Data and Sentiment Analysis of Amazon Book Reviews

Authors: Suman Chauhan, Yougender Chauhan, Nagender Chauhan, Viswanth Reddy Sama
Department of Information Systems, California State University, Los Angeles
CIS5200-03 System Analysis and Design
schauha7@calstatela.edu, ychauha4@calstatela.edu, nchauha5@calstatela.edu, vsama@calstatela.edu

Abstract: This project uses the Dataset of a well-known e-commerce website Amazon.com to analyze the Sentiment of the review text and summary text of the reviews (posted by customers) of books posted on Amazon. The primary focus of this project is to analyze the reviews of the books that are posted on Amazon.com using data analysis techniques such as Sentiment Analysis, Context N gram, data visualization etc. This project provides a clear flow of handling big data files, data engineering and analysis processes using Hadoop and Hive/Beeline. In addition to that, visualization of this analyzed data is conducted using Excel and Tableau, depicting visuals such as Line charts, bar charts, pie charts, hybrid charts etc.

1. Introduction

This project uses Hadoop and Hive/Beeline to keep and process Amazon book review dataset. Amazon.com is originally found by Jeff Bezos in 1994 and has grown rapidly to become one of the most successful e-commerce businesses in the world. With a rapid rate, Amazon.com has expanded in the world and has become one of the most popular retailing websites in the world. The success is mainly due to its customer friendly website interface and innovative tools that aid the customers such as providing lists of best sellers, popular books, and the recommendation system. Reviews are generated in the corresponding product when the customers leave their feedback and rating on the website.

The dataset is consisting of two files Books_rating.csv & books_data.csv Books_rating.csv has information about 3M book reviews for 212404 unique book and users who gives these reviews for each book. This file also consists of columns such as Id, Title, price, profilename, review/summary, review/text, review/helpfulness, review/score etc. Books_data.csv has details of 212404 books such as genres, authors, cover, description etc. with columns consisting of such as Title, description, authors, publisher, categories etc.

When a user posts a review on amazon.com, they have the option to post the review text and summary text. Review text, as the name suggests is an elaborate review typically ranging from 1-2 paragraphs whereas review summary text is a crisp description ranging between 1-3 sentences.

2. Related Work

Although Amazom.com is a popular e-commerce platform, there are quite a few works publicly available based on data from Amazon.com dataset. One such project previously performed by Amrit Pal Singh and Gurvinder Singh is “Analysis of Amazon Product Reviews using Big Data – Apache Pig tool”[1].

The aim of their work was to compare ratings of the products during the two lifespans i.e., between 1996-2011 and 2012-2014 and let the user make decision about a product by comparing the ratings. In their project, different categories of datasets were analyzed. Another aspect of their project was to observe how ‘Amazon’ analyses the reviews on various products. Using the review generated information; Amazon can decide whether to continue listing a product or to remove it from its store [1].

One of the other works “Exploratory Data Analysis of Amazon.com Book Reviews” by Timothy Wong was to explore whether earlier reviews receive more votes and favorable votes over time and other related aspects. Th Project’s primary goal was to observe whether earlier reviews tend to receive higher helpfulness ratings because of the duration of the review, instead of the review’s content. Also, it explained the nature of the dataset using summary statistics and exploratory data analysis; in particular, the project was focus on perspectives that are related to favorable votes and total votes [2].

3. Specifications

The dataset comprises of the dataset is consisting of two files Books_rating.csv & books_data.csv. To perform Sentiment Analysis, we used a predefined dictionary. The size of dataset is 2.9GB. Amazon review Dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

Table 1 shows files and size of the files from dataset.

Table 1 Data Specification

Data Set	**Size (Total 2.9GB)**
Books_rating.csv	2.8 GB
books_data.csv	181.35 MB

The below table show the specification for Oracle cluster and Hadoop specifications

Table 2 H/W Specification

Number of nodes	5 (2 master & 3 worker)
OCPUs	8
CPU speed	1995.309 M Hz, 4 core CPU
Memory	390.7 GB

4. Architecture

The Project architecture is illustrated below in (Figure 1). The Data Source being the first phase, we downloaded the dataset from Kaggle. Then the second phase was Data Processing and Analysis for which we used Hive/beeline,

Hadoop HDFS. The third phase is more about the downloading of analyzed data, for which we used the SCP Command, get cand bash terminal. Finally, for the Data Visualization we used tools such as Excel and Tableau.

