FINAL ASSIGNMENT: DATA ANALYSIS OF AMAZON'S DATA SET

CENG 3521, DATA MINING

Senem Yalın - Almira Gürkan
senemyalin@posta.mu.edu.tr - almiragurkan@posta.mu.edu.tr

Sunday 17th January, 2021

Abstract

Data mining is the process of automatically discovering useful information in large data repositories. We are drowning in data, but starving for knowledge. Natural language processing (NLP) is an exciting branch of artificial intelligence (AI) that allows machines to break down and understand human language. In this assignment, we used NLP techniques to interpret text data that we are working with, for our analysis. In addition, Recommendation System and Neural Network are used for prediction.

1 Introduction

This assignment has 7 steps which are Data Cleaning, Exploratory Data Analysis, Sentiment Analysis, Text Generation, Regression, Neural Network Process, Recommendation Process. After these steps, we had forward inferences.

2 Assignments

When we are doing text analysis part, we used several NLP libraries in Python including TextBlob along with the standard machine learning libraries including pandas and scikit-learn. We got results from the data which we analysed and plotted these results in many different ways. Also we did Recommendation Analysis. After that, we printed the suggestions.

2.1 Data Pre-processing

2.1.1 Getting the data

We had 2 data sets, so we merged our data according to the columns that we will use. Our data which we merged, has the columns below.
overall - rating of the product
reviewTime - time of the review (raw)
reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin - ID of the product, e.g. 0000013714
reviewText - text of the review
title - name of the product
brand - brand name

2.1.2 Cleaning the data

We made text lowercase, removed punctuation and removed words containing numbers. Also we got rid of some additional punctuation, non-sensical text that was missed the first time around and stop words on ”reviewText” column.

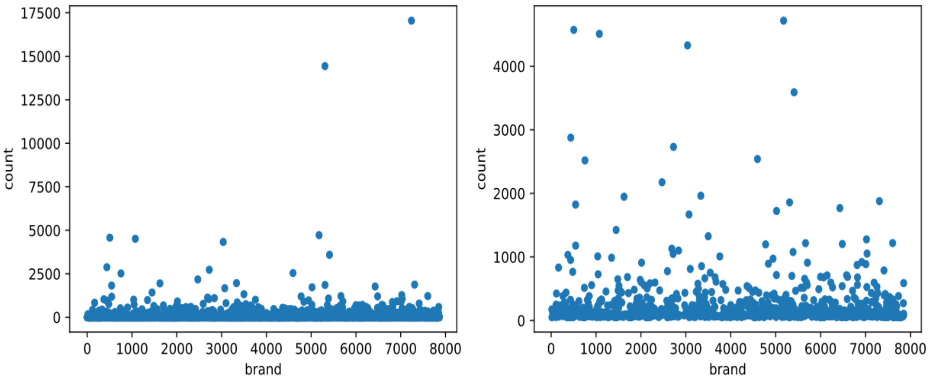
2.1.3 Organizing the data

We organized the cleaned data into a way that is easy to input into other algorithms.
1-We assigned numbers to brands so that the machine could understand, and we gave the same numbers to those with the same brand.

2.2 Brand Count

2.2.1 Exploratory Data Analysis

1-We grouped brands according to the number of products they sell.
2-We saw that how many products were purchased from which brand and how many brands are there.
3-Because our data makes sense, we removed the most and least purchased brands from this data.
4-We plotted our 2 results.



2.3 Comparing Analysis

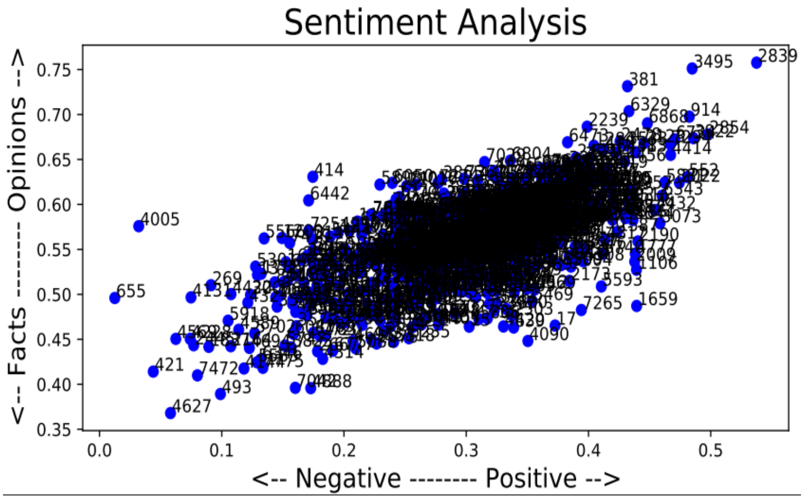
2.3.1 Sentiment Analysis

1-We applied the ”sentiment.polarity” and ”sentiment.subjectivity” methods of the ”textblob” library to find the polarity and subjectivity of brand’s ”reviewText”.
2-Each word in ”reviewText” column is labeled in terms of polarity and subjectivity.
3-We took the average subjectivity and polarity of product reviews that have the same brand and assigned them to that brand.

