Milestone 1

*Anak Agung Ngurah Bagus Trihatmaja, Anupam Gupta, Ayush Bandari, Xiaomeng Peng*

## Abstract:

In the milestone 1, we preprocessed the data and did exploratory data analysis for Amazon Fine Food Reviews [1]. The purpose of the analysis is to find out the potential prior knowledge of the data and hope it can help for future unsupervised analysis.

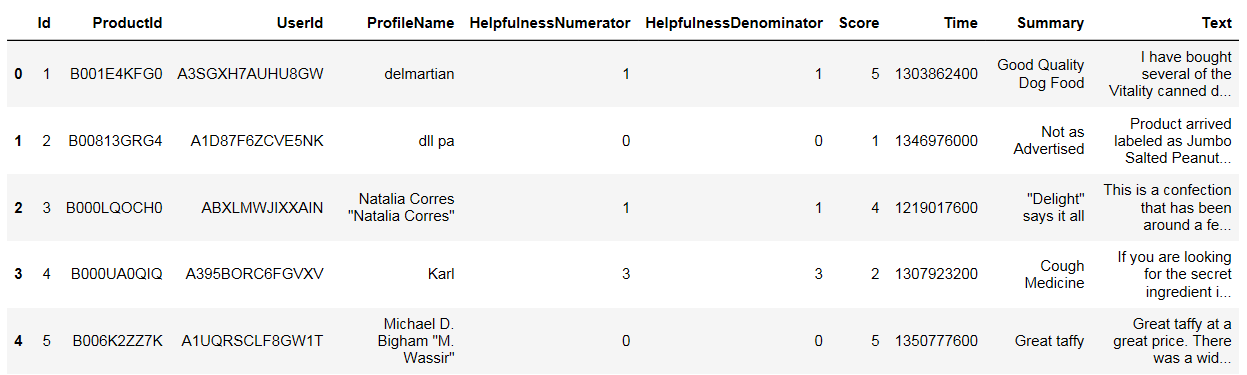
## Dataset Overview:

The dataset has total 568,454 reviews collected from Amazon fine food category from 1999 to 2012. For each review, it has following attributes:

Table 1: Attribute of reviews

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Id | Row Id |
| Unique identifier for the product | Unique identifier for the product |
| User Id | Unique identifier for the product |
| Profile Name | Profile name of the user |
| Helpfulness Numerator | Number of users who found the review helpful |
| Helpfulness Denominator | Number of users who voted for whether they found the review helpful |
| Score | Rating between 1 and 5 |
| Time | Timestamp for the review |
| Summary | Brief summary of the review |
| Text | Text of the review |

The reviews are illustrated in Figure 1.

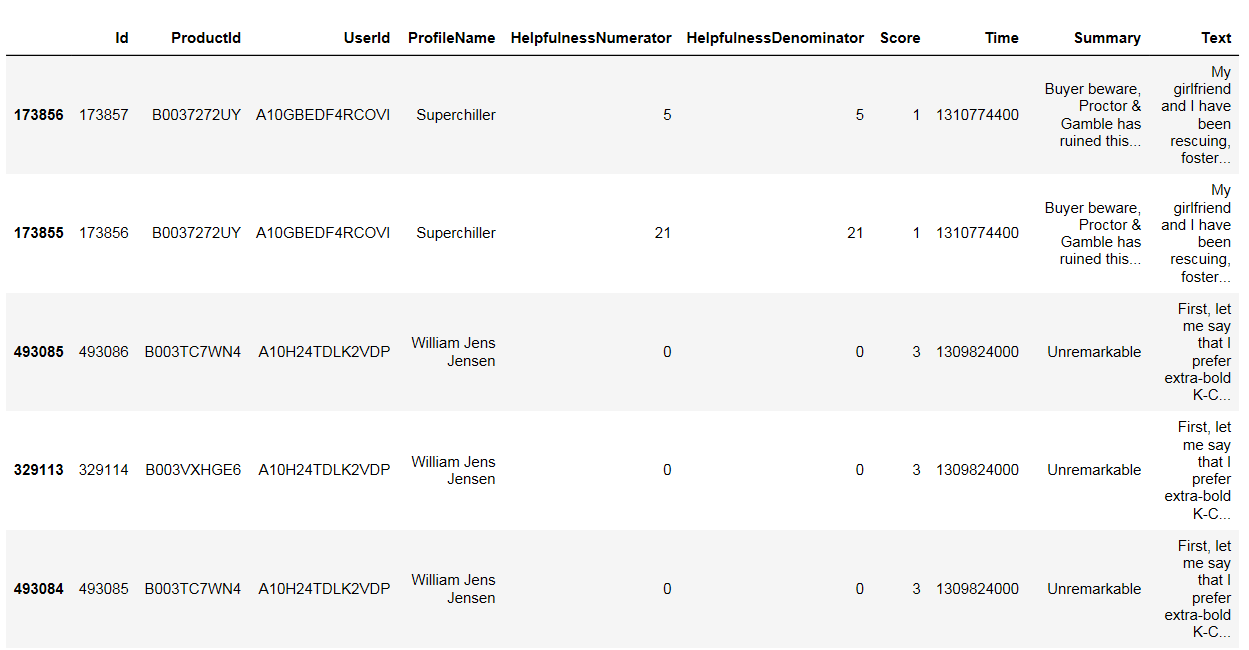


*Figure 1.* Illustration of the dataset

## Data Cleaning:

1. **Remove duplicated reviews**:

