Northeastern University
College of Professional Studies

ALY 6040 – DATA MINING APPLICATIONS

Prof. Justin Grosz

Module 5: Technique Practice

05/18/2023

Pooja Kairamkonda

INTRODUCTION

Natural language processing (NLP), commonly referred to as text analytics or text mining, is a potent technology used to glean valuable knowledge and information from unstructured text data. Text mining can be used to assess customer feelings, extract topics, and comprehend customer comments at scale when applied to Amazon reviews. One important aspect of text mining in Amazon reviews is topic extraction. By applying topic modeling techniques like Latent Dirichlet Allocation (LDA), businesses can uncover the main topics or themes discussed by customers. This categorization allows for a better understanding of customer preferences and common issues or trends associated with the products.

In E-commerce industry, a large scale data gathering happens with customer purchase reviews of each product. These reviews play an important role to the business analysis to review what customers are actually looking for and target specific products by improving business strategies. This project focuses on topic modelling technique that can further provide review summarization. The technique uses automated approaches to distill massive amounts of Amazon evaluations into succinct, insightful summaries. This enables organizations to rapidly understand the overall sentiment, key topics and critical issues conveyed in a group of reviews without having to manually read them all. At the end, the project provides recommendations to the company based on the topic model results.

DATA PRE-PROCESSING

The given dataset ‘Reviews.csv’, contains 568454 records and 10 variables. This dataset consists of Amazon product purchase reviews provided by customers. In the section, we will discuss the data preprocessing steps that involves cleaning and transforming raw text data into

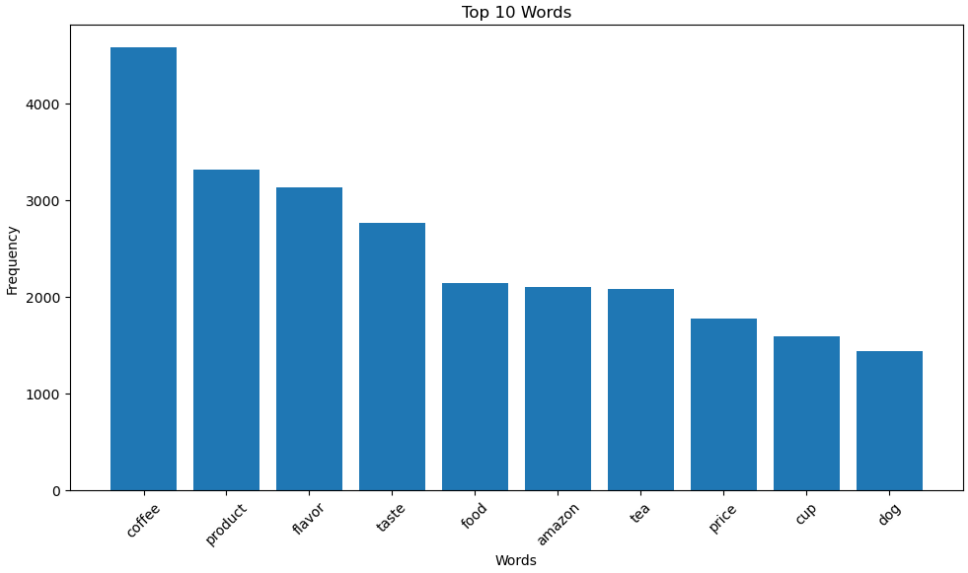
required format suitable for further analysis by using the Natural Language Toolkit libraries and functions. Since the reviews are written by customers it contains lot of noisy data such as punctuation, special characters, or numbers that do not contribute to the overall meaning. These unwanted elements were removed initially in order to focus only on the relevant content. The entire text must then be divided into manageable chunks in order to reveal the specifics of our study. To do this, tokenization—the process of separating words or tokens from the input text -was carried out. These words were categorized based on their part-of-speech tagger. The tagging assigns grammatical tags to each word in a sentence, such as noun, verb, adjective, etc. For our analysis, mainly nouns were focused and with the help of tokenization and pos tagging all the nouns were collected. Further, an additional set of stop-words was included to ensure that these additional words are included when removing stop words from the text during text analysis. Finally the preprocessed text data object was set into a document-term matrix, where ach row represents a document (text) and each column represents a term (word), with the cell values indicating the frequency of each term in each document. This matrix was used to further analysis.

