ratings, and average rating of the books that belong in the top 5 book formats

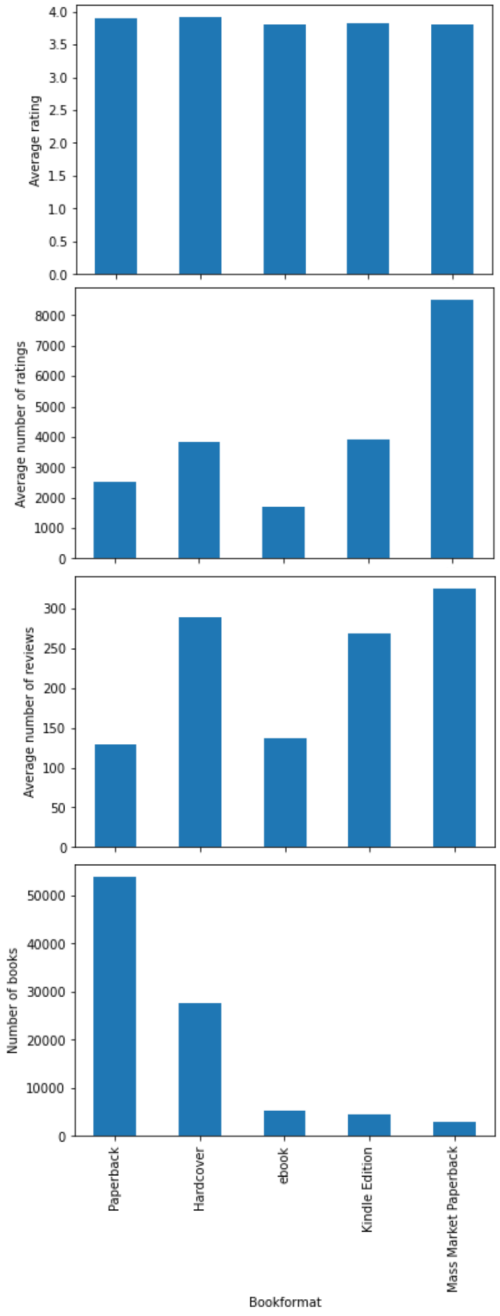


Figure 8: Average number of reviews and average rating of the top 20 books with most ratings