Goodreads' Books and Reviews

Diogo Almeida
up201806630@edu.fe.up.pt

Pedro Queirós
up201806329@edu.fe.up.pt

ABSTRACT

Over time, the amount of data available in the internet has grown in an unimaginable rate, reinforcing the need to have mechanisms to connect all the available information. This document concerns the data preparation process of our information system on Goodreads' books and its reviews. To obtain a dataset with relevant and suitable information~~ for the theme,~~ data refinement and enrichment were performed. Furthermore, the dataset was analyzed for a better understanding of the available data, with some statistics being made for that same purpose.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Dataset , Data Extraction, Data Preparation, Data Cleaning, Book, Review, Domain Conceptual Model, Pipeline, Python, Pandas

1 INTRODUCTION

Books have been playing an important role in society since the early times of humanity. Nowadays, books can be read in many different formats and have numerous purposes: to transmit knowledge about a specific subject, to tell a story, to serve as a record for future generations, among others.

The current panorama of book search systems is ~~pretty decent~~ in regards to the information that it is able to retrieve, letting users mainly search for titles, authors or genres. The main goal of this project is to complement these types of search systems with a search engine that allows users to also search for books based on ratings, reviews, descriptions, awards, ~~etc.,~~ in order to provide an easier and better experience when trying to find a book to read that fits their preferences.

This article describes the first development phase of this search system and is divided into several sections that characterize the several steps taken:

The first section, **Data Extraction and Enrichment**, describes the process of data gathering and enrichment as well as detailing the source of the information. This is an essential step to guarantee that relevant data for the problem is used to create the first version of the dataset.

The next section is **Data Preparation**, that details the procedure of cleaning and refining the collected data in order to have a consistent and practical dataset that is easier to handle.

The **Domain Conceptual Model** section describes how data will be organised and structured in the domain. It is followed by **Possible Search Tasks** ~~section,~~ in which ~~all~~ the possible queries to do in the system are listed.

Finally, the **Dataset Characterization** section details how tools like plots and graphs are essential to better explore and understand all the collected information.

2 DATA EXTRACTION AND ENRICHMENT

In order to extract the best information for the project, various open data platforms were consulted. After some research, Kaggle [6] proved to be the best solution. Kaggle is a platform for data scientists and machine learning practitioners to publish datasets and explore and build models. The dataset from Kaggle, **'Goodreads Books - 31 Features'** [7], which has around 150 MB, contains around ~~50.000~~ books with 31 features each. These features include several relevant aspects about a book, like its title, author, description, publisher, various details about its rating (1-5 stars) and more.

After an initial review of the Kaggle dataset, it was decided to enrich the project's data further with more text-rich fields~~, therefore,~~ another dataset was created using some reviews from each book in the Kaggle dataset. These reviews were scraped from the Goodreads website [5] and include the review's author, text and publication date. The information was scraped and exported to a .csv file using a Python script with Requests [1] and BeautifulSoup [2] ~~librarys.~~ This .csv file ended up with a size of around 372 MB, containing around 8 reviews for each book.

3 DATA PREPARATION

Data preparation was made on the dataset [7] to make sure that it is consistent and ready for an information retrieval phase. Before starting this operation the data was analysed carefully to ensure that only the relevant data for the project was selected based on the possible search tasks that could be done.

Firstly, ~~remove~~ the columns that don't have relevant information or that have a significant amount of NaN values, for instance, the "asin" (Amazon Standard Identification Number) column.

Secondly, books that didn't have an original title got this column filled with their title.

Thirdly, the books without all the relevant information, for instance title, publisher or author, were removed from the dataset.

Next, due to bad encoding in the original dataset, there were some strange characters appearing in books that had column values that weren't in English, ~~this~~ was fixed so all characters from all languages appear correctly.

After that, books with descriptions containing less than 25 characters ~~are~~ also removed for not having enough information.

Books that weren't in the reviews dataset were removed. Although the reviews dataset was scraped using the books from the original dataset, it wasn't possible to scrape all of them due to errors, so they had to be removed from the original as well.

Finally the referenced books in the "recommended_books" or "books_in_series" columns that weren't in the same dataset were also removed. Some columns (awards, genres and their votes, and characters) were extracted to a new .csv file to make it easier to analyse the final data that will be used ~~for the information search system.~~ These columns were also removed from the original dataset in order ~~to not~~ have duplicated information. All of these operations were made using Pandas [3], a Python library.