Figure 1 – Architecture Diagram

5. Implementation Flowchart

To begin with, the raw dataset, which has both files Books_rating.csv & books_data.csv, was downloaded from a trusted source (Kaggle). To perform Sentiment Analysis a dictionary is required and thus we downloaded the Predefined Dictionary using the wget command. The whole process of date manipulation is shown in the below flowchart (Figure 2).

There are three data logs in csv format (Books_rating.csv, books_data.csv & dictionary.csv) that were uploaded to the Hadoop File System using the Put command. After that, HiveQL/Beeline was used as querying language to create the tables' schema, clean data, create a summary table and export the results. Then we implemented the Data analysis techniques such as Sentiment Analysis, N gram and Context N gram etc. Finally, the analyzed file was downloaded using the SCP command which was used in Excel and Tableau for visualizations.

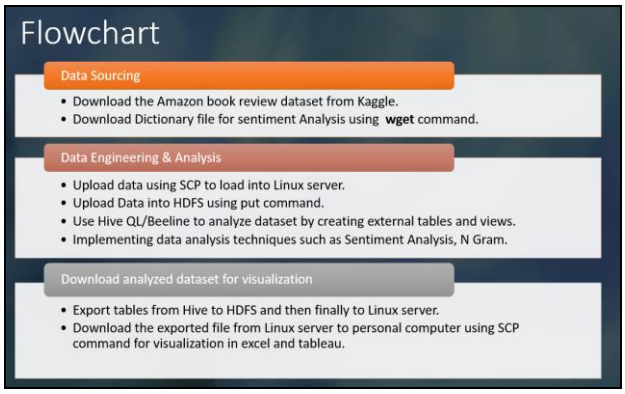

Figure 2 - Implementation Flowchart

6. Data Engineering

For our data set, basic data cleaning was required. For which used the functions such as REGEXP_EXTRACT(), CAST(), REGEX_REPLACE(), from_unixtime(), tblproperties('skip.header.line.com') etc.

After uploading the dataset to HDFS we created the following two tables using HIVE QL. The two tables are as follows:
Books_data (Figure 3) and Books_Rating(Figure 4)

col_name	data_type
title	string
description	string
authors	string
image	string
preview	string
publisher	string
publish_date	bigint
info_link	string
categories	string
ratings_count	int

Figure 3 – Books_data Table Description

col_name	data_type
id	bigint
title	string
price	string
user_id	string
profile_name	string
r_helpfulness	string
r_score	int
r_time	bigint
r_summary	string
r_review	string

Figure 4 – Books_ratings Table Description

7. Analysis and Visualization

Once the data was cleaned, we performed the data analysis techniques such as N gram, context N gram etc.

For the most reviewed movie, The Hobbit, we performed trigram to find most 3 frequent words in review column. The trigram is showed as below (Figure 5)

trigram
{"ngram":["lord","of","the"],"estfrequency":2670.0}
{"ngram":["of","the","rings"],"estfrequency":2457.0}
{"ngram":["the","lord","of"],"estfrequency":1871.0}
{"ngram":["the","hobbit","is"],"estfrequency":1693.0}
{"ngram":["read","this","book"],"estfrequency":1257.0}
{"ngram":["this","book","is"],"estfrequency":1020.0}
{"ngram":["read","the","hobbit"],"estfrequency":939.0}
{"ngram":["one","of","the"],"estfrequency":795.0}
{"ngram":["hobbit","is","a"],"estfrequency":734.0}
{"ngram":["this","is","a"],"estfrequency":685.0}

10 rows selected (45.684 seconds)

