Amazon Book Reviews

ANA SOFIA TEIXEIRA, RICARDO NUNES RIBEIRO, SÉRGIO MANUEL CARVALHAIS, and TOMÁS AGANTE MARTINS

ABSTRACT

In an era characterised by an exponential proliferation of data, the significance of information management cannot be overstated. This paper introduces an innovative approach to design and implement an information processing and retrieval system for books based on reader reviews. The primary objective of this endeavour is to enable efficient fulfilment of prospective search tasks, such as locating desired books based on reader opinions, thematic discussed, and genres.

KEYWORDS

Book Reviews, data collection, data source, data preparation.

1 INTRODUCTION

While existing book search systems rely on conventional criteria, our project introduces a novel technical solution. We aim to develop a search engine that enables users to search books based on reader opinions and detailed content descriptions. This article serves as a comprehensive overview of the initial development phase of our project, focusing on the technical aspects. Our goal is to create a technically robust book search system that enhances the user experience by considering reader opinions and detailed content information.

We begin by sourcing and curating relevant data sets, assess data quality and authority, and conduct exploratory data analysis. Additionally, we document our data processing pipeline. Our work involves characterising data set properties, defining the conceptual model for our book data, and meticulously identifying and characterising information needs for the rest of the development of the project.

2 DATA SOURCES

In our quest to obtain data aligned with our research objectives, Kaggle [1] emerged as the ideal source. The data set of interest, 'Amazon Books Reviews' [2], comprises two key files. The first file details 212,404 distinct books, sourced from the Google Books API. The second file contains a treasure trove of Amazon reviews, totalling 3 million entries from May 1996 to July 2014. The book details file encompasses essential information like title, authors, publisher, and user ratings. In contrast, the reviews file provides insights into book titles, pricing, user details, review scores, and more.

What's noteworthy is that this data set is licensed under Creative Commons Zero (CC0), effectively dedicating it to the public domain and allowing unfettered usage and modification in accordance with copyright law. This open data approach encourages collaboration and innovation within the research community.

Authors' address: Ana Sofia Teixeira, up201806629@fe.up.pt; Ricardo Nunes Ribeiro, up202310095@fe.up.pt; Sérgio Manuel Carvalhais, up202007544@fe.up.pt; Tomás Agante Martins, up201704976@fe.up.pt.

3 DATA COLLECTION AND PREPARATION

3.1 Data Preparation Pipeline

For the data collection phase, we first downloaded a .zip folder from the Kaggle website [2]. After having the folder locally on our devices, we unziped it and inside found the two files we required: *Books_ratings.csv* and *books_data.csv*.

From the name of both files, we quickly understood what was contained in them. the first one compiled all the reviews and ratings from the users associated with a given book. We became aware that there were 3 million entries of reviews in this file. The second one was more focused on storing the individual book information (212404 records of books), among them, their authorship and categorisation.

After having collected the data from the data source, we proceeded with some tasks in order to prepare our data for the next phases according to the pipeline depicted in Figure 1.

In the data exploration phase, and through the use of the Python Pandas library, we were able to collect some initial data regarding the data sets. For instance, the column names of both tables; the number of missing values; the number of unique values in the columns (e.g.: there were 10883 unique values in the column 'categories' in *books_data.csv*). Later on, in the Data Cleaning phase, we decided that, in the context of our project, there was no need to store some of the information present in the files. Moreover, the files' size was considerably large, therefore, to facilitate the computing operations in our personal computers, we deleted those columns we deemed unnecessary: "Id", "Price", "User_id", "profileName", "review/helpfulness", "review/score" and "review/time" from *Book_rating.csv*; and "image", "previewLink", "infoLink" and "ratingsCount" from *books_data.csv*. Later on, we renamed the remaining columns in *Books_rating.csv* according to the following scheme: "Title" $\mapsto$ "book_title", "review/summary" $\mapsto$ "summary", "review/text" $\mapsto$ "text"; and "Title" $\mapsto$ "book_title", leaving all the others with the same name. As for *books_data.csv*, we renamed "Title" to "book_title" and left all others unchanged ("description", "authors", "publisher", "publishedDate", 'categories'). In the end of this phase, we wrote the new tables into new .csv files for organisation purposes, being them *reviews.csv* and *books.csv*.

