NLP Assignment: Kindle Store Reviews Analysis

Wenqi Jiang

1/28/2022

Introduction

The Kindle Store is an online e-book e-commerce store operated by Amazon as part of its retail website and can be accessed from any Amazon Kindle, Fire tablet or Kindle mobile app. At the launch of the Kindle in November 2007, the store had more than 88,000 digital titles available in the U.S. store. This number increased to more than 275,000 by late 2008, and exceeded 765,000 by August 2011. In July 2014, there were over 2.7 million titles available. As of March 2018 there are over six million titles available in the U.S.

One of the store's novelties was one-click buying, which allowed consumers to rapidly purchase an e-book. The Kindle Store employs a recommendation system that examines the user's purchase history, browsing history, and reading behavior before recommending stuff it believes the user would enjoy. As a result, it is critical to analyze user reviews in order to promote books and enhance sales.

Data Description

This 5-core dataset of product reviews from Amazon Kindle Store category from May 1996 - July 2014. Contains total of 982619 entries. Each reviewer has at least 5 reviews and each product has at least 5 reviews in this dataset. Each observation contains 10 attributes that include information such as reviewer ID, product ID etc. The table below provides brief descriptions of each attribute and its type.

Variable	Type	Description
X	Long	Index
reviewerID	String	ID of the reviewer, e.g. A2SUAM1J3GNN3B
reviewerName	String	name of the reviewer
asin	String	ID of the product, e.g. 0000013714
helpful	Array	helpfulness rating of the review, e.g. 2/3
reviewText	String	text of the review
overall	Integer	rating of the product
summary	String	summary of the review
unixReviewTime	Timestamp	time of the review (unix time)
reviewTime	Date	time of the review (raw)

This dataset is taken from Amazon product data, Julian McAuley, UCSD website. http://jmcauley.ucsd.edu/data/amazon/

Experiments

0. Load Packages

```
library("tm") # for text mining
library(tidyverse) # for data processing
library(plyr)
library("SnowballC") # for text stemming, reduces words to their root form
library("syuzhet") # for sentiment analysis
library(udpipe) # tokenization, Parts of Speech Tagging, Lemmatization and Dependency
library(lattice) # for bar plot etc.
```

1. Data Pre-processing

a. Get samples from raw data

Since there are nearly one million data, limited by local computing power and resources, it was selected a sample of one percent of the total data base on the `overall` rating.

b. Check the most frequent, identical review texts

```
## # A tibble: 9,826 x 3
##    reviewText                                                  n_reviews     pct
##    <chr>                                                           <int>   <dbl>
##  1 "-- Italian review below-- Rating: 4,5/5We had left Ciardi~         1 1.02e-4
##  2 ".  I enjoy reading about Mail Order Brides and this one w~         1 1.02e-4
##  3 "....but what we have here is an unbiased look at J.Osteen~         1 1.02e-4
##  4 "...and inside I found the most interesting articles about~         1 1.02e-4
##  5 "...and not very sexy.  I couldn't even get lost in the ri~         1 1.02e-4
##  6 "...and now I'm glad I don't have to re-read the other 4 n~         1 1.02e-4
##  7 "...if like me you don't want to have to recharge your Kin~         1 1.02e-4
##  8 "...it may have been my favorite of the series! After read~         1 1.02e-4
##  9 "...that's a story all by itself. This baby's a novella - ~         1 1.02e-4
## 10 "...they were Edaion's slaves now.Leocardo Reyes finds him~         1 1.02e-4
## # ... with 9,816 more rows
```

From the output, there are no duplicate reviews, or empty reviews, which means that the raw data has been processed. So those reviews just need to be normalized in the next step.

Reviews Analysis

1. The top 30 words

We build a Term Document Matrix(TDM) to represents documents vectors in matrix form in which the rows correspond to the terms in the document, columns correspond to the documents in the corpus and cells correspond to the weights of the terms. So we can easily count out the top 30 most frequently used words in reviews.

Conclusion

By analyzing the reviews as described above, we were able to have a good understanding of basic data processing, as well as text normalization (including the removal of numbers, punctuation, stop words, etc.), and then calculate a matrix to count word frequencies, in addition to analyzing the sentiment of reviews, as well as keywords, by using existing NLP analysis tools (`tm`, `SnowballC`, `syuzhet` and `udpipe` etc.). Among the such reviews, “trust”, “anticipation”, “joy” are the most common sentiments, and ” family members”, “science fiction”, and ” small town” are the most common keywords, which can be easily seen in the plot.