Figure 5 – Trigram on The Hobbit book reviews

After that, we performed a context N-gram(fivegram), with the first 3 words being, “this book is” and found that next two consecutive words were as follows (Figure 6)

```
+-------------------------------------------------+
|                    fivegram                     |
+-------------------------------------------------+
| {"ngram":["about","a"],"estfrequency":50.0}     |
| {"ngram":["one","of"],"estfrequency":35.0}      |
| {"ngram":["a","must"],"estfrequency":24.0}      |
| {"ngram":["a","classic"],"estfrequency":20.0}   |
| {"ngram":["a","great"],"estfrequency":20.0}     |
| {"ngram":["the","best"],"estfrequency":20.0}    |
| {"ngram":["a","true"],"estfrequency":19.0}      |
| {"ngram":["a","wonderful"],"estfrequency":16.0} |
| {"ngram":["amazing","i"],"estfrequency":16.0}   |
| {"ngram":["a","timeless"],"estfrequency":15.0}  |
+-------------------------------------------------+
10 rows selected (16.923 seconds)
```

Figure 6 – Fivegram on The Hobbit book reviews

To visualize the analyzed data, we used both Excel and Tableau. In this paper we will go through all the key insights that we performed as part of this project.

7.1 Visualization in Excel

For the first insight, we created a table named “top10books” to get top 10 rated books (i.e., books with the greatest number of reviews), by their average rating and ordered by the count of reviews. With this analysis being performed, we figured that the Hobbit was most reviewed book.

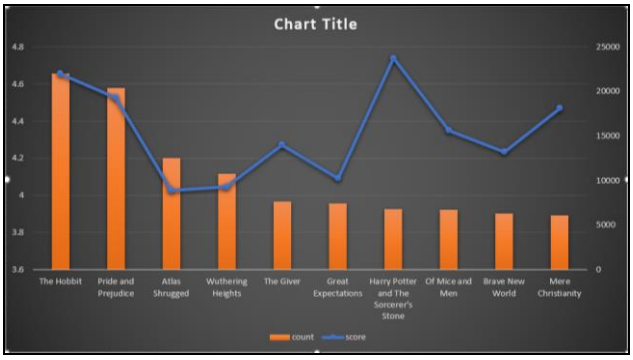

Figure 7 – Top ten reviewed books based on the ratings [Clustered column - line chart]

For the second insight, we visualized the data for Least rated book based on avg score (i.e., books with least star ratings) and create a view with title, average score and count of number of reviews for each book, we query this view to get most reviewed and least rated for more than 500 reviews

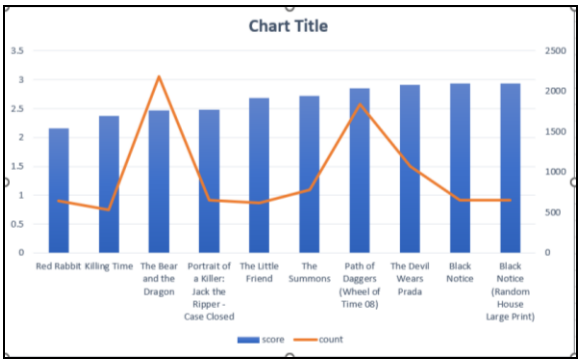

Figure 8 – Least rated books based on the avg score [Clustered column - line chart]

For the third insight, we visulaized the analyzed data for Top 10 categories of books by count

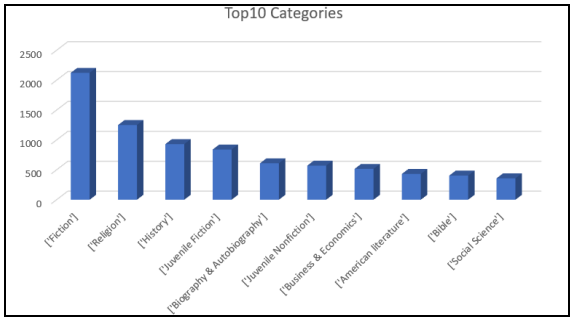

Figure 9 – Top 10 categories of books [3D column chart]

7.2 Visualization in Tableau.

For this insight, we visualized review sentiment and summary sentiment on title wherein we found that there is a deviation in review sentiment and summary sentiment for the same title.