Polarity: How positive or negative a word is. -1 is very negative. +1 is very positive.

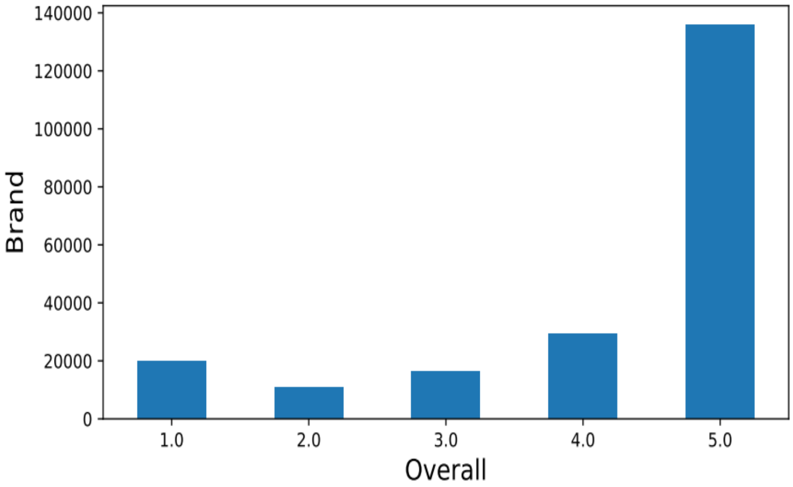
Subjectivity: How subjective, or opinionated a word is. 0 is fact. +1 is very much an opinion.

4-We plotted these results.



2.3.2 Rate Analysis

1-We took the average of product rates that have the same brand and assigned them to that brand, so we found the average number of rate a brand gets.
2-We grouped overall according to brand then plotted as bar chart.

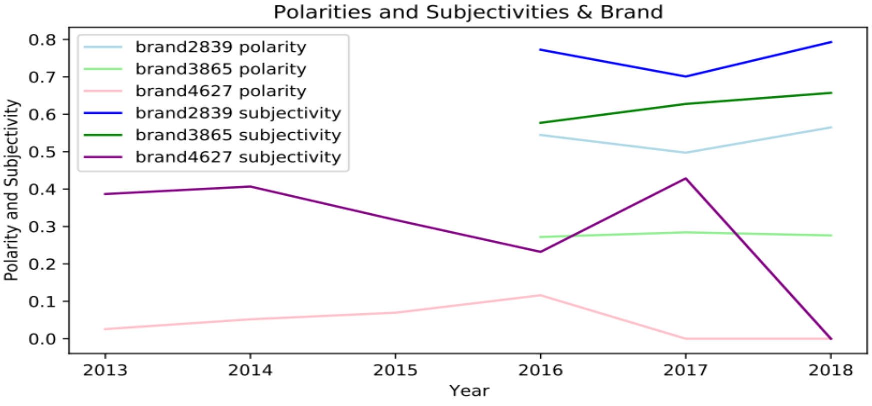


2.3.3 Results

When we compare two plotting results, saw that brands are generally positive on first plotting and these brands got 5 point for overall on second plotting, so our analysis is consistent.

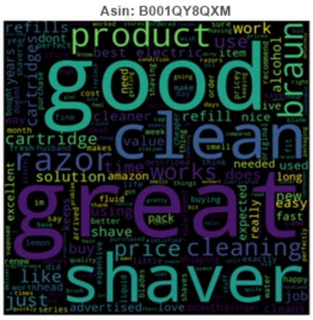
2.4 Changes of Polarity and Subjectivity Over the Years

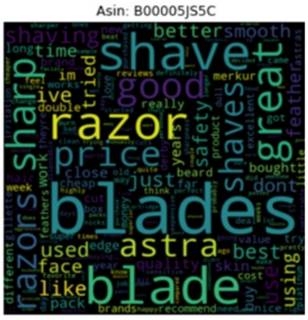
1-We chose 3 brands according to Sentiment Analysis which is above. One of them has the most polarity and subjectivity, one of them has the middle ones and one of them has less ones. We wanted to see polarity and subjectivity changes year by year.
2-We plotted these results.

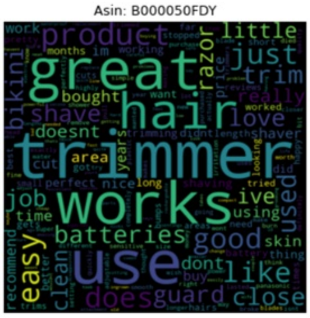


2.5 Most Used Words in ”reviewText”

1-We looked at how many products are purchased, then chose 3 products which are most sold.
2-We organized our data according to these 3 products.
3-We combined the comments made for the same product in the same dataframe cell.
4-We split our cells separately, then put them into different dataframe for these 3 products.
5-We grouped our words according to how many are these.
6-Since dataframe type does not work on Wordcloud,so we put these datas into ”dictionary” and plotted them as word cloud.