ANALYSIS & INTERPRETATION

In this section we will discuss in detail about the analysis performed on text which was collected during initial processing. Here, the set of nouns obtained, was represented using a word cloud which is on of the ideal ways to represent text mining data. The would cloud is displayed as shown below:

From the above figure, the bigger & bolder words represents the mostly occurred or highlighted words from the customer reviews that included food, product, meat, stew, case, mouthful, etc. The occurrence of these words suggests that customers frequently mention these food-related terms in their reviews. This suggests that customers' opinions and experiences with certain food goods, such as their quality, flavor, or variety, may be relevant to the reviews of other products on Amazon. The presence of words like product and mouthful suggests that customers express their opinions regarding the quality of the products they purchased. The specific food items mentioned in the word cloud such as, meat, sugar, grape, beer, stew explains the customer's interest in the particular products. The majority of the time, it has been seen that customers choose to buy food and grocery products from Amazon Fresh (the food and grocery division) because it gives prime membership benefits and favorable prices on these items. This analysis will be helpful to the division to maintain fresh food stock all the time and replace/try replacing new products and analyze the reviews of their experiments.

Additionally, we have displayed the top most frequency of nouns obtained in the collection as shown below.



Next part of our analysis included, Topic extraction through LDA (Latent Dirichlet Allocation). We have used this technique to provide the company with insights to uncover the main topics or themes discussed by customers. Each topic is represented by a set of topically significant terms, ranked based on their contribution to the respective topic. By analyzing these topics, businesses can identify common customer preferences, popular features, recurring issues, or emerging trends related to their products or services. The choice of number to topic being generated by the model was kept to 3 and results generated are as follows:

Topic 0: This subject appears to be connected to flavors and tea. Several words are used, including "tea," "taste," "flavor," "crackers," "sugar," and "chocolate." This subject covers discussions about various tea flavors, personal preferences, and perhaps some associated goods or dishes.

Topic 1: Customer's main interest was related to topic 'Cofee'.The words included, "coffee," "cup," "flavor," "cups," "taste," and "keurig." It indicates conversations on different coffee flavors, preferences for various brewing techniques (like the Keurig), and possibly the cost of coffee goods.

Topic 2: This topic is focused on customer's reviews on 'Dog treats and food'. The topic highlights the words like "treats," "dog," "food," "dogs," "treat," and "cookies." Customer reactions on dog treats, preferred dog foods, suggested products, and price considerations provides an overview of this topic.

The interpretation of the topics is based on the most significant terms associated with each topic, as indicated by their corresponding weights is shown below:

Topic 0		**Topic 1**		**Topic 2**	
word	weight	word	weight	word	weight
tea	0.019	coffee	0.08	treats	0.028
taste	0.017	cup	0.027	dog	0.028
flavor	0.017	flavor	0.019	food	0.026
product	0.012	cups	0.015	dogs	0.018
crackers	0.012	taste	0.013	product	0.015
sugar	0.01	amazon	0.012	treat	0.015
mix	0.009	product	0.011	cookies	0.011
water	0.008	price	0.009	bag	0.009
chocolate	0.008	chocolate	0.009	price	0.008
amazon	0.007	keurig	0.008	amazon	0.008

RECOMMENDATIONS & CONCLUSION

We relate the findings to reality using the analysis we completed, and we can provide recommendations to the business regarding customer reviews. It is evident that product quality and flavor are of utmost importance to customers. The presence of topics related to tea, coffee, and flavors suggests that customers value a satisfying taste experience. Sellers should put a premium on product quality and flavor consistency in order to match client expectations. Delivering goods that meet or beyond consumer expectations necessitates regular quality control checks, the sourcing of high-quality ingredients, and proper production procedures.

To capitalize on Topic 1 analysis that revolves around ‘Coffee’, sellers can consider expanding their coffee product line to include diverse blends, flavors, and brewing formats. Businesses can adapt to various customer tastes and stand out in a competitive market by providing a wide range. The presence of term ‘Price’ can help the company build pricing strategies and evaluate the perceived value of their product. Adjustments in pricing, occasional promotions, or bundle deals can be implemented to incentivize customer purchases and enhance the perceived value of the products. Based on the Topic 3 analysis of Pet treats and food, to cater to pet owners' needs, businesses should invest in high-quality ingredients and focus on the nutritional value of their pet treats and food. Offering a variety of options to accommodate different dietary preferences and restrictions is essential.

In conclusion, I would suggest the business can help segment clients based on their preferences, interests, or sentiments indicated in reviews among the numerous ways where text mining is valuable for analysis. Businesses can customize their marketing campaigns, customised recommendations, or customer service to distinct consumer segments by grouping customers with similar review patterns or sentiments. The study can further be enhanced by an sentiment analysis where sentiment expressed in reviews, such as positive, negative, or neutral sentiments, we can gain an understanding of customers' opinions and satisfaction levels.

