NORTHEASTERN UNIVERSITY

Student – Sajal Gangrade

Technique Practice – Module 5

Class Number - ALY 6040

Class Name – Data Mining Application

CRN Number – 20231

Date: 03/23/35

Professor – Justin Grosz

Text Mining on Amazon Reviews

INTRODUCTION

In the current digital age, customer feedback holds significant value for businesses. Amazon, as one of the leading e-commerce platforms, receives millions of product reviews from customers worldwide. These reviews contain valuable information about customer satisfaction, product quality, pricing, and overall shopping experience.

Manually reading through thousands of reviews is not practical. Therefore, this assignment focuses on applying Text Mining and Natural Language Processing (NLP) techniques to analyze the Amazon product reviews dataset. By leveraging these methods, we aim to extract key themes, common product attributes, and customer sentiments present within the reviews. The insights gained from this analysis can help businesses like Amazon better understand their customers' needs, identify product strengths and weaknesses, and make informed decisions to improve their services.

PROBLEM STATEMENT

The primary problem addressed in this assignment revolves around understanding and interpreting customer feedback from Amazon product reviews. These reviews are written by customers who share their experiences, opinions, and satisfaction levels regarding various products. The Amazon review dataset contains thousands of reviews, each consisting of textual comments, product information, ratings, and other metadata.

To address this challenge, we will apply text mining techniques to uncover patterns and key themes within the reviews. Specifically, we aim to identify:

- Frequently mentioned product attributes,
- Recurring customer sentiments (both positive and negative),
- Common product categories or issues.

By transforming unstructured text into meaningful insights, this analysis will help Amazon's product team to gain a deeper understanding of customer preferences and concerns. It will allow them to make data-driven decisions to improve product quality, marketing strategies, and overall customer experience.

DATA CLEANING

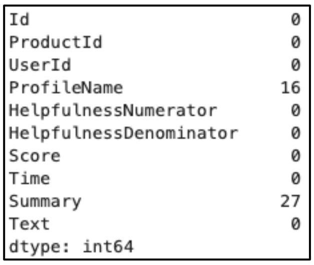
Before performing any text mining or NLP analysis, it was essential to clean and preprocess the dataset to ensure accuracy and reliability. Unstructured text data often contains unnecessary elements such as punctuation, stop words, and irrelevant tokens that do not contribute meaningful insights. Therefore, careful attention was given to preparing the data in a systematic way.

Checking for Missing Values:

Before conducting text mining and NLP analysis, we first checked the dataset for any missing values to ensure its completeness and reliability. But we identified a small number of missing values in two specific columns:

Profile Name: 16 missing values
Summary: 27 missing values

Importantly, the Text column, which contains the main body of the customer reviews, had no missing values.



```
Id                       0
ProductId                0
UserId                   0
ProfileName             16
HelpfulnessNumerator     0
HelpfulnessDenominator   0
Score                    0
Time                     0
Summary                 27
Text                     0
dtype: int64
```

Given that the Profile Name column merely holds the reviewer's name, and the Summary column contains a short title summarizing the review, we made a conscious decision to proceed without dropping or filling in these missing values. Removing rows based on non-essential fields would have unnecessarily reduced the dataset size without contributing to the accuracy or depth of our analysis. Our primary focus was on the detailed Text column, which had complete data.

Filtering the Dataset:

To focus on meaningful and balanced feedback, we filtered out reviews with a score of 2 or lower. These reviews often contain extreme negativity, which might skew the analysis if not handled carefully. We chose to include reviews with a Score greater than 2, representing neutral to positive customer experiences. Additionally, we limited the dataset to the first 15,000 reviews to manage computational resources and ensure smooth processing during model training.

Handling Problematic Words:

While inspecting the review text, we noticed several common words that appeared frequently but offered little analytical value. Examples include casual terms like "like," "im," "dont," "yeah," and HTML tags like "br." These words, although not part of standard stop word lists, could introduce noise in the analysis and distort the visualizations. Therefore, we manually added them to the custom stop words list to ensure a clearer focus on meaningful content. Addressing these problematic words was crucial in improving the quality and interpretability of our text mining results.

Text Cleaning and Tokenization:

We applied several text preprocessing steps to clean the review text:

1. Lowercasing all text to ensure uniformity.
2. Removing punctuation and digits**,** as these do not add value in understanding customer sentiment.
3. Tokenizing the text**,** breaking it down into individual words for easier analysis.
4. Extracting only nouns using part-of-speech tagging. This decision was intentional, as nouns typically capture the core elements of customer feedback, such as product names, attributes, or specific issues.

Stop Words Removal:

In addition to standard English stop words (like "the," “and”, "is"), we identified additional words frequently present in casual language but offering little analytical value. Words such as "like," "im," "don’t," "yeah," "br" (HTML tag remnants) were added to the stop words list. Removing these helped reduce noise and sharpen the focus on meaningful content.

Key Preprocessing Decisions:

- **Why extract only nouns?**
  Nouns typically represent key topics or entities customers mention—like product names, features, or issues. By focusing on nouns, we filtered out filler words while retaining the core information.
- **Why remove additional stop words?**
  Words such as "im" or "dont" appear frequently but do not contribute to understanding product-specific feedback. Eliminating them improves clarity in visualizations and topic modeling results.

ANALYSIS

After preparing and cleaning the dataset, we applied various text mining techniques to uncover meaningful patterns. Our analysis focused on visualizing word frequencies, identifying hidden topics, and grouping reviews based on similar language usage. Each technique was chosen carefully to provide a comprehensive understanding of customer feedback.

Word Frequency Analysis

We began by analyzing the frequency of words used across all reviews. Using a document-term matrix, we computed the occurrence of each word after stop words and irrelevant terms had been removed. To present this information clearly, we created a bar chart illustrating the top 15 most frequently used words (see Appendix A, Figure 1).

The bar chart highlights terms such as **"good," "coffee," "great," "taste," "product,"** and **"flavor"** as being among the most common. This result confirms that customers often focus on product quality, taste, and specific product categories in their feedback. Presenting this data in a bar chart format provided a quantitative view of word frequencies, making it easy to compare how often certain words were mentioned.

Word Cloud Visualization:

To complement the bar chart, we generated four word clouds. The first two visualize the top 30 words in each of the two topics identified (Appendix B, Figures 2 and 3), while the next two focus on the top 10 words for each topic (Appendix B, Figures 4 and 5).

This dual approach allowed us to:

- Provide a broad perspective of the most frequently discussed terms.
- Offer a simplified, high-level view of the dominant words.

For Topic 1, words such as **"dog," "treats," "food," "taste," "product,"** and **"dogs"** were most prominent, indicating customer discussions centered around pet food, particularly dog treats.

For Topic 2**,** the most visible words were **"coffee," "flavor," "tea," "cup," "product,"** and **"taste,"** pointing towards reviews related to beverages, especially coffee and tea products.

These visualizations reinforce the patterns identified in the bar chart and offer an intuitive snapshot of customer focus areas.

Topic Modeling:

To dive deeper into the themes present in the reviews, we applied Latent Dirichlet Allocation (LDA) for topic modeling. This technique uncovered two distinct topics based on word co-occurrence patterns, as reflected in the word clouds.

The choice of topic modeling was intentional. It allowed us to automate the discovery of key themes without manually labeling the data. The resulting topics clearly showed:

- **Topic 1:** Dominated by pet-related products, particularly dog food and treats.
- **Topic 2:** Focused on beverage products, primarily coffee and tea.

This aligns well with the nature of products commonly reviewed on Amazon and supports our initial goal of understanding customer feedback trends.

Clustering Analysis:

To explore if the reviews naturally group based on word usage, we applied **K-Means Clustering** after reducing the dimensionality of the data using Truncated SVD**.** This step helped simplify the dataset while retaining important patterns. We chose to create two clusters, aligning with the two topics identified earlier. The resulting scatter plot (Appendix C, Figure 6) clearly shows two distinct groups of reviews. One cluster predominantly reflects reviews related to pet food products like dog treats, while the other focuses on beverages such as coffee and tea. This clustering analysis not only confirms the findings from our topic modeling but also provides a visual representation of how customer feedback is organized around specific product categories. It demonstrates that customer reviews are not random, but form structured, meaningful patterns, offering valuable guidance for Amazon's product and marketing teams.

Word Frequency Table:

Additionally, we included a table summarizing the top 15 most frequent words and their respective counts (Appendix D, Table 1). This table offers a factual, numeric perspective and supports the visual findings from the bar chart and word clouds.

Key Learnings:

Several key insights emerged from the analysis:

1. **Customer focus areas:** Customers consistently emphasize product quality, taste, and price across reviews.
2. **Product categories:** Pet food (specifically dog treats) and beverages (particularly coffee and tea) are major discussion points.
3. **Positive sentiment dominance:** Words like "good," "great," and "love" suggest a generally favorable tone among reviewers

Does the Analysis Match the Context?

Yes, the analysis aligns closely with the context of the dataset. Amazon's diverse product range includes items such as pet food and beverages, making it logical for reviews to heavily feature terms related to these categories. The appearance of words like **"dog treats,"** **"coffee,"** and **"taste"** reflects genuine customer concerns and interests, validating the effectiveness of our text mining approach.

INTERPRETATION AND RECOMMENDATIONS

The text mining and NLP analysis of the Amazon product reviews revealed consistent and valuable insights into customer behavior and preferences. One of the most critical steps in this process was tokenization, which involved breaking down each review into individual words or tokens. Tokenization served as the foundation of the entire analysis, allowing us to clean the text effectively, calculate word frequencies, remove irrelevant words, and prepare the data for deeper techniques such as topic modeling and clustering. This approach aligns with the core NLP concepts discussed in class and proved essential in making the large volume of unstructured text analyzable.

Through our analysis, we identified two dominant themes in customer feedback: pet food products, specifically dog treats, and beverage products, primarily coffee and tea. Words like "taste," "dog," "treats," "coffee," "flavor," and "product" consistently appeared, highlighting customer focus on product quality, taste, and price. The clustering analysis visually confirmed these patterns by grouping reviews naturally based on product categories.

Based on these findings, we recommend the following actions:

1. **Maintain Product Quality and Consistency:**
   Customers frequently mention product taste and quality. Ensuring consistency in these aspects, particularly in popular categories like pet food and coffee, can help maintain positive customer sentiment.
2. **Tailor Marketing Strategies by Product Category:**
   Since the analysis shows clear focus areas, marketing efforts should emphasize the key attributes that customers care about, such as taste and product quality for coffee, or nutritional value and packaging for pet treats.
3. **Regular Monitoring of Customer Feedback:**
   Applying similar text mining techniques regularly can help Amazon stay updated on changing customer preferences, detect emerging issues early, and make data-driven product improvements.
4. **Use Clustering Insights for Targeted Engagement:**
   The clear clustering of reviews suggests potential for segmenting customers and tailoring communications or recommendations based on the products they review most frequently.

In conclusion, the combination of tokenization, word frequency analysis, topic modeling, and clustering provided a well-rounded understanding of customer feedback. These techniques allowed us to extract actionable insights and offer recommendations that align with Amazon’s business objectives and product offerings.

REFERENCES

- McAuley, J., & Leskovec, J. (n.d.). *Amazon Fine Food Reviews* [Data set]. Kaggle. https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. https://www.nltk.org/
- Mueller, A. (n.d.). WordCloud Documentation. Retrieved from https://amueller.github.io/word_cloud/
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://scikit-learn.org/stable/index.html

APPENDIX

Appendix A: Visualizations

Figure A1: Bar Chart of Top 15 Most Frequent Words

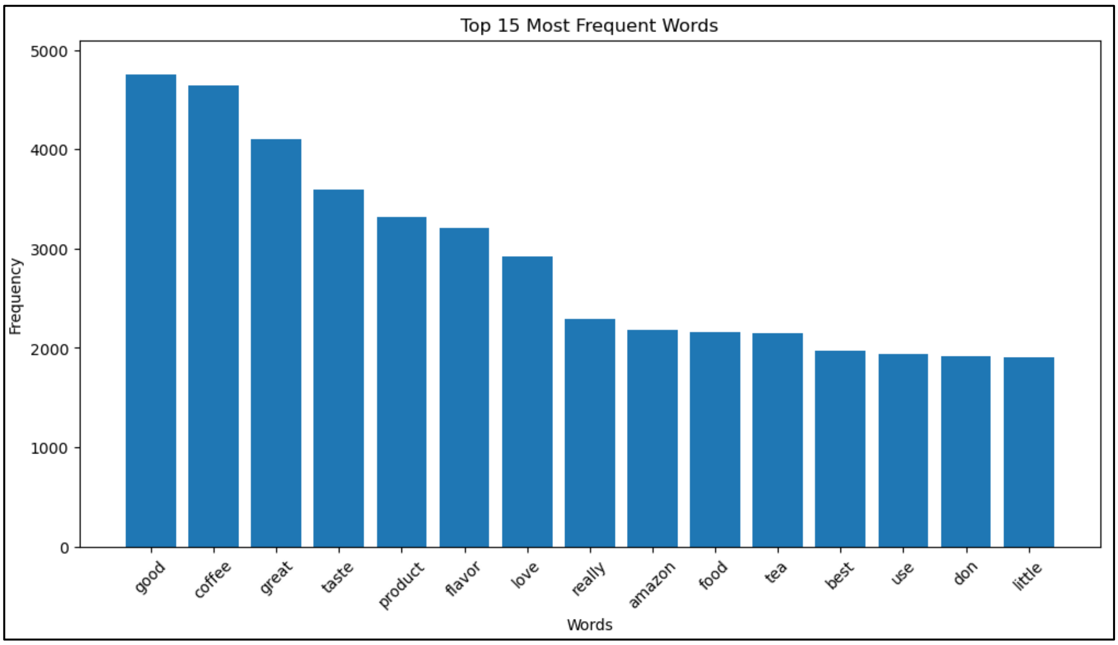


Figure A2: Word Cloud for Topic 1 (Top 30 Words)

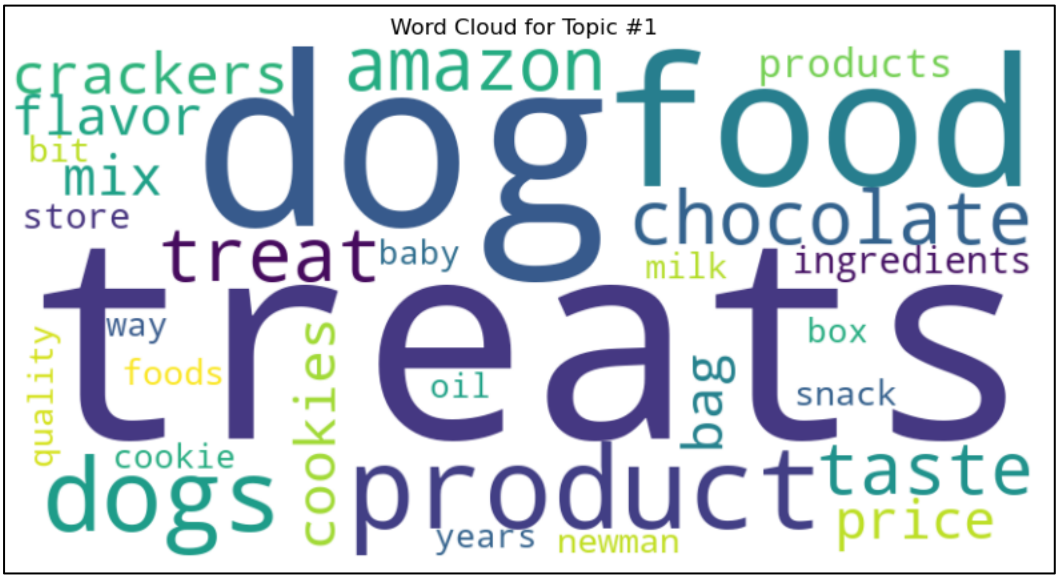


Figure A3: Word Cloud for Topic 2 (Top 30 Words)

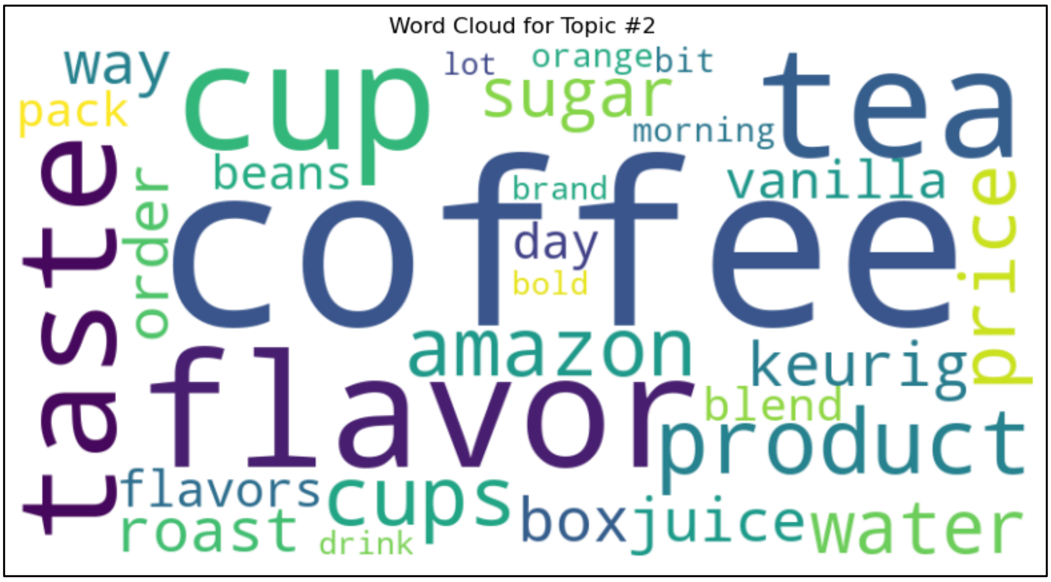


Figure A4: Word Cloud for Topic 1 (Top 10 Words)

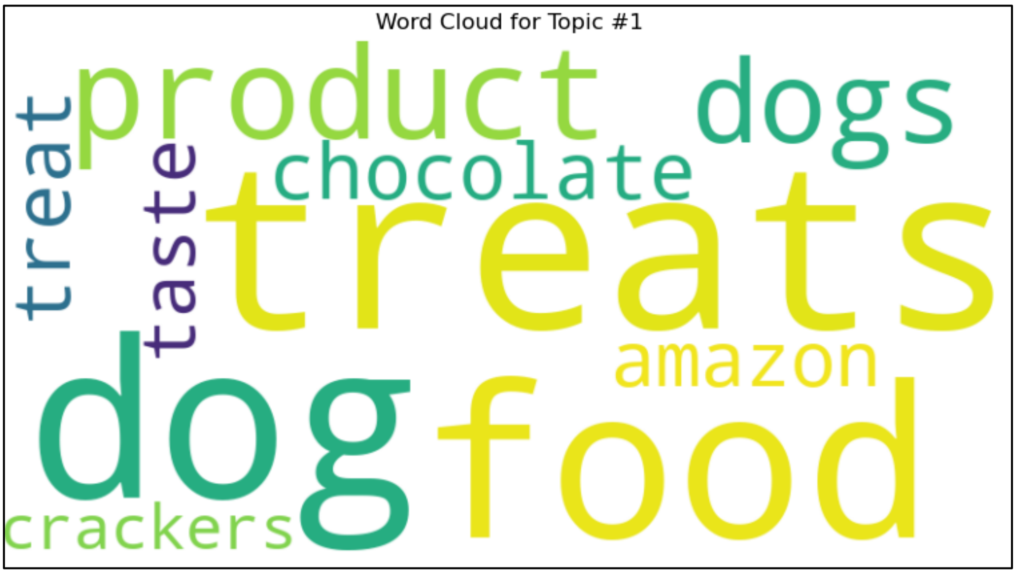


Figure A5: Word Cloud for Topic 2 (Top 10 Words)

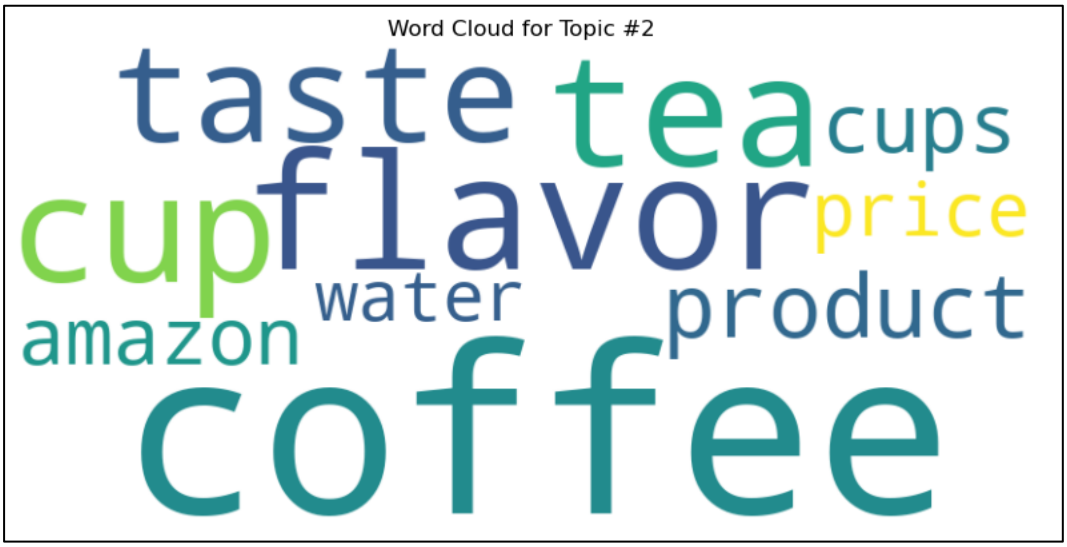