Lastly, using a Jupyter Notebook, we created some plots in order to have a deeper look at the data at hands (cf. 3.3 Data Characterisation). Figure 1 represents a formal overview of the entire process.

3.2 Conceptual Data Domain Model

Having the data in such a way we could better handle and analyse, we were able to come up with a Conceptual Data Model. The model (Figure 2) identifies the entities we consider relevant in the context of our problem as well as their respective attributes. We point out the following:

- *book* entity
  - book_title: stores the name of the book

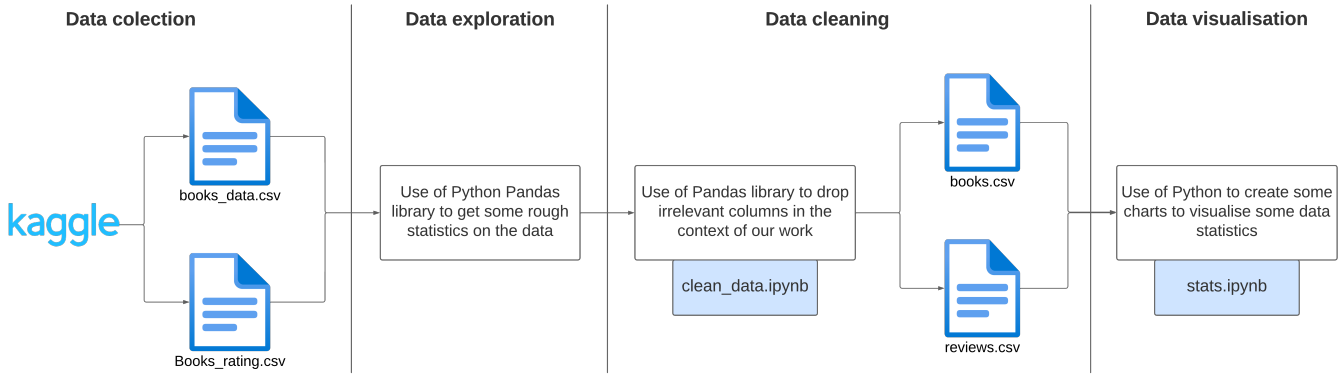

Fig. 1. Data Pipeline

- *review* entity
  - summary: holds a summary/preview of the full review
  - text: stores the full review written by the user
- *author* entity
  - author: stores the name of the author
- *category* entity
  - category: records the name of the category the book falls into
- *publisher* entity
  - publisher: holds the name of the company who responsible for publishing the book

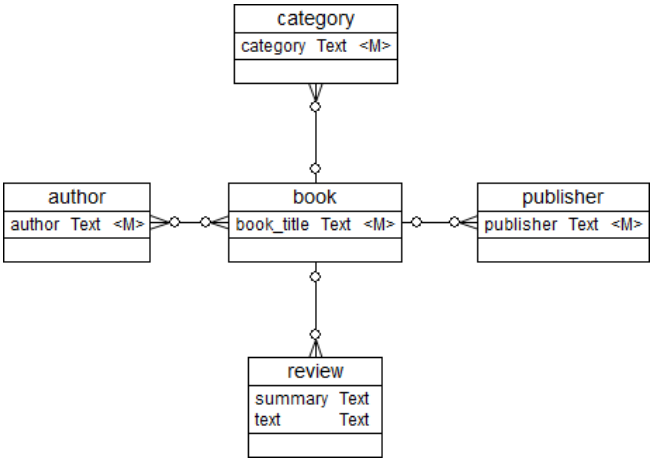

Fig. 2. Conceptual Data Model

4 DATASET CHARACTERISATION

In terms of our collection, the .csv files we obtained were extensive. They contained information regarding full-text reviews with reference to the books reviewed and also every information imaginable about the book, like Amazon internal reference number, no. of pages, the full book, etc. To ease out our task and for the presentation of our topic to be clearer, we produced some charts using the python scripts in our pipeline.