From Figure 2, we can see the the dataset has duplication reviews for one review. For each review, two records may have different Helpfulness Denominator and Helpfulness Numerator. In that case, we assume the record is cumulated and only kept the record with largest number of Helpfulness Denominator.



*Figure 2*. Duplicated reviews

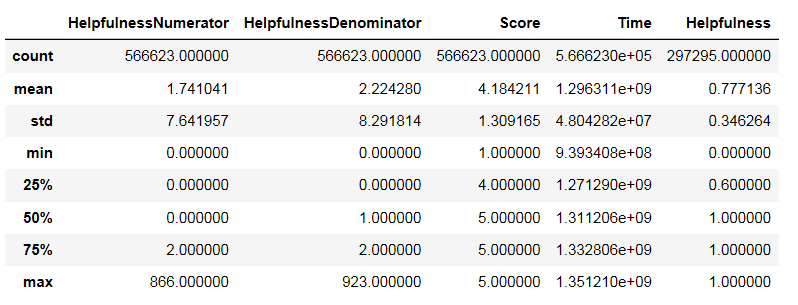
1. **Remove the empty reviews**:

The purpose of the project is to do semantic analysis for customer reviews. When reviews are not applicable, that review is not helpful. We removed the empty reviews.

1. **Text preprocessing**:

For text preprocessing, for this milestone, we converted the text into count vector and TF-IDF. We limited our vocabulary into 1000 words and with minimum document frequency of 800. We will play around later with this number. Together with that, we also run stemming and removing the stopwords as a standard pre-processing for text.

## Exploratory Numerical Data Analysis:

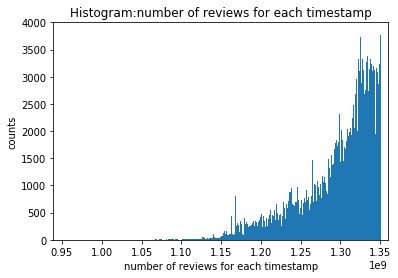
Helpness is defined as Helpness = HelpfulnessNumerator / HelpfulnessDenominator, which indicates the helpness of the review. The summary of the dataset is shown in Figure 3. 

*Figure 3.* Summary of the dataset

From the summary, we can see only around 50% reviews have votes for helpness. To better understand the dataset, following preliminary analysis have been done.

1. **Trend of the number of reviews**

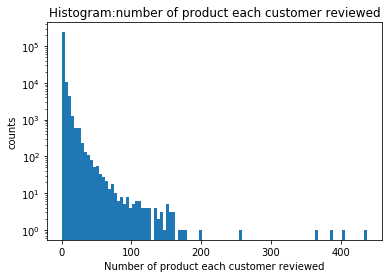
The time span of the dataset is from 1999 to 2012. Figure 4 shows trend of the number of the reviews for each timestamp. As timestamps increase, the number of reviews increases. This is meet our intuitive thinking. Because from 1999 to 2012, the number of Amazon customer is increasing exponentially, the number of reviews is increasing accordingly.



*Figure 4*. Trend of the number of reviews

1. **The customer of reviews**

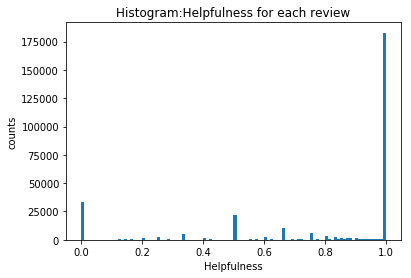
Every unique customer, which is indicated by user Id, reviewed multiple products. Each review has different helpfulness. Figure 5 shows for each customer, how many products he/she reviewed. Within 256059 customers, majority of the customers only reviewed one product. Third quartile (Q3) is 2. This is a good indicator for semantic analysis. It means the reviews are written by various people. The result should be more general compared to the situation that all reviews are written by limited number of customers.



*Figure 5.* Distribution of number of products per customer

1. **The helpfulness of reviews**

Although not all reviews are voted for helpfulness, there are still 297,295 reviews helped customers to make the decision. Figure 6 directly shows the distribution of helpfulness. Unsurprisingly, around 60% of the reviews are noted as 100% helpful. It is interesting that there are a lot of reviews are voted as 50% helpful. It may caused by the way reviewers wrote, which is worth to discover in the future.