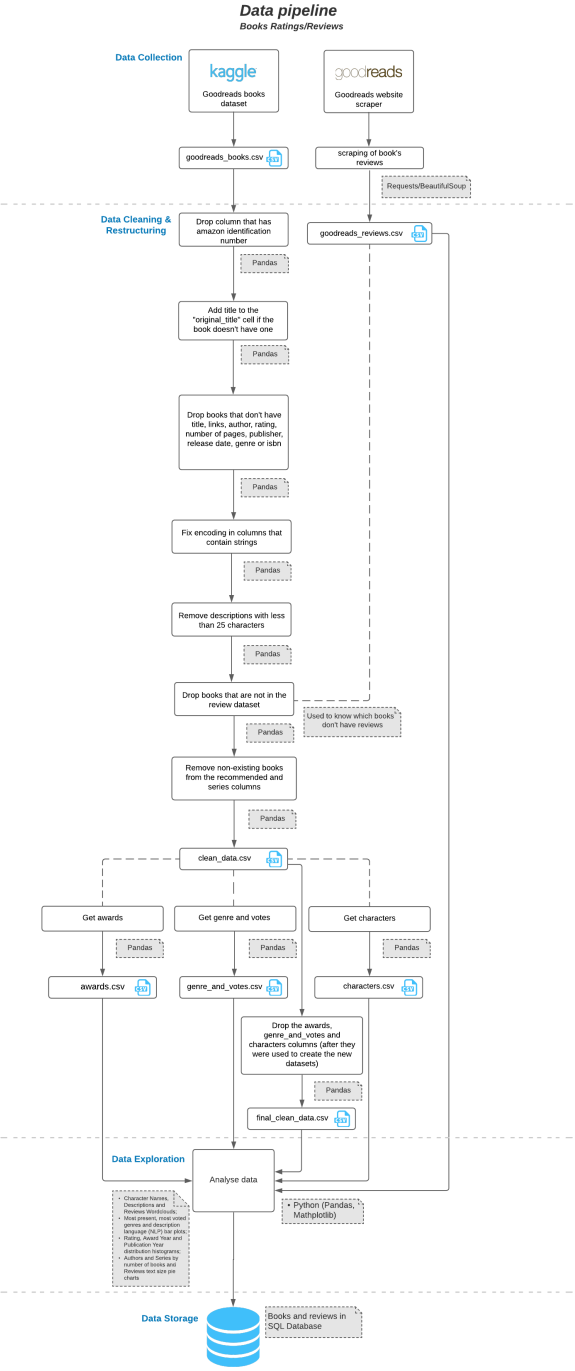
Data pipeline
Books Ratings/Reviews
Data Collection
kaggle
Goodreads books dataset
goodreads
Goodreads website scraper
goodreads_books.csv
scraping of book's reviews
Requests/BeautifulSoup
Data Cleaning & Restructuring
Drop column that has amazon identification number
Pandas
goodreads_reviews.csv
Add title to the "original_title" cell if the book doesn't have one
Pandas
Drop books that don't have title, links, author, rating, number of pages, publisher, release date, genre or isbn
Pandas
Fix encoding in columns that contain strings
Pandas
Remove descriptions with less than 25 characters
Pandas
Drop books that are not in the review dataset
Used to know which books don't have reviews
Pandas
Remove non-existing books from the recommended and series columns
Pandas
clean_data.csv
Get awards
Get genre and votes
Get characters
Pandas
Pandas
Pandas
awards.csv
genre_and_votes.csv
characters.csv
Drop the awards, genre_and_votes and characters columns (after they were used to create the new datasets)
Pandas
final_clean_data.csv
Data Exploration
Analyse data
• Character Names, Descriptions and Reviews Wordclouds;
• Most present, most voted genres and description language (NLP) bar plots;
• Rating, Award Year and Publication Year distribution histograms;
• Authors and Series by number of books and Reviews text size pie charts
• Python (Pandas, Mathplotlib)
Data Storage
Books and reviews in SQL Database

Figure 1: Data Pipeline

4 DOMAIN CONCEPTUAL MODEL

A domain conceptual model was made to represent the main entities of the domain. The book class is the center of the conceptual model and it has the following attributes: Goodreads id, the title, the original title, the links for the book's cover, Goodreads page, Amazon and Worldcat, number of ratings, average rating, number of pages, isbn, isbn13 and the book's description. Every book was written by at least one author and published by one publisher. A book can belong to a series (collection), can have settings (which are the locations where the story unfolds) and can have characters. Each book has several genres which have several votes from the readers. Furthermore, they also have the number of votes for the ratings from 1 to 5 and some reviews made from some Goodreads users. Finally, books still have awards with the respective year when they have ~~win~~ it and recommended books based ~~in~~ their genre or rating.

