Barcelona School of Economics

Data Science Methodology Program

22DM014 - Intro to Natural Language Processing

Analysis of Amazon Books and Reviews

Authors:
Chen, Joshua
El Daou, Mahmoud
Gallegos, Rafael

Professors:
Mueller, Hannes
Garcia, Luis

Contents

List of Figures

List of Tables

1 Introduction

On July 16th, 1995, Amazon opened as an online book retailer, promoting itself as "the world's biggest bookstore". Since then, Amazon has clearly expanded into other ventures but still remains the largest seller of books in the world. Amazon generates $28 billion worldwide every year with a selection of over 32 million books. With Kindle and Audible, Amazon also has majority control of the ebook and audiobook markets as well. While Amazon is no longer primarily known as predominantly a bookseller, they have succeeded in becoming the most dominant bookstore in the world. The world's largest bookseller gives us the opportunity to acquire and use lots of book data to identify book trends and reader preferences.

2 Literature Reviews

3 Dataset

In this project, we explore the Amazon Books Reviews dataset, which can be found on Kaggle at the following link: Amazon Books Reviews Dataset

Contained here are two large data sets contain information about Books and information about Reviews.

Feature	Description
Title	Book Title
Descripe	Description of the book
authors	Name of book authors
image	URL for book cover
previewLink	Link to access this book on Google Books
publisher	Name of the publisher
publishedDate	The date of publish
infoLink	Link to get more information about the book on Google Books
categories	Genres of books
ratingsCount	Averaging rating for the book

(a) Book Dataset

Feature	Description
id	The Id of the Book
Title	Book Title
Price	The Price of the Book
User_id	Id of the user who rates the book
profileName	Name of user who rates the book
helpfulness	Helpfulness rating of the review, e.g., 2/3
score	Rating from 0 to 5 for the book
time	Time of giving the review
summary	The summary of a text review
text	The full text of a review

(b) Review Dataset

There are 214,404 books in the book dataset and 3,000,000 reviews in the review dataset - with information scraped from the Google Books API. Clearly, two very large datasets, we subsetted the datasets and reviews to only the most relevant for our purposes.

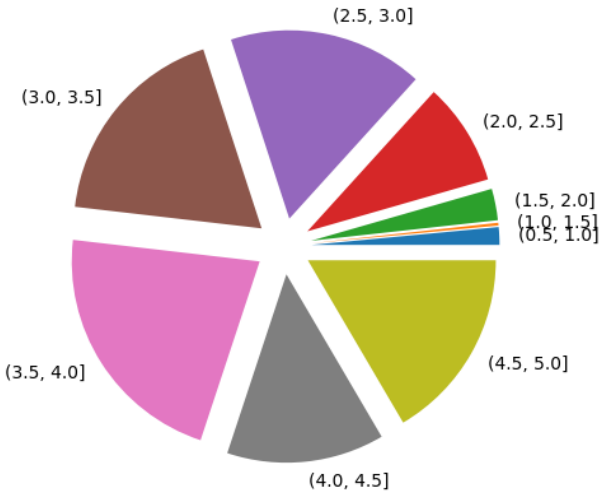

Figure 4.1: Percent of Books by Average Score

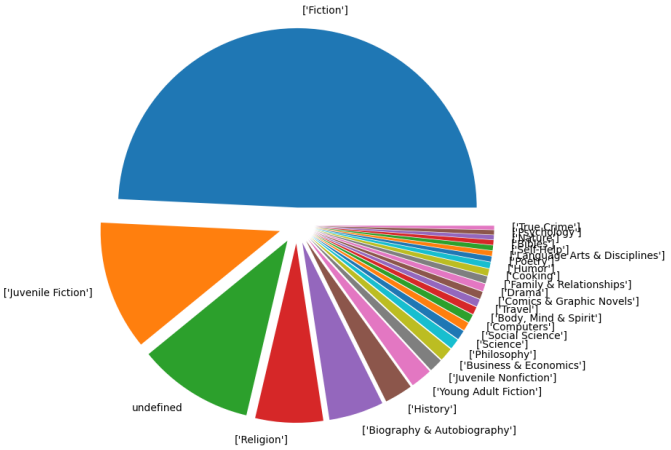

Figure 4.2: Percentage of Books by Genre

First we dropped all the books without any ratings or descriptions as these are irrelevant for our purpose. This reduced the amount of books to 45,127. We then dropped any books with less than 10 reviews, considering them insignificant enough. Furthermore, using the *Langid* package in Python, we removed any books that were identified to be any language other than English. This leaves us with 6,399 books.

We only kept the reviews that were for one of our 6,399 books and used the same method to remove non-English reviews. Because Goodreads allows other users to review other people's reviews, we used this to further subset only the most relevant reviews. Any reviews with less than 10 responses were therefore removed. This brought our review count to 94,573.

4 Exploratory Data Analysis

First, to better understand the dataset we performed some exploratory data analysis. By calculating the average review scores for each book, we find that the largest representation of reviews are within the 3.0 - 4.0 range, with a significant majority of the books receiving a rating above 3.0. The median rating was a 3.6 and the mean was also around 3.6. While this in balance may be reflective of our selection bias of only the most reviewed books (i.e. the most popular books), it is typical for people to tend towards higher reviews as people selectively consume things they tend to enjoy rather than sampling randomly (ex. MovieLens dataset).

By genre we see that the majority of the books are fiction. However, this is only broken down into Juvenile Fiction, Young Adult Fiction, and Fiction whereas non-fiction books are broken down into many topics/subtopics. Genres with less than 15 books were grouped into "Other" as there were a number of hyper-specfiic or uninformative genres. Particularly amusing ones include "Fairy", "Diners", "Bachelors", and "Human-computer Interaction".

In Figure 4.3, we have the counts of books by the number of reviews they received (binned by

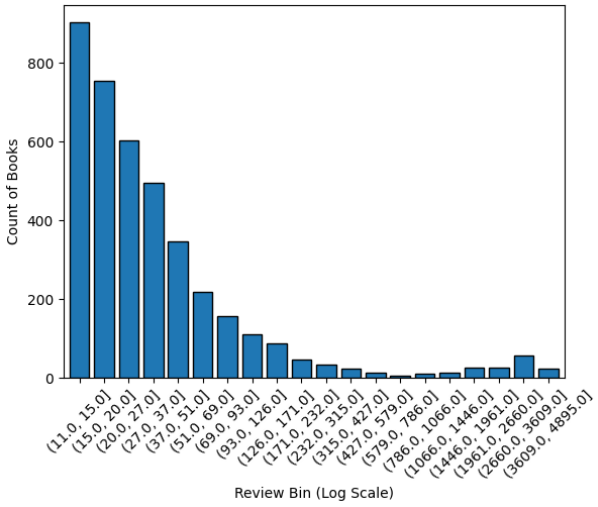

Figure 4.3: Count of Books per Review Bin

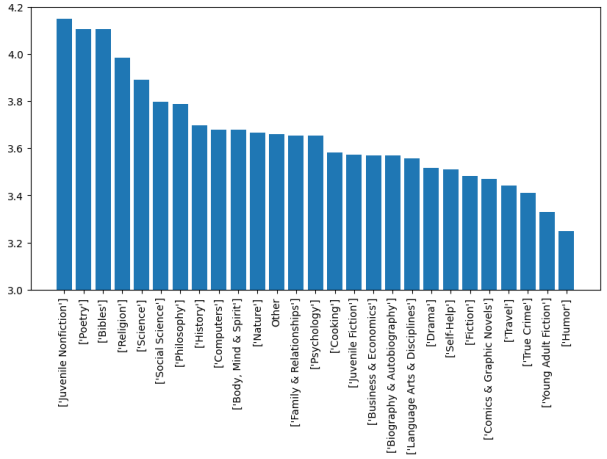

Figure 4.4: Average Review per Genre

a log 10 scale). Here we see that a significant majority of the books have received less than 100 reviews. However, there is a significant tail with the maximum review count reaching 4,895. Figure 4.4 shows the average review breakdown by genre. Interestingly, this chart implies that fiction books tend to have lower reviews than nonfiction books.

5 Descriptions

For each of the books in the dataset, there is a corresponding description. To process this description, we removed all non-alphabet characters, tokenize the text, lemmatize the words, and remove stopwords (with the default *nltk* stopwords with the addition of some of our own). These descriptions range from empty to 813 tokens with a mean of 67.19 and a median of 58. For exploration, here are some wordclouds to explore the descriptions using a document term matrix. What becomes immediately apparent in the Figure 5.1, which is created from 1-3 ngrams with no trimming of terms, is that the most common descriptions have very little to do with the actual books but rather to do with generalizations of the book or author. Particularly the phrase (New York Time Bestselling Author stands out).

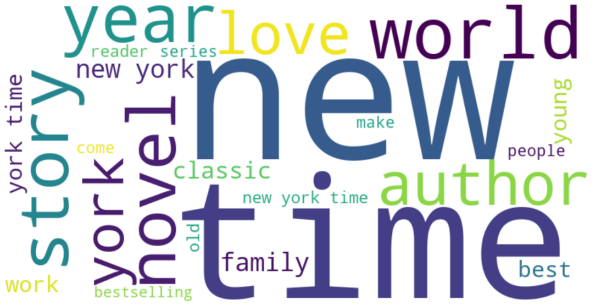

Figure 5.1: ngram - (1,3), no min or max

Figure 5.2: (1,3), min - 100, max - 0.3

This is to be expected because very specific words such as 'Hogwarts', 'Magic', and 'Bach' will not appear in most of the documents. For a more interesting and perhaps informative exploration, we set the minimum document frequency to 100, meaning a term must appear in at least 100 descriptions. This threshold is somewhat arbitrary, as word clouds only include the most frequent terms. Additionally, we set the maximum document frequency to 30 percent, indicating that a term appears in less than 30 percent of the books.

Additionally, here are the wordclouds broken down by the 9 most popular genres. These provide a very interesting cursory perspective on each of these genres. For example, it seems like the philosophy section is dominated by Netizsche's "Thus Spoke Zarathustra" and, for some reason, motorcycles. Most of the other ones follow along expectations, although Juvenile Nonfiction does seem to carry overtly religious undertones.

Figure 5.3: Wordclouds of Term Frequencies by Genre

5.1 New York Times Bestsellers

The prevalence on "New York Times Bestseller" or "New York Times Bestselling Author" then begs the question, do people actually like these books? After a quick linear regression of Review Score to Ratings Count, we see a negative correlation. However, for a more informative look, we ran a quantile regression to see the relationship of Ratings Count with Review Score. Except for the top 10th percentile, it seems that increasing the popularity

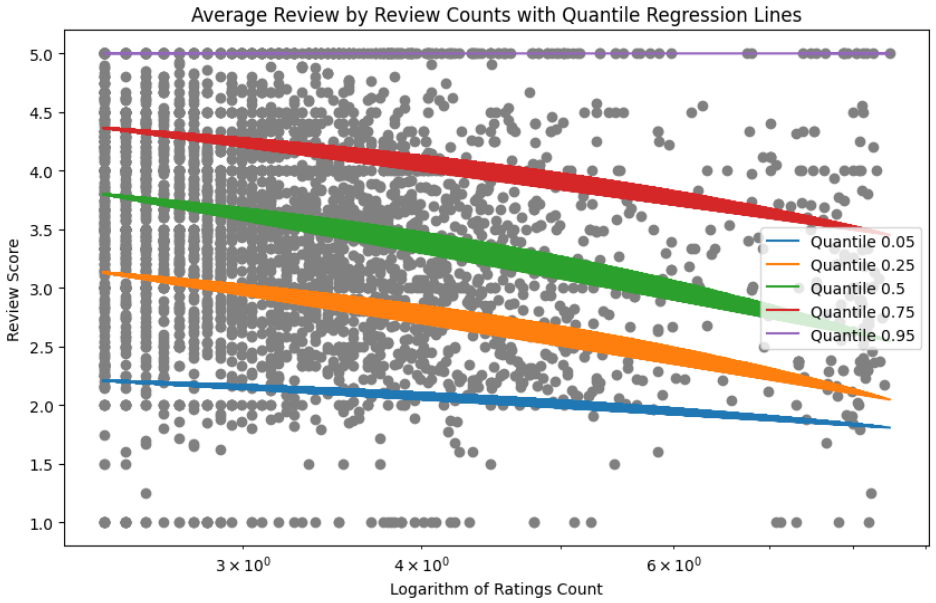

Figure 5.4: Quantile Regression of Average Reviews per Book over Count of Reviews

of books has an adversarial effect on the average review of the book. For a book to come from a bestselling author or New York Times bestseller, it will be interesting to see if this trend holds true. To study this, we will compare the effect of NYT bestseller against other popular books.

Too extract which books are New York Times bestsellers (or written by New York Times Bestselling Authors), I implemented a modified Dictionary Method. I had three dictionaries, one indicating the presence of New York Times, one indicating whether it was bestseller or bestselling, and one with terms referring to the author. The thought behind this is that I will be able to differentiate between descriptions highlighting whether it was the author or the book that was bestselling. Based on this identification system, there are 386 books with 254 of them mentioning the author, and 132 without.

6 Reviews

In the reviews, there are two important text features - the review summary and the full review itself.

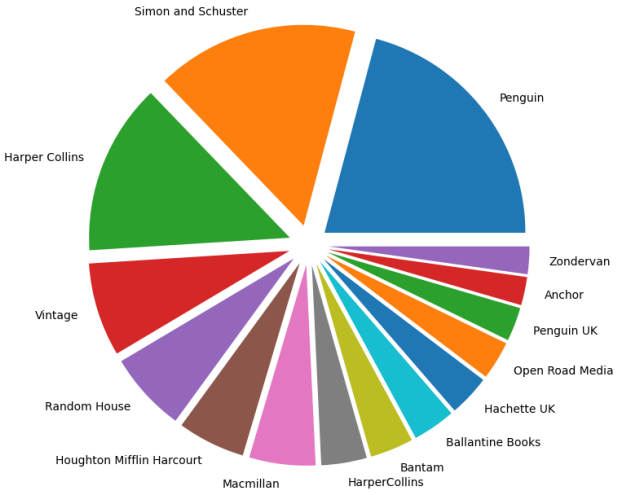

Figure 10.1: Percentage of Books by Publisher

6.1 Sentiment Analysis

7 Predicting Review Score from Text

8 Review Description Alignment

9 Conclusions

10 Other Figures