Goodreads Books and Reviews

Data Preparation

Master in Informatics and Computing Engineering, FEUP
Group 53, Information Processing and Retrieval

Diana Freitas, Diogo Samuel Fernandes, Juliane Marubayashi
up201806230@edu.fe.up.pt,up201806250@edu.fe.up.pt,up201800175@edu.fe.up.pt

ABSTRACT

The following article describes in detail the preparation and characterisation of a group of datasets about Goodreads books and their respective reviews.

It documents the development of a Data Processing Pipeline, from the selection of the datasets, to the Exploration of the processed data. All the intermediate processes are also explained, namely the assessment of the Data Quality, the Data Cleaning methodology, and the representation of the processed data with a Conceptual Model.

KEYWORDS

Databases, Goodreads Dataset, Data Preparation

1 INTRODUCTION

It wasn't that long ago when people used to get together to decide the next book to read. By engaging in social interactions, either in personal or professional environments, readers could get book recommendations and avoid the difficult task of choosing the right book by chance.

Nowadays, with the influence of the internet in the globalization process, the access to information is at our fingertips. The ease of access to information, however, also comes with a great cost: the increasing amount of data has made it difficult to effectively make decisions. According to the article *Information overload: the science behind bad decisions* [4] "Information overload shuts our brains down", "The more choices we have, the more challenging the decision". Given that *Big Data* [1] increased our level of indecision, it became central to filter the information according to our needs.

Currently, the idea of data filtering is widely spread, and websites such as Goodreads, the world's largest website of book details and recommendations, are becoming more and more popular.

By taking into account this information overload issue and the need to filter the available data, specially in the scope of Book recommendations, the purpose of this project is to gather and process a subset of information about Goodreads books, with the ultimate goal of providing valuable book details and recommendations to readers, through the development of an information search system.

This article covers the multiple tasks involved in the collection, preparation and characterisation of datasets regarding Goodreads books, reviews, genres and authors, justifying the decisions that were made throughout the process.

Supervised by João Paulo Madureira Damas.

Goodreads Books and Reviews, PRI, FEUP
2021.

Firstly, the main stages of the process will be introduced with a brief explanation of the pipeline's structure. In the following sections, each of the pipeline's stages (Data Gathering, Cleaning and Exploration) will be explained in more detail. The methodology used to assess the quality of the data and the issues that were uncovered will also be addressed in a separate section. Finally, a description of the conceptual model of the data domain is also included, followed by a brief conclusion about the obtained results.

2 PIPELINE

A data pipeline consists of a chain of data processing elements, that represent the flow and transformation of the data. In the pipeline that resulted from this project, three stages can be clearly identified:

- Data Gathering
- Data Cleaning
- Data Exploration

The **Data Gathering** stage, simply shows the four original **json** datasets - books, authors, reviews and genres; which were downloaded from UCSD book graph[2].

This stage is followed by the **Data Cleaning** process, which consists of two consecutive tasks:

Firstly, it is responsible for modifying the data of each dataset individually, by removing irrelevant columns, converting data types or removing observations with missing values. After applying the desired transformations to the data, each dataset was stored in **csv** format. This output format was selected, because it consumes less disc space and it is easier to work with. The format conversion could not be performed at the beginning of the pipeline, since there were nested **json** objects.

The second task of the **Data Cleaning** stage combines the processed datasets by using the ids of the books and of the authors. When combining the datasets, some observations that had references to others that were dropped in the previous step were removed (e.g. Some reviews referred books that were removed for not having a description). Also, as a result of this processing step, some many-to-many relations, in particular, the relations between genres and books, and between authors and books were extracted. These relations were represented, in the original data, as columns that stored an array of foreign keys to the other relation.

Last but certainly not least, the **Data Exploration** process, unlike the other pipeline stages, does not modify the datasets. However, its performance and value are significantly impacted by the previous stages of the pipeline. This last stage of the pipeline refers to the use of data visualization and statistical techniques to better

characterise the datasets and identify valuable information about the nature of the data.