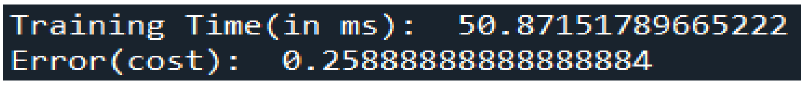






2.6 Neural Network Process

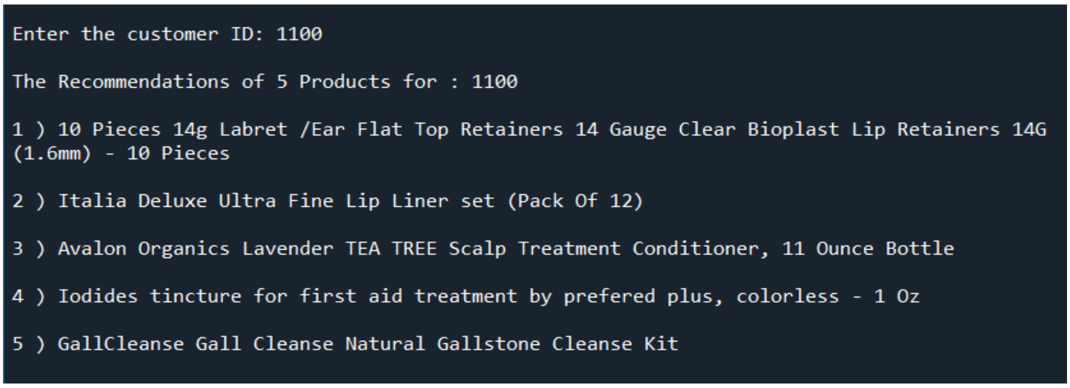
We taught the product information and tried to guess which brand it was.
1-We assigned a number to ”asin”, ”brand” and ”reviewerID”, so that the machine could understand.
2-We split our data such that randomly selected 70 tuples are used for training while 30 tuples are used for testing.
3-We applied MLPClassifier on training data.
4-We determined the number of rows and we gave value to the hidden layer size for the error rate is at the optimum value.
5-We predicted Brand of product.
6-We printed error rate and traning time on the console.



```
Training Time(in ms):  50.87151789665222
Error(cost):  0.2588888888888884
```

2.7 Recommendation Process

We wanted to predict the products a person could buy by looking at the products and rates they purchased.
1-We assigned a number to ”asin”, ”brand” and ”reviewerID”, so that the machine could understand.
2-We calculated the normalized rating for a customer. This data would be used to calculate the final score for the customer later.
3-We have used cosine similarity function of sklearn to calculate the similarity.
4-We used two methods commonly used for this :
 a)The customer average over the row.
 b)The product average over the column.
5-We calculated the similarity between the users.
6-We created ”find similar customers” function which takes the similarity matrix and the value of n as input and returns the nearest n neighbors for all the customers. We chose n as 30.
7-We created ”customer product score” function which uses our above discussion to calculate predictions.
8-We asked the user customer ID and printed the recommendations of 5 products for this customer.



```
Enter the customer ID: 1100

The Recommendations of 5 Products for : 1100

1 ) 10 Pieces 14g Labret /Ear Flat Top Retainers 14 Gauge Clear Bioplast Lip Retainers 14G
(1.6mm) - 10 Pieces

2 ) Italia Deluxe Ultra Fine Lip Liner set (Pack Of 12)

3 ) Avalon Organics Lavender TEA TREE Scalp Treatment Conditioner, 11 Ounce Bottle

4 ) Iodides tincture for first aid treatment by prefered plus, colorless - 1 Oz

5 ) GallCleanse Gall Cleanse Natural Gallstone Cleanse Kit
```

3 Conclusion

In conclusion, we worked on NLP(Natural Language Processing), Neural Network Processing and Recommendation System. We learned how to analyze text-based data. In this analysis, we cleared the text by removing the punctuation and stop words. As a result of our research, we saw that every word has polarity and subjectivity, interpreted according to these values.

Then we worked on application of a neural network with multi layer perceptron. By using MLPClassifier we taught the machine the product information and had it guess which brand it was.

In addition, we experienced Recommended System. We focused on (User-Based Collaborative Filtering) UB-CF which is a memory-based method. The main idea behind UB-CF is that people with similar characteristics share similar taste.

The motivation behind coding this assignment is to deep dive into the algorithm and understand how data mining process work. We learned the techniques we will use while carrying out these studies by researching and experimenting. As we do such works, we gain the ability to produce solutions to existing problems over a real scenario as much as possible.