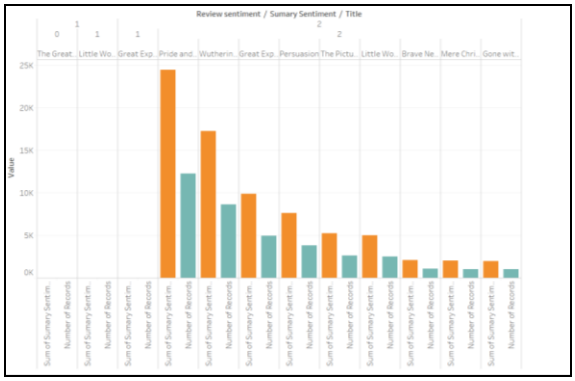

Figure 10 –review vs summary sentiment [side by side bar chart]

For this insight, we created a treemaps in tableau for title, year of review and score for that particular year

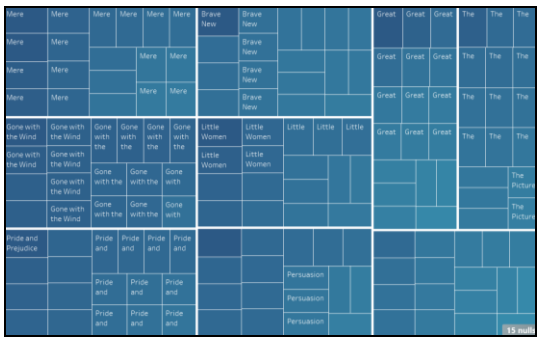

Figure 11 – Average review score based on each year

For this is the insight, we visualized review sentiment and summary sentiment on title and score wherein we categorized the sentiment of review and summary based on review score i.e. (1-5) for top rated books. In this example, we can see wuthering Heights have about 18k positive review sentiment and only 9k positive summary sentiment with rating score being 5.

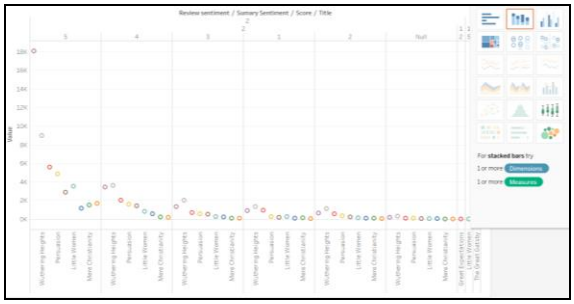

Figure 12 – Reviews Vs Summary sentiment based on review score [side by side circle chart]

Limitation of this analysis is that we found for the same review there were deviation in Review text vs Summary text Sentiment. It is highly unlikely that a user posts a positive review text, and the summary text is negative. Thus, the below pie chart (Figure 13) depicts the limitation of the algorithm used to perform the sentiment analysis on our dataset.

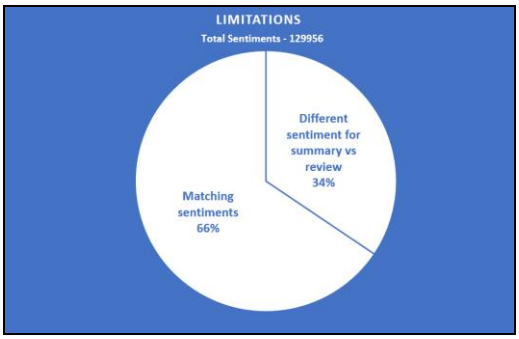

Figure 13 – Matching Sentiment Vs. Different Sentiments [Pie chart]

8. Conclusion

With the analysis performed on the amazon book review dataset we could conclude that

- The Hobbit was the most reviewed book
- Red Rabbit was least rated book
- Fiction was top category of books
- Performed N gram on the hobbit book
- Deviation of Summary Vs Review text

9. GitHub Link

GitHub Link:
https://github.com/Chauhan67/Amazonbookreviews

References

[1] Related Work
https://www.mecs-press.org/ijieeb/ijieeb-v11-n1/IJIEEB-V11-N1-2.pdf
[2] Related Work
https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Timothy.Thesis.pdf
[3] Data Source
https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews?select=Books_rating.csv
[4] Dictionary Link
https://github.com/dalgual/aidatasci/raw/master/data/bigdata/dictionary.tsv
[5] Lab tutorials used:

- labSentimentAnalysisTextNgrams
- labTwitterSentimentAnalysisLab
- labTableau_oracle_v2
- lab2HiveSensorDataAnalysisLab_aws