The next goal is for our pipeline to extract the content in a structured way in order for it to fit the conceptual model we idealised for the final search engine and respective database.

Given that the main focus of this project is regarded with user reviews, we decided to better analyse this field in our data set. To do so, we created some charts that reflect the characteristics we consider to be more relevant.

4.1 Lexical Diversity

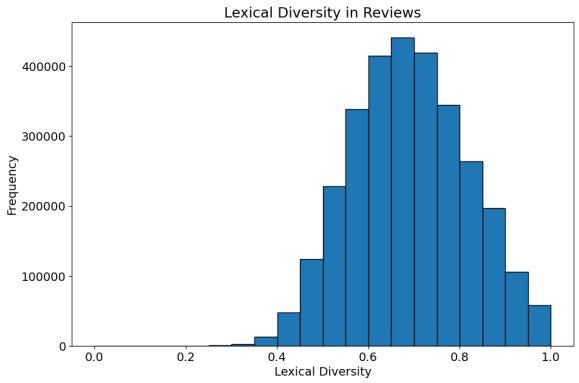

Fig. 3. Reviews' Lexical Diversity Distribution

The chart in Figure 3 gives an overall idea of how many different words reviewers use in their reviews. From the chart, we highlight that more than 1.2 million reviews (around 40% of the whole data set) have a lexical diversity ratio between 65% and 75%, meaning most words in the review text field were only written once.

4.2 Most Common Words

Another useful analysis we found useful was to explore what were the most common words that were written in the whole data set

(Figure 4). In line with out expectations, "book" is, in fact, the most frequent word in the database - more than 5 million repetitions -, followed by "read", "one", "story", and, closing the top five most common words, "like".