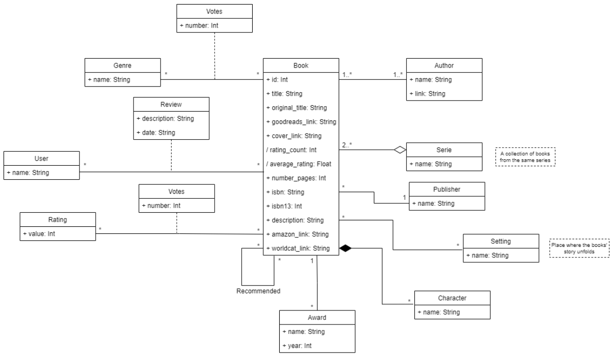

Figure 2: Conceptual Model

5 DATA CHARACTERIZATION

With the aim of better understanding and classifying all the collected data, charts were developed regarding some of the dataset's most interesting and striking features~~, which will structure many of search system's tasks and parameters.~~ These features and respective charts are analyzed ~~thoroughly~~ ~~in the subsections~~ ~~bellow.~~

A. Awards and Books

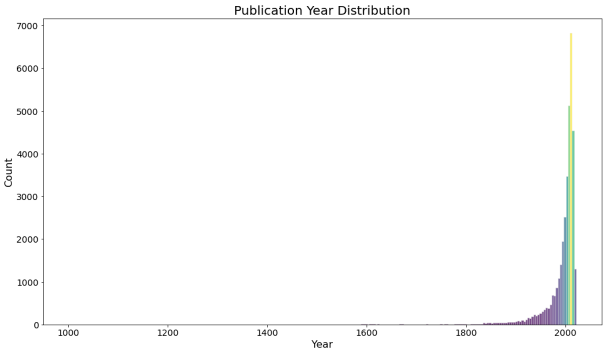

Figure 3: Publication Year Distribution

Regarding the year a book was published, it can be observed in the histogram that the dataset mainly focuses on modern releases, containing many books published after the 2000's. Another interesting note is that there seems to be an ~~exponential~~ increase in included books published from the 1800's onwards.

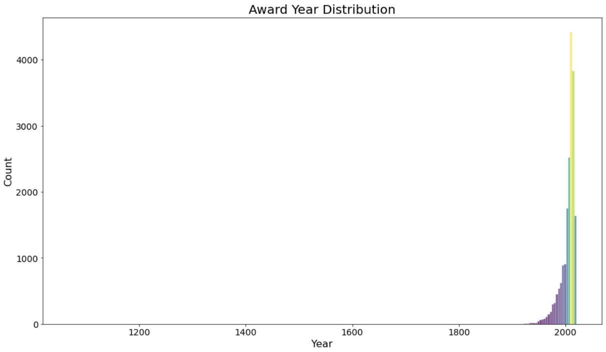

Figure 4: Award Year Distribution

The year an award was given also follows the same trend as the book publication year distribution seen previously, with most awards being handed out after the 2000's. This ~~fact~~ makes sense since most books present in the dataset were published around this time.

About 8,948 of the total of 35,347 books (25.3%) were given an award. Since the dataset contains 19,186 awards, a lot of these awarded books have ~~won multiple trophies:~~ 4,018 to be exact (44.9% of the awarded books). The fact that a ~~decent~~ chunk of books got an award is a good indicator of positive book ratings, a feature that will be analyzed later.

B. Book Genres

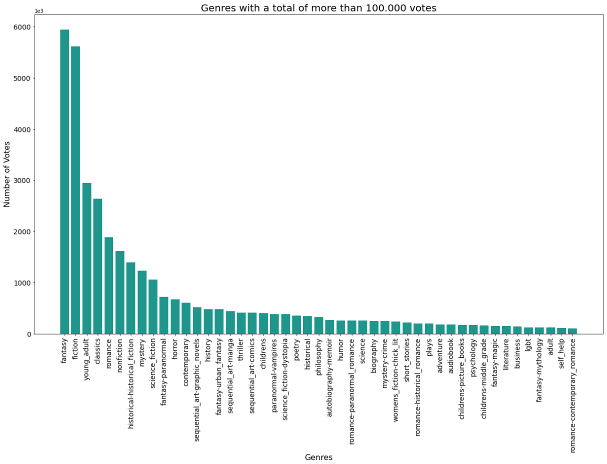

Figure 5: Genres with a total of more than 100000 votes

In the Goodreads website ~~(~~the source of the books dataset~~)~~, users can vote in the genres they think the book fits into. Upon analyzing the plot, the conclusion is that the "fantasy" and "fiction" genres are the most voted, with almost 6,000,000 votes each, 3,000,000 more than the third most voted genre "young_adult".

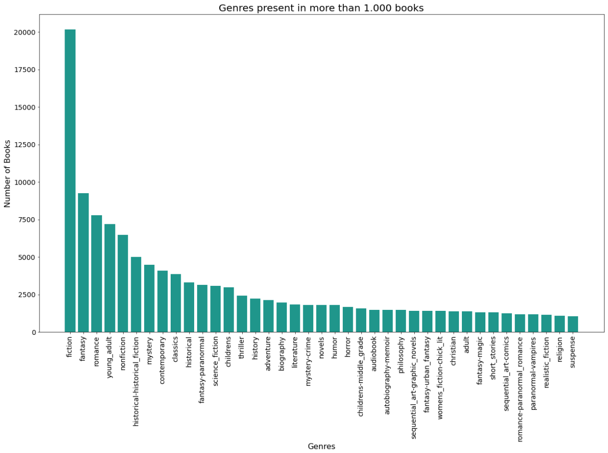

Figure 6: Genres present in more than 1000 books

This bar plot is similar to the previous one, however, it now counts the genres that are present (have at least one vote) in more than 1,000 books. In this chart, we can observe that the "fiction" genre is present in a lot more books than the "fantasy" genre, although both genres were very close in votes, as analyzed in the previous chart (with "fantasy" actually having more votes than "fiction"). From this ~~examination~~ we can deduce that, in our dataset, "fantasy" genre books are fewer, but more popular (have more votes and therefore more users reading it) than "fiction" genre books, that need presence in more books to match the amount of votes that the "fantasy" genre has. In fact, the votes to books ratio in the "fantasy" genre is superior to every other genre.

C. Book Ratings

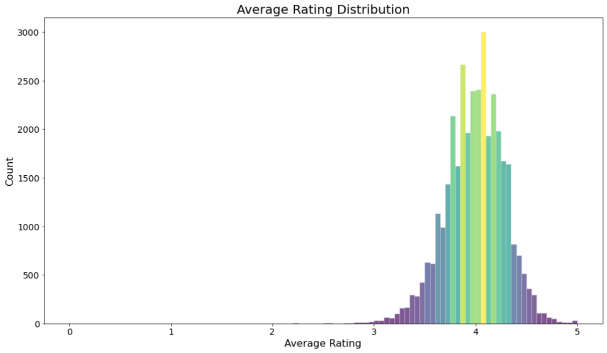

Figure 7: Average Book Rating Distribution

By observing the Average Rating Distribution histogram we can see that, in a classification of 0-5 stars, most books in the dataset have a rating of around 4 stars. This large amount of positive ratings was expected, since in the book awards analysis done before, we saw that there was a good amount of awarded books, which generally means that those books are acclaimed by both critics and readers/users.

D. Authors

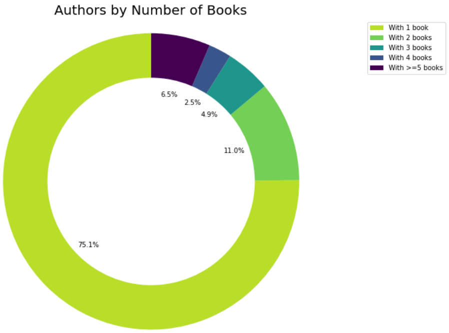

Figure 8: Authors by Number of Books

From observing this pie chart we ~~reach the conclusion~~ that the majority of authors present in the dataset (75.1%) only published 1 book. Since the dataset contains 35,347 books and 19,171 authors, this means 14,405 books were written by authors that only wrote 1 book. Therefore, 20,942 books, which is more than half of the books dataset (59.2%), were written by only 4,781 authors, which averages ~~out~~ at about 4/5 books per author, despite that, according to the pie chart, only 6.5%+2.5%=9% (1716) of authors have written more than 3 books.

E. Book Descriptions and Reviews

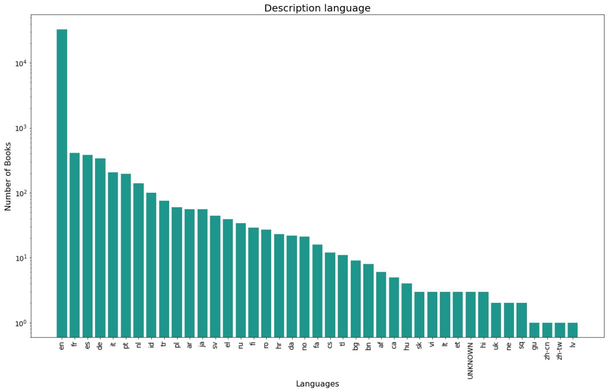

Figure 9: Book Descriptions Language

In this bar plot, we explore the different languages the descriptions were written in. The language for the descriptions was detected using the Python library *spacy-langdetect* [4]. From the chart ~~(that~~ uses a logarithmic scale~~)~~, we conclude that the vast majority of descriptions are in English, with French, Spanish, German and other languages following after with ~~significantly~~ less descriptions. Analyzing the book description language was an attempt of obtaining a rough approximation of each book's original idiom, although some of the descriptions end up being translated to English in the website, especially for popular books.

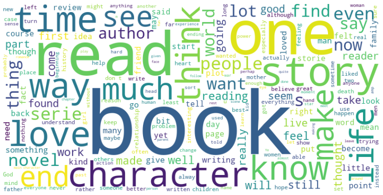

Figure 10: Book Reviews Word Cloud

For the book reviews, we decided to generate a word cloud plot to try and see ~~the general tone of them.~~ Naturally, there are a lot of words associated with the elements of books such as "read", "character", "story" or even the word "book" itself. What is interesting is the large amount of positive words, such as "love", "good", "great", "well", ~~etc.,~~ which might be related to the high rating distribution the books in the dataset have, as analysed before.

6 POSSIBLE SEARCH TASKS

Following an extensive data analysis of the main features that made the group acknowledge the dataset's relevant information, we can now consider some possible search tasks to be used:

- Search for a book by name, author, publisher, number of pages, ISBN, publication date, series or similar books.
- Search for a book by number of genres/characters or specific genres/character names.
- Search for a book by its description: using the description's text length or language.
- Search for a book by its popularity (higher number of rating votes) or rating (higher average rating).
- Search for a book by number of awards or award names.
- Search for an author by name, specific book written and number of books.
- Search for the most common genres (that appear in more books) and genres of popular books (that have more votes).
- Search for reviews by text length, user or date.

7 CONCLUSIONS AND FUTURE WORK

All of the planned objectives for the first phase of the project were successfully completed, giving the group a better understanding of all the collected data and of what is needed to develop the search system.

Even though, as explained before, there sometimes were some difficulties in enriching and refining the original data, these were eventually overcome, resulting in a large and coherent dataset that has been thoroughly analyzed and is now ready to be fully integrated in the search engine that is to be built in future work.

REFERENCES

[1] URL: https://docs.python-requests.org/en/latest/ (visited on 11/07/2021).
[2] URL: https://beautiful-soup-4.readthedocs.io/en/latest/ (visited on 11/07/2021).
[3] URL: https://pandas.pydata.org/ (visited on 11/02/2021).
[4] URL: https://spacy.io/universe/project/spacy-langdetect (visited on 11/15/2021).
[5] Goodreads.com. *Goodreads | Meet your next favorite book.* URL: https://www.goodreads.com/ (visited on 11/03/2021).
[6] Kaggle. *Your Machine Learning and Data Science Community.* URL: https://www.kaggle.com/ (visited on 10/27/2021).
[7] Austin Reese. *Goodreads Books - 31 Features.* July 2020. URL: https://www.kaggle.com/austinreese/goodreads-books (visited on 11/02/2021).