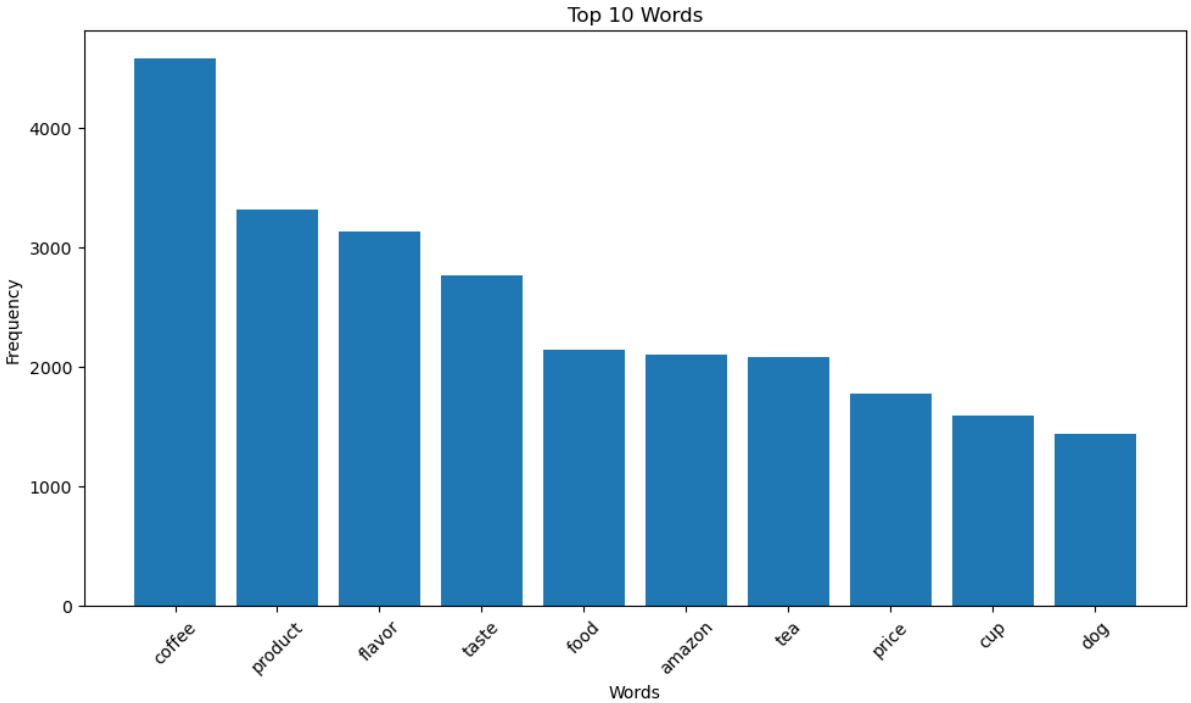
REFERENCES

1. What is Text Mining, Text Analytics and Natural Language Processing?. Retrieved from - https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing
2. Poonam Rao (Jul 19, 2021). Text Analytics & Text Mining: the next Big thing in Data Science. Retrieved from - https://medium.com/nerd-for-tech/text-analytics-and-text-mining-6bcc8e83f473
3. Sowmya Vivek (Aug 22, 2018). Analyzing Customer reviews using text mining to predict their behaviour. Retrieved from - https://medium.com/analytics-vidhya/customer-review-analytics-using-text-mining-cd1e17d6ee4e

APPENDIX

1. Word cloud of nouns

2. Top 10 words in the reviews



3. Topic modelling results

```
[(0,
  '0.019*"tea" + 0.017*"taste" + 0.017*"flavor" + 0.012*"product" + 0.012*"crackers" + 0.010*"sugar" + 0.009*"mix" + 0.008*"wat
er" + 0.008*"chocolate" + 0.007*"amazon"'),
 (1,
  '0.080*"coffee" + 0.027*"cup" + 0.019*"flavor" + 0.015*"cups" + 0.013*"taste" + 0.012*"amazon" + 0.011*"product" + 0.009*"pri
ce" + 0.009*"chocolate" + 0.008*"keurig"'),
 (2,
  '0.028*"treats" + 0.028*"dog" + 0.026*"food" + 0.018*"dogs" + 0.015*"product" + 0.015*"treat" + 0.011*"cookies" + 0.009*"bag"
 + 0.008*"price" + 0.008*"amazon"')]
```