3 DATA GATHERING

Initially, when selecting the data that would be used for the project, the idea was to extract it from Goodreads.com using Web Scrapping, in order to get up to date book details and reviews. However, Goodreads' Terms of Use clearly state that Data Gathering is no longer allowed, and that new developer keys for the public developer API are no longer being issued.

However, some datasets were available from the time when scrapping was allowed and when the Goodreads API was still active. From the available sources, we decided to use the datasets from *Ucsd book graph* [3], which provides Goodreads datasets for academic use, collected in late 2017 and updated in 2019. These datasets include book meta-data, user-book interactions and book reviews, and can be merged together by matching book/user/review ids.

4 DATA QUALITY

When selecting the datasets that would be used for the project, it was important to assess the quality of the data and how well suited it was to serve the project's purpose. Therefore, during this Data Assessment stage, a search was performed in order to find datasets that were rich in textual data, which, considering the selected theme, would probably be found in reviews or book descriptions. Also, the most common data quality problems were assessed, by identifying the quantity of missing or inconsistent values, by addressing the validity of the data, in particular, checking if the data conformed with type, unique and foreign-key constraints, and by searching for uniformity problems. Event though the four selected datasets had few data quality issues, some problems, that would need to be fixed in the Data Cleaning stage were detected, namely:

- The date format was not consistent in all the datasets;
- Many integer ids were stored as strings;
- There were many books without any genre associated (missing values).

5 DATA CLEANING

After exploring the multiple datasets that were available regarding Goodreads books, four of them were selected for the project - books, authors, genres and reviews. There was also a user dataset available in the same source, however, as most of the relevant user information was not available due to data protection, the available client details were discarded. Once the datasets were chosen, their volume was evaluated and their most relevant attributes were identified, which was essential to determine the size of the subsets to use, which attributes to drop, and which columns to aggregate or clean. The Data Cleaning process, which will be described and justified in detail below, was first performed on each dataset individually, using the **json** python module. The datasets were then stored in **csv** files in order to simplify the integration.

5.1 Books

The original books dataset (**books.json**) consists of about 2.3M observations, each with 29 attributes. This dataset was initially reduced to a subset of 11 thousand observations, to allow a faster processing. All the selected books have a description, which will be used in forthcoming milestones for text search. Furthermore, the publication_year, publication_month and publication_day were merged into a single date column. Some columns were also dropped mainly due to one of the the following reasons:

- they were references to observations of datasets that were not used (series and work_id);
- they were related to commercial details about the book (asin and kindle_asin);
- they contained derived information that didn't make sense when using subsets of the original datasets (text_reviews_count and ratings_count);
- they contained information that was already represented in another dataset (e.g. instead of using the attribute popular_shelves, which represented the genres attributed by the users, a genres dataset, that was also available, was used);
- they did not add relevant information to the data (country_code, link, url, title_without_series).

Also, the similar_books attribute was dropped, because the similarly between books can easily be evaluated when searching the dataset, by filtering by genre or author, for example.

5.2 Authors

For the authors dataset, onlythe essential information was kept, dropping some attributes that were derived from the original and larger dataset, in particular the text_reviews_count and the ratings_count.

5.3 Genres

In Goodreads, the genres of each book are defined by the users, and therefore, for each genre associated with a book_id, the original dataset included the number of users that categorized it with a certain genre. In this first Data Cleaning phase, the number of users that classified the book with each genre was discarded, only keeping the names of the genres associated with each book_id. Also,all the observations that represented books without any genre classification were removed, and a subset of 11 thousand observations was selected to ensure an easier processing.

5.4 Reviews

The reviews dataset included more than 1.3M English book reviews of 25475 books. A subset of 10 thousand reviews was used, dropping the user_id and some details related with the management and statistics of each review, in particular the date_updated, started_at, read_at, n_comments and n_votes. Also, as this dataset included reviews with spoilers, which are labeled using '(view spoiler)[' and '(hide spoiler)]', the spoilers were kept, but their delimiting tags were removed.

