Data Visualization Project (COM-480)

Studio Bibli

Student's name	SCIPER
Guillaume Parchet	283294
Arnaud Gaudard	269672
Roxane Pangaud	283177

Milestone 1 (8th April, 5pm)

For the project, we wanted to talk about a subject that all of us enjoy. After some discussions and research, we decided to do something about books. Indeed, since books have covers, it can lead to good visualisation.

Dataset

The main data basis of our work will be the "Best Books Ever Dataset". Published in November 2020, it can be found on Zenodo.

The original data has been collected from GoodReads - Best Books Ever and contains 25 variables and 52478 records (books). By default, the dataset contains the following features:

Attributes	Definition	Completeness
bookId	Book Identifier as in goodreads.com	100%
title	Book title	100%
series	Series Name	45%
author	Book's Author	100%
rating	Global goodreads rating	100%
description	Book's description	97%
language	Book's language	93%
isbn	Book's ISBN	92%
genres	Book's genres	91%
characters	Main characters	26%
bookFormat	Type of binding	97%
edition	Type of edition (ex. Anniversary Edition)	9%
pages	Number of pages	96%
publisher	Editorial	93%
publishDate	publication date	98%
firstPublishDate	Publication date of first edition	59%
awards	List of awards	20%
numRatings	Number of total ratings	100%
ratingsByStars	Number of ratings by stars	97%
likedPercent	Derived field, percent of ratings over 2 starts (as in GoodReads)	99%
setting	Story setting	22%
coverImg	URL to cover image	99%
bbeScore	Score in Best Books Ever list	100%
bbeVotes	Number of votes in Best Books Ever list	100%
price	Book's price (extracted from Iberlibro)	73%

After a quick inspection of the data, it seems to be qualitative (low amount of missing data on most important fields, no corrupted data, etc...). It will require some pre-processing though, such as filtering non-english books or grouping them by language and harmonize the publisher names, author names, or settings.

We will also create new variables like the number of words in the title, the length of the series, some key words of the description and the number of awards won.

Problematic

This project has two mains purposes. The first one is to implement a book library and a dynamic top favorite list based on the genre. The aim here is for the visitor to be oriented for his next reading : * With the top favorite list * By clicking on a book that they like in the library * There will be other book recommendations, based on how similar the books' descriptions are

Then, we will visualise the books market dynamics, to understand better if the external parameters of a book (prize, number of pages, length of title, type of cover, publisher, number of sequels/prequels...) influence its popularity, if a genre tends to have more pages, be more expensive or more popular than others, or if a publisher publishes more likable books than others and what genre they publish. It might also be interesting to check if books parts of series display a tendency in ratings (for example if the ratings start falling after n^{th} book published).

We are aware that since we will only consider english books, some similarities in the book descriptions in English, may not be reflected in French (and vice-versa), so the recommendations generated may differ from ones generated with French descriptions (or others).

Exploratory Data Analysis

One can find the detailed procedure of our Exploratory Data Analysis in the Jupyter Notebook.

In short, we mainly used Pandas to find the distribution of some interesting categorical columns (author, language, genre, publisher, setting). We wanted to know if the categories were uniformly spread, or if there were many rarely-used categories and a few common ones. It turns out that the latter is true for every categorical column, though at different amplitudes.

For the publishers, we had noticed that many entries were very very similar but not exactly equal. We call them quasi-duplicates. We tried to do some statistical analysis to show that, but neither the method nor the results were really conclusive. Still, we gave some examples of the issue. These quasi-duplicates are a problem if we want to make a meaningful visualization later on. This worsens the previously mentioned non-uniform spread of values, too.

For the languages, a large majority of books in the dataset were written in English (42661 books), and only a few thousands of books are available in all the other languages. We'll surely select only English books in our visualization, so that it's easier to make similarity analyses between their descriptions.

We tried seeing if the author had a significant correlation with the book's rating, and we found a correlation of about 0. Also, we saw that the large majority of ratings are high. We count only 310 ratings under 3 out of 5, out of the 68'000

available ratings. Their mean is 4.04 and their standard deviation is 0.35, so we will probably rescale them.

Finally, we checked if there were any obvious correlations between the different numerical statistics from the books. The analysis is yet fairly simple and only yields minor insights (summarizing correlation visualization below, more details in the notebook).

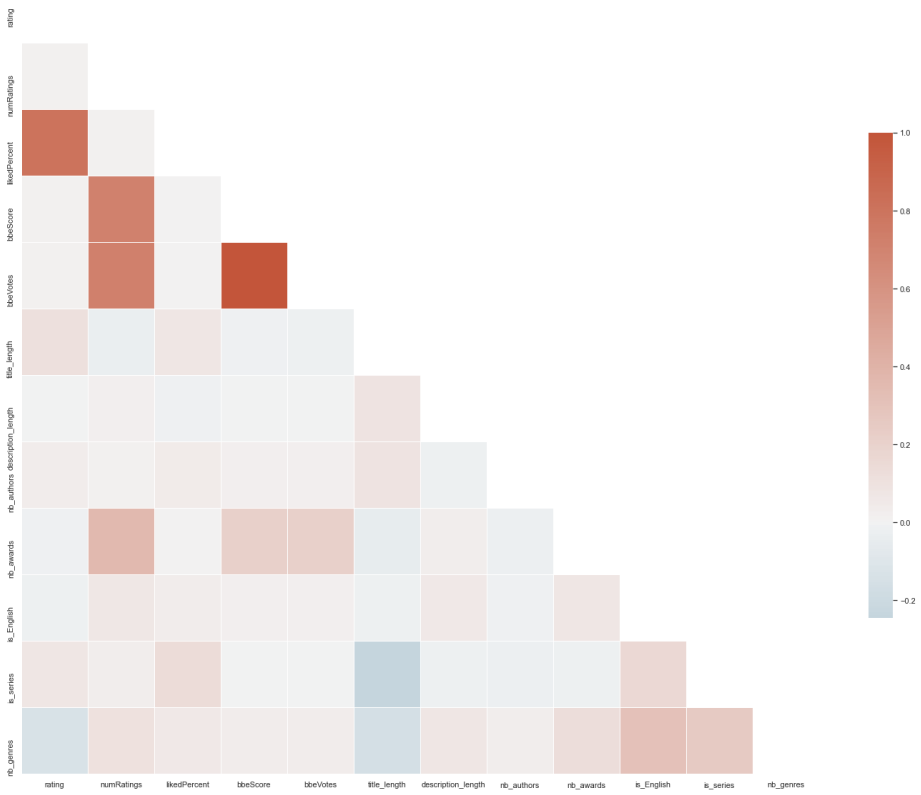

Figure 1: Correlation of numerical columns

With this picture, we see that we do not have strong correlation in first sight. So there might be little relevant to show graphs about how a variable influence another one.

Related work

What has been already done with this data: The "Best Book Dataset" has been used by Stacy Stephanie McDonald for her Master's paper. She used this Dataset to simulate a library dataset as she was developing a Library Management System. You can see her paper here.

Originality of our approach: Our approach is interesting as it focuses on recommendation based on a similarity measure on the books' descriptions. Also, in our website, we don't want to sell books. We want to display them. So the way we will visualize them will be totally different.

Sources of inspiration for the project: Here are some websites that inspire us for our web design. They do not use the same dataset as we.

- What Should I Read Next? - In this website, the search bar and the way it is showing the recommendations in a list are great;
- PlaySuisse - The way the films are shown is instinctive and fluid. It would be very innovative to do so with books. Also, when we put our mouse on a film without clicking on it, we have more infomation that shows, such as the title, the genre and the duration. We can imagine to do the same for the books with the title, the author, the genre and the number of pages.
- Whichbook - In the bottom, under the "Trending books", is the type of visualisation we want. Here also, when you select one of those books, the way it displays informations about that book is what we want to achieve.