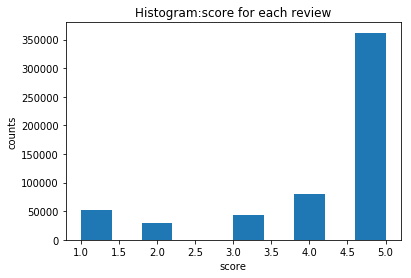


*Figure 6.* Distribution of helpfulness

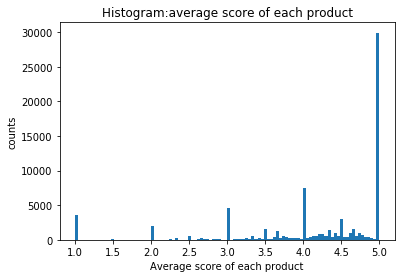
1. **The score of reviews**

The score of the product indicates the customer rated quality. The counts for each scores are shown in Figure 7. More than 60% of the reviews rated 5 score for reviews. The customer tends to rate high score for product. This phenomenon also can be seen in Figure 8. Every product has multiple reviews and we calculate the average score for one product. In general, the average score is also skewed to high score. Note that there are relatively more score counts fall in the 4 to 5 range, especially in 4.5. It means when customer rate product, score 4 and 5 is not clearly separated, which also same with our daily experiences. For clearer picture, in figure 9, we divide the dataset into two parts: positive and negative reviews. Positive reviews are the reviews who has score at least 3. While negative reviews are the reviews which get score less than three.

This highly skewed score distribution can indicate that while we do the semantic analysis, the size of the clusters should be different. This characteristic may affect the result of K-means algorithm. We can validate the assumption in the next milestone.



*Figure 7.* Scores distribution of reviews



*Figure 8.* Average score distribution of reviews



*Figure 9.* Positive reviews

## Exploratory Text Data Analysis:

For text, we will see what words contribute the most. After converting to count vector, we get the to 10 words in our review fields (Table 2).

Table 2: Word counts of reviews

|  |  |
| --- | --- |
| **Words** | **Count** |
| veri | 963 |
| use | 953 |
| tast | 887 |
| product | 689 |
| one | 603 |
| love | 514 |
| like | 496 |
| great | 388 |
| good | 380 |
| flavor | 345 |

To get a better understanding about the reviews, we run LDA for both count vector and TF-IDF representation of the review. However, we could not see any result clear distinction between each topic.

Table 3: LDA result over the count vector representation

|  |  |
| --- | --- |
| 10 Components | Top words |
| 1 | bar, love, eat, free, gluten, snack, protein, taste, great, organ |
| 2 | use, salt, mix, butter, make, taste, peanut, like, great, popcorn |
| 3 | love, find, amazon, store, great, use, buy, price, local, get |
| 4 | like, taste, flavor, chocolate, good, chip, cookie, eat, snack, bag |
| 5 | order, product, amazon, price, box, buy, veri, good, ship, purchase |
| 6 | flavor, hot, make, great, good, like, veri, love, add, taste |
| 7 | tea, coffee, flavor, cup, like, coffee, taste, good, drink, one |
| 8 | dog, treat, get, love, one, like, time, day, use, give |
| 8 | taste, like, use, water, product, drink, sugar, flavor, oil, tri |
| 10 | food, cat, dog, eat, like, feed, love, chicken, one, drive |

Table 4: LDA result over TF-IDF

|  |  |
| --- | --- |
| 10 Components | Top words |
| 1 | use, make, cook, add, great, taste, like, good, rice, mix |
| 2 | amazon, price, store, find, great, buy, love, bar, local, product |
| 3 | tea, flavor, green, hot, taste, like, drink, good, sauce, love |
| 4 | coffe, coffee, cup, flavor, like, strong, tast, roast, good, blend |
| 5 | dog, food, cat, treat, love, eat, like, one, get, chew |
| 6 | chip, salt, popcorn, flavor, like, tast, bag, chips, good, great |
| 7 | chocolate, taste, like, flavor, candi, sugar, bar, sweet, good, syrup |
| 8 | order, product, veri, box, arrive, receive, amazon, time, package, great |
| 9 | cookie, gluten, love, like, eat, free, taste, great, good, bread |
| 10 | drink, water, taste, like, use, product, tea, bottle, help, good |

Table 3 is the result of LDA over the count vector representation, while table 4 shows the LDA result over TF-IDF. We can see the category for each topic. For example, the first topic on the first table talks about snack, most likely protein bar while the second topic is likely about salt and butter. We couldn't tell much from this.

Since we also have positive and negative review, we will implement TSNE in the future, to see if we can run sentiment analysis using this algorithm. Furthermore, we can also use word vector representation and see if it helps.

## Conclusion:

In this project milestone, we preprocessed the data and briefly explored the dataset. Also LDA analysis over both word counts feature and TF-IDF feature have been done. Some interesting topics and natural of the problems are found. We hope these problem characteristics can be used for future work.

## References

[1] http://snap.stanford.edu/data/web-FineFoods.html