5.5 Combining the datasets

The resultant datasets were then combined in the following order:

(1) The genres dataset was combined with the books dataset by book_id, dropping all books that did not have any genre observation and all genres observations that were not related to any book. Then, the **genres.csv** was adapted to map the name of each genre to an id. Finally, by combining this datasets, we extracted a many-to-many relation between books and genres, which was also saved as a **csv** file.
(2) The authors of each book, which were originally stored in a list in the authors column of **books.csv**, were combined with the books dataset by author_id, resulting a many-to-many relation which maps author_id with book_id. Also, all authors that were not associated with any books of the chosen subset were discarded.
(3) The reviews were combined with the books, in order to discard reviews that were not related to any book of the chosen subset.

6 DATA EXPLORATION

In the last step of the data processing pipeline, the exploration of the processed datasets uncovered some patterns, and characteristics of the data, by means of visualization, with distribution and correlation plots and by generating tables that expose descriptive statistics - data types, number of missing values, maximum, minimum, mean, among others. This process does not aim to explore every bit of information of the Goodreads database, but rather to improve the understanding of the data properties and relations, which will be determinant for the success of the final search system. In the next few sections the most interesting results for each of the tables will be discussed and illustrated.

6.1 Books

The books dataset contains a collection of 9483 books, with a representative amount of exemplars available in physical format.

To attest the variety of books in the dataset, the range and distribution of some of the attributes that refer to book meta-data, such as the number of pages, was analysed. If the dataset was only composed by books with a number of pages between 1 and 100, for example, it could mean that the sample lacked diversity. As the plot shows [1], this scenario wasn't materialized.

In addition to verifying the diversity of a database, studying the composition of the data is also important. The books dataset contains many missing values, namely the isbn, language_code, format, publisher, num_pages, isbn13 and edition_information. The missing information wasn't treated neither dropped, since it's common that some books don't have an isbn identification, for example, and this fact does not influence the quality of the book.

Finally, a word-cloud was used in order to capture the main ideas of the books meta-data.

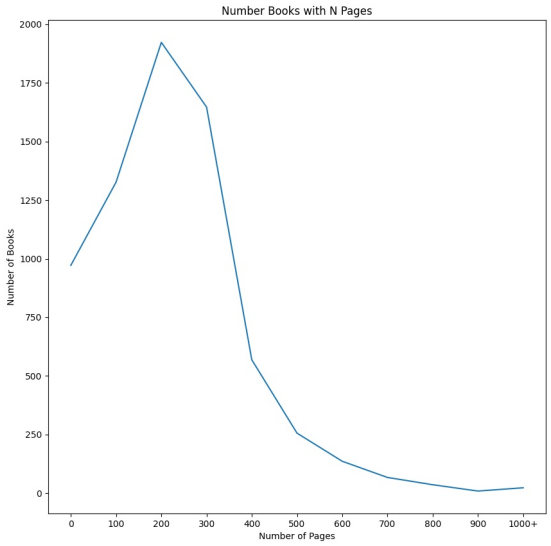

Figure 1: Distribution of the number of pages

6.2 Authors

The author database consists of 11508 authors, of which only 231 have an average rating, which implies that not all authors are popular.

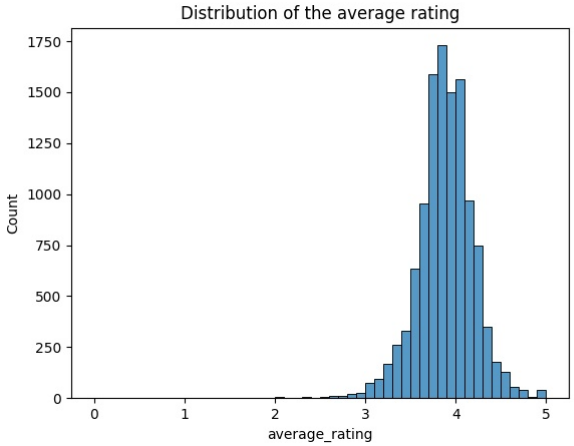

Figure 2: Average rating histogram

Moreover, each author has written at least one book, and, even thought the majority of the authors has composed just one book, there are some outliers, particularly some books with 50 and 51 authors. Fortunately, the presence of outliers doesn't cause a negative impact in this research, thus these observations don't have to be dropped. In fact, they provide us valuable data.

