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.

In 2013, Amazon acquired Goodreads, a social cataloging platform for books. Goodreads now has more than 150 million users and has become a cornerstone (for better or worse) in the book community. However, the partnership between world's largest bookseller and the largest community of readers, gives us the opportunity to acquire and use lots of book data to identify book trends and reader preferences.

2 Literature Reviews

There has been extensive research conducted in the fields of sentiment analysis, prediction and text classification. Per instance, Almjawel et al. (2019) do Sentiment Analysis of Amazon Books' Reviews and explore different visual systems that aids users to get information about many books.

Regarding text classification, determining sentiment polarity stands as a core issue within sentiment analysis and the objective is to ascertain whether a given text segment is positive, neutral or negative. Srujan et al. (2018) examine customer book reviews from Amazon.com, conducting a comparative analysis of various classifiers such as, K-Nearest Neighbours (KNN), Random Forest (RF), Naive Bayes (NB) and a sentiment analysis made with NRC Emotion Lexicon words and term frequency-inverse document frequency (TF–IDF).

Beyond the classification text models appear the regression and prediction models using sentiment analysis. Per instance, Ganu et al. (2009) showed that using textual information results in better general or personalized review score predictions than those derived from the numerical star ratings given by the users. Additional, the text-based recommendation allows users to get recommendations on specific aspect, and soft clustering-based approaches that group users based on their reviewing styles and interest similarities.

Some 'mix' techniques is worked in Asghar (2016) that do Review Rating Prediction as a 5 multi-class classification problem, and build sixteen different prediction models by combining four feature extraction methods, (i) unigrams, (ii) bigrams, (iii) trigrams and (iv) Latent Semantic Indexing, with four machine learning algorithms, (i) logistic regression, (ii) Naive Bayes classi

cation, (iii) perceptrons, and (iv) linear Support Vector Classiffication.

Additionally, Qu et al. (2010) introduce a bag-of-opinions method. In this framework, an opinion within a review is composed of three elements: a root word, associated modifier words from the same sentence, and any negation words. Each opinion is allocated a numerical score, which is determined through ridge regression. For testing with domain-specific reviews, a review's rating is forecasted by summing the scores of all opinions within the review and integrating this with a domain-specific unigram model.

Finally, we use a Quantile Regression defined by Koenker (2005). This regression considers the effect of explanatory variables on the entire conditional distribution of rating. This is particularly useful in understanding the impact of independent variables not just at the center (mean) of the distribution but throughout the distribution. For example, it can show how a change in an independent variable affects the lower, median, and upper parts of the distribution of the dependent variable.

We broadly outline the quantile regresion. Suppose that the random variable Y has cumulative distribution function (CDF) $F_Y(y) = P(Y \leq y)$. The τ-th quantile of Y is defined as $Q_\tau(Y) = \inf\{y : F_Y(y) \geq \tau\}$, where $0 < \tau < 1$ is the quantile level. Besides, if Y is a response variable and $\mathbf{x}$ is a d-dimensional predictor. Let $F_Y(y|\mathbf{x}) = P(Y \leq y|\mathbf{x})$ denote the conditional CDF of Y given $\mathbf{x}$. Then the τ-th conditional quantile of Y is defined as

$$Q_\tau(Y|\mathbf{x}) = \inf\{y : F_Y(y|\mathbf{x}) \geq \tau\}.$$

From the definition of a quantile, we can see that $Q_{0.5}(Y)$ is the median, also we referred to the third quantile, while $Q_{0.2}(Y)$ is our first quantile or the 20th percentile, the second quantile, $Q_{0.4}(Y)$, corresponding to the 40th percentile, the forth, $Q_{0.6}(Y)$, to the 60th percentile and last percentile (fifth percentile), $Q_{0.8}(Y)$, to 80th percentile, respectively.

Then, the linear conditional quantile function $Q_\tau(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_\tau$, is estimated by solving,

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\boldsymbol{\tau})),$$

where $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$, and $\mathbb{I}$ is the indicator function.

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.

There are 214,404 books in the book dataset and 3,000,000 reviews in the review dataset. Clearly, two very large datasets, we subsetted the datasets and reviews to only the most relevant for our purposes.

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

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

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".

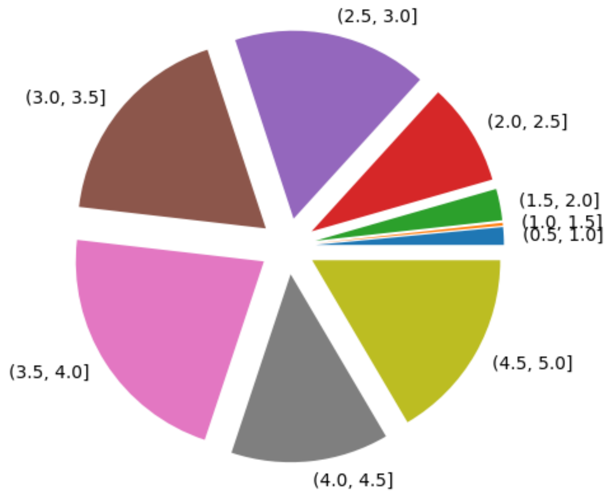

Figure 4.1: Percent of Books by Average Score

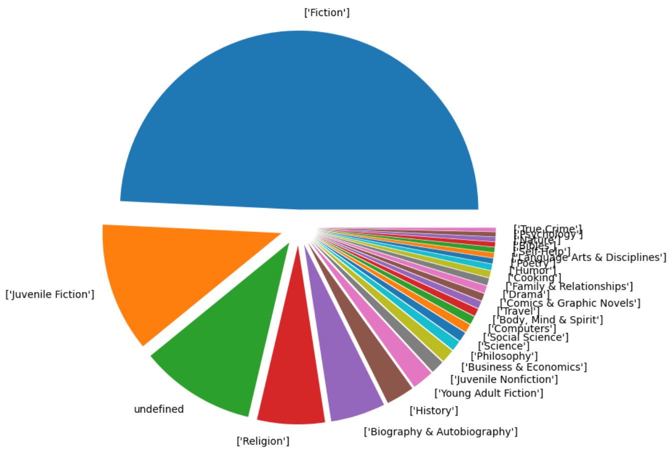

Figure 4.2: Percentage of Books by Genre

We also have the breakdown by publishers. While this was not utilized in this paper, an interesting extension of this project would be to use the description data to see if there are any significant differences among how different publishers market their books.

Finally, we have the counts of books by the number of reviews they received (binned by 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.

4.1 Descriptions

4.1.1 New York Times Bestsellers

4.2 Reviews

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

4.2.1 Sentiment Analysis

5 Predicting Review Score from Text

6 Review Description Alignment

7 Conclusions

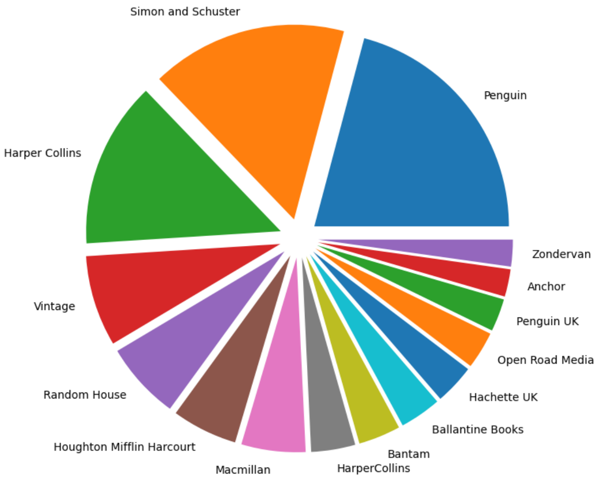

Figure 4.3: Percentage of Books by Publisher

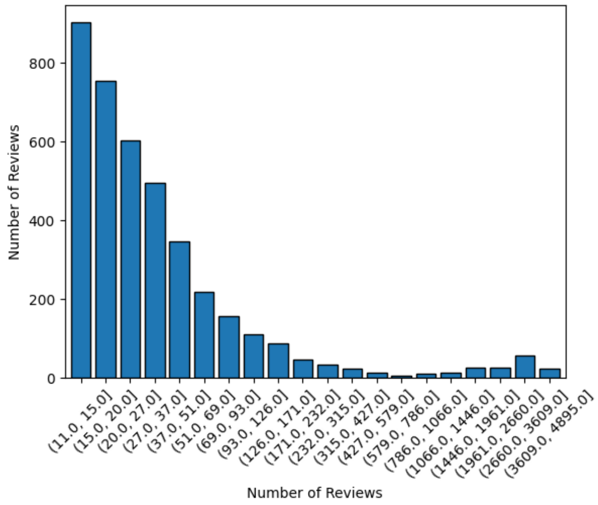

Figure 4.4: Percentage of Books by Publisher

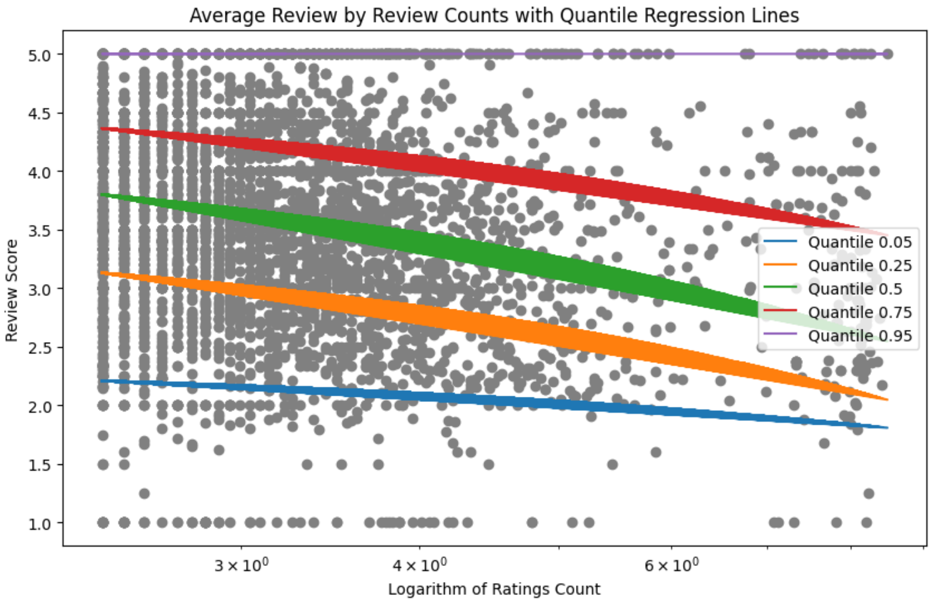

Figure 4.5: Quantile Regression of Average Reviews per Book over Count of Reviews

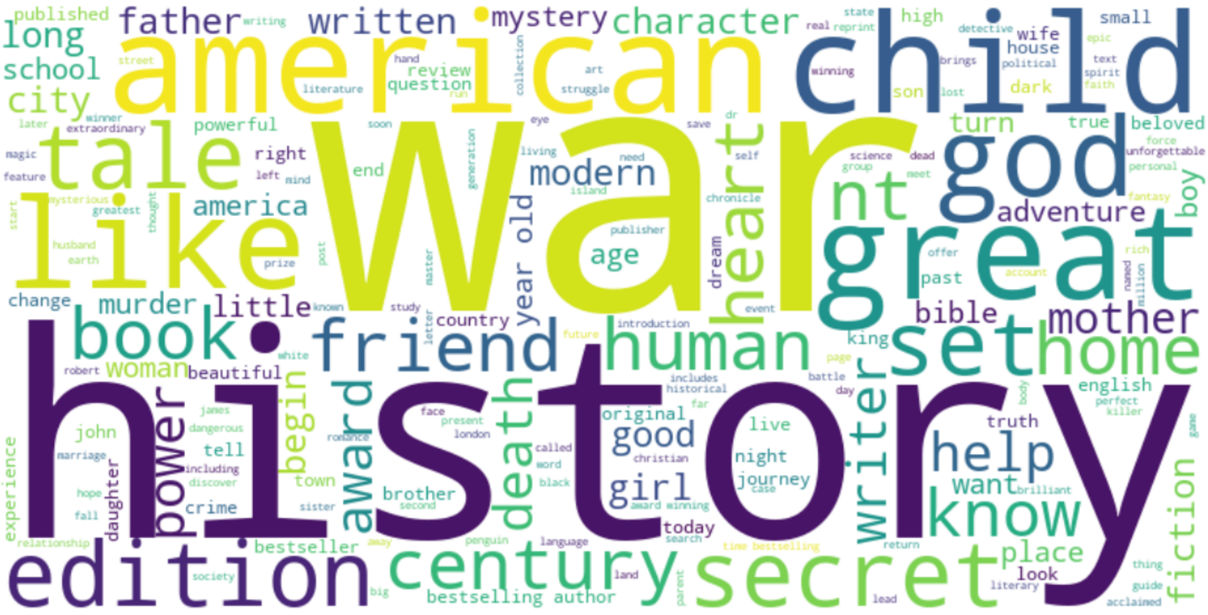

Figure 4.6: Bag of Words - Book Descriptions - MinDF = 20, MaxDF = 0.01

References

1. A. Almjawel, S. Bayoumi, D. Alshehri, S. Alzahrani, and M. Alotaibi (2019). "Sentiment Analysis and Visualization of Amazon Books' Reviews," in *Proceedings of the 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, Riyadh, Saudi Arabia, pp. 1-6.

2. N. Asghar (2016). "Yelp Dataset Challenge: Review Rating Prediction."

3. G. Ganu, N. Elhadad, and A. Marian (2009). "Beyond the Stars: Improving Rating Predictions using Review Text Content," in *Proceedings of the International Workshop on the Web and Databases.*

4. R. W. Koenker (2005). *Quantile Regression.* Cambridge University Press, Cambridge, UK.

5. L. Qu, G. Ifrim, and G. Weikum (2010). "The bag-of-opinions method for review rating prediction from sparse text patterns," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, Association for Computational Linguistics, USA, pp. 913–921.

6. K. S. Srujan, S. S. Nikhil, H. Raghav Rao, K. Karthik, B. Harish, and H. Keerthi Kumar (2018). "Classification of Amazon Book Reviews Based on Sentiment Analysis," in V. Bhateja, B. Nguyen, N. Nguyen, S. Satapathy, and D. N. Le, eds., *Information Systems Design and Intelligent Applications*, vol. 672 of *Advances in Intelligent Systems and Computing*, Springer, Singapore.