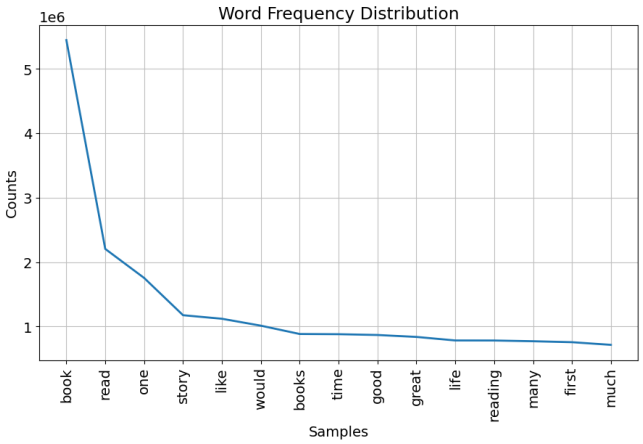

Fig. 4. Most Common Words

4.3 Word Count Distribution in Reviews

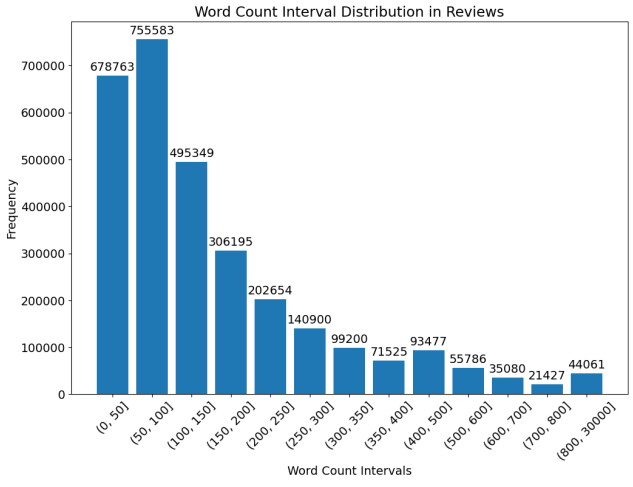

Fig. 5. Word Count Interval Distribution in Reviews

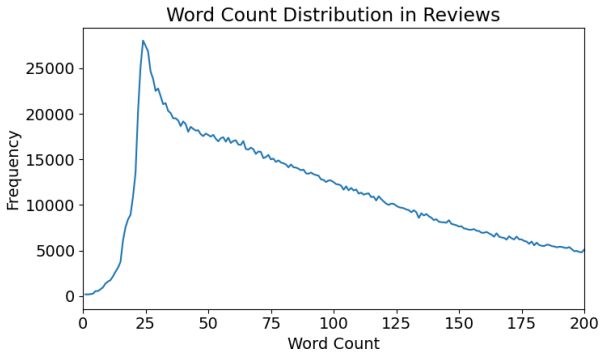

Fig. 6. Word Count Distribution in Reviews

As the Figure 5 chart demonstrates, most reviews contain between 50 and 100 words (around 25% of all reviews). A closer look at the first four intervals (Figure 6) allows us to see that the most frequent number of words in reviews is around 25 words. However, it came to our attention the fact that there is a significant amount of reviews that contain more than 400 words.

4.4 Number of Sentences Distribution in Reviews

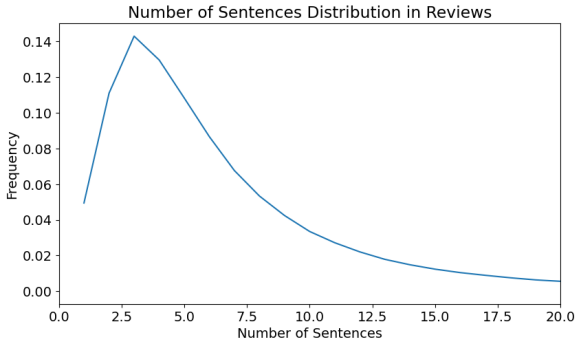

Fig. 7. Number of Sentences Distribution in Reviews

Lastly, in Figure 7, we plotted the distribution of the number of sentences written in the reviews. As the figure shows, around 14% of the reviews are made up of roughly 3 sentences, being this the most common sentence size in the data set.

5 PROSPECTIVE SEARCH TASKS

5.1 Description

The objective of the project is to provide the user with an advanced book search so that they can find the reading they are looking for based on several factors. As such, the user must be able to find books not only by their title, but also through their author, publisher, category, theme, reviews and ratings, all combined. An example of a possible search would be science fiction books that do not address the topic of space.

5.2 Information needs

To perform this search, all the information contained in the files obtained after cleaning the data from Kaggle website will be necessary. Thus, we will have at our disposal technical information about the books (such as title, author, publisher, publication date, category, description) contained in the books.csv file and also reviews and ratings of these in the reviews.csv file.

Based on this, there is a need to convert the tables to SQL and implement some queries. Some examples of these queries are:

- Books by category and author
- Books with certain category without some theme
- Books from some author with rating above some number

6 CONCLUSION

The main goals for the first phase of the project were accomplished successfully given that there is a better understanding of the chosen domain, the already existent data and data sets in it and which ones are relevant for this purpose.

The only setback we had during this milestone was the data preparation using Excel, that was midway changed to Python, other than that, we achieved our end goal of creating a data set with complete and coherent data about book reviews.

As for future work we plan on creating an organised database to help data retrieval and later on to develop our search engine.

REFERENCES

[1] [n.d.] *Datasets*. URL: https://www.kaggle.com/datasets.
[2] Mohamed Bekheet. *Amazon Books Reviews*. URL: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews. (accessed: 29.09.2023).