Amazon Reviews

Eploratory analysis using Hadoop

Vinyas Kaushik Tumakunte Raghavendrarao | Engineering Big Data| December 15, 2018

Contents

[Summary 2](#_Toc532675800)

[Dataset 2](#_Toc532675801)

[Analysis performed 3](#_Toc532675802)

[1. Put Merge and Random Filtering Pattern 3](#_Toc532675803)

[2. Average of Reviews per product 3](#_Toc532675804)

[3. Average of reviews per product category 3](#_Toc532675805)

[4. Count of reviews per product 3](#_Toc532675806)

[5. Count of reviews per product category 3](#_Toc532675807)

[6. Standard deviation and Median of reviews per product 3](#_Toc532675808)

[7. Standard deviation and Median of reviews per product category 3](#_Toc532675809)

[8. Filtering based on any field 3](#_Toc532675810)

[9. Top products by Year 3](#_Toc532675811)

[10. Top 5 products by Year 3](#_Toc532675812)

[Visualization 4](#_Toc532675813)

# Summary

Amazon Customer Reviews (a.k.a. Product Reviews) is one of Amazon’s iconic products. In a period of over two decades since the first review in 1995, millions of Amazon customers have contributed over a hundred million reviews to express opinions and describe their experiences regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for academic researchers in the fields of Natural Language Processing (NLP), Information Retrieval (IR), and Machine Learning (ML), amongst others. Accordingly, we are releasing this data to further research in multiple disciplines related to understanding customer product experiences. Specifically, this dataset was constructed to represent a sample of customer evaluations and opinions, variation in the perception of a product across geographical regions, and promotional intent or bias in reviews.

# Dataset

**Link to dataset**: https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

The dataset contains the customer review text with accompanying metadata, consisting of three major components:

1. A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to facilitate study into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products at scale. (130M+ customer reviews)
2. A collection of reviews about products in multiple languages from different Amazon marketplaces, intended to facilitate analysis of customers’ perception of the same products and wider consumer preferences across languages and countries. (200K+ customer reviews in 5 countries)
3. A collection of reviews that have been identified as non-compliant with respect to Amazon policies. This is intended to provide a reference dataset for research on detecting promotional or biased reviews. (several thousand customer reviews). This part of the dataset is distributed separately and is available upon request – please contact the email address below if you are interested in obtaining this dataset.

This dataset contains the following columns:

* **marketplace - 2 letter country code of the marketplace where the review was written.**
* **customer\_id - Random identifier that can be used to aggregate reviews written by a single author.**
* **review\_id - The unique ID of the review.**
* **product\_id - The unique Product ID the review pertains to. In the multilingual dataset the reviews for the same product in different countries can be grouped by the same product\_id.**
* **product\_parent - Random identifier that can be used to aggregate reviews for the same product.**
* **product\_title - Title of the product.**
* **product\_category - Broad product category that can be used to group reviews (also used to group the dataset into coherent parts).**
* **star\_rating - The 1-5 star rating of the review.**
* **helpful\_votes - Number of helpful votes.**
* **total\_votes - Number of total votes the review received.**
* **vine - Review was written as part of the Vine program.**
* **verified\_purchase - The review is on a verified purchase.**
* **review\_headline - The title of the review.**
* **review\_body - The review text.**
* **review\_date - The date the review was written.**

# Analysis performed

### Put Merge and Random Filtering Pattern

Combined different datasets from the link which contains data from different regions or different categories of products and used random filtering pattern to create a data set which contained around 6Million randomly selected data from the above datasets.

### Average of Reviews per product

### Average of reviews per product category

### Count of reviews per product

### Count of reviews per product category

### Standard deviation and Median of reviews per product

### Standard deviation and Median of reviews per product category

### Filtering based on any field

### Top products by Year

### Top 5 products by Year

1. Sentiment Analysis of the Reviews based on NAÏVE BAYES using MAHOUT
2. Recommendation System- Item based recommendation using MAHOUT
3. Pig Analysis – Top 25 Products by year
4. Pig Analysis – Top 10 products based on reviews
5. Pig Analysis – Joining based on Products
6. Pig Analysis – Filter and Merged the data

# Visualization

All the above analysis are visualized using Tableau.