6.3 Genres

The genres table contains 16 types of genres, from which the most common one is fiction, with over 6000 samples, and the least common is poetry.

An interesting fact is that the histogram for each of the genres follows a shifted normal distribution, where the center comes around the rating 4.

6.4 Reviews

By exploring the review dataset, which is composed of 510 reviews, it was concluded that only 111 books contain a review. Following the same pattern of the genres and authors, the reviews' rating histogram follows a shifted normal distribution with mean around the number 4.

Finally, by analysing the word-cloud generated from the text of the reviews, the presence of some words that might influence the book rating, such as "love", "enjoyed", "long", among others, was noticed. However, after analysing the correlation between those words and the ratings, no correlation was detected.

Figure 3: Reviews word-cloud

7 CONCEPTUAL MODEL

The datasets that were used in the project are directly related to the main entities of the Goodreads Conceptual Model - Book, Author, Genre and Book Review. Each book has a title, a description, and many other attributes that describe essential information about it, like the publisher, number of pages and isbn, among others. Also, a book is either an ebook or a physical book, and can have many authors and many reviews. Each review has a date, a rating and a text field, which may be used for text-search. A book may belong to multiple literary genres, which are simply characterised by their name.

The conceptual model in figure 4 represents the Goodreads' data domain.

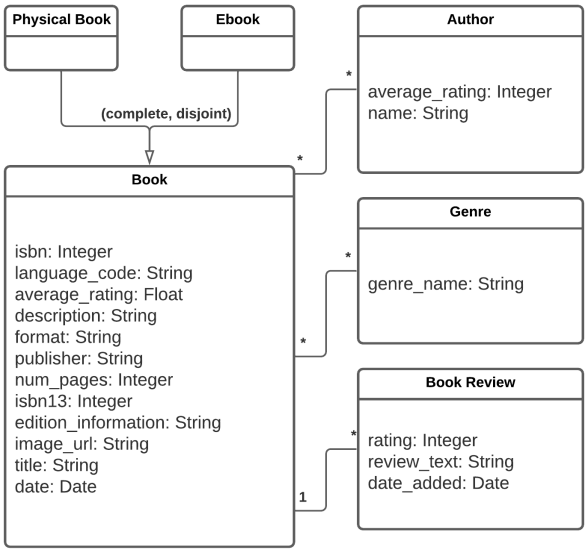

Figure 4: Conceptual Model

8 CONCLUSION

To sum up, the goals proposed in the first phase of the project, which involved the collection, preparation and characterisation of the selected datasets, were achieved with great success. As a result, the final datasets, which represent a subset of the original data, are free of irrelevant attributes and, when combined, clearly represent the main entities of the Goodreads domain - Books, Authors, Reviews and Genres; as well as the relations between them.

The technologies that were selected for the Data Preparation and Exploration, particularly some **Python** libraries like **json**, **csv**, **Pandas**, **Matplotlib** and **Seaborn**, had a great impact on the success of this phase, given that they are intuitive and provide methods to deal with most of the steps involved in a Data Processing tasks.

As future work, an information retrieval tool will be used on the project's datasets and the data will be explored with free-text queries.

REFERENCES

[1] [n. d.]. Big Data. https://pt.wikipedia.org/wiki/Big_data
[2] [n. d.]. Information overload. https://en.wikipedia.org/wiki/Information_overload
[3] [n. d.]. UCSD Book Graph. https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/books?authuser=
[4] Joe McCormack. 2016. Information overload science behind bad decisions. https://thebrieflab.com/blog/information-overload-science-behind-bad-decisions/
[5] Mengting Wan and Julian McAuley. 2018. Item Recommendation on Monotonic Behavior Chains. (2018). https://doi.org/10.1145/3240323
[6] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. (July 2019), 2605–2610. https://doi.org/10.18653/v1/P19-1248

A PIPELINE FLOW