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** 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** 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** 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** 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** 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**. 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.

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

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

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

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

Lizhen Qu, Georgiana Ifrim, and Gerhard 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, 913–921.

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

3 Dataset

4 Exploratory Data Analysis

4.1 Descriptions

4.1.1 New York Times Bestsellers

4.2 Reviews

4.2.1 Sentiment Analysis

5 Predicting Review Score from Text

6 Review Description Alignment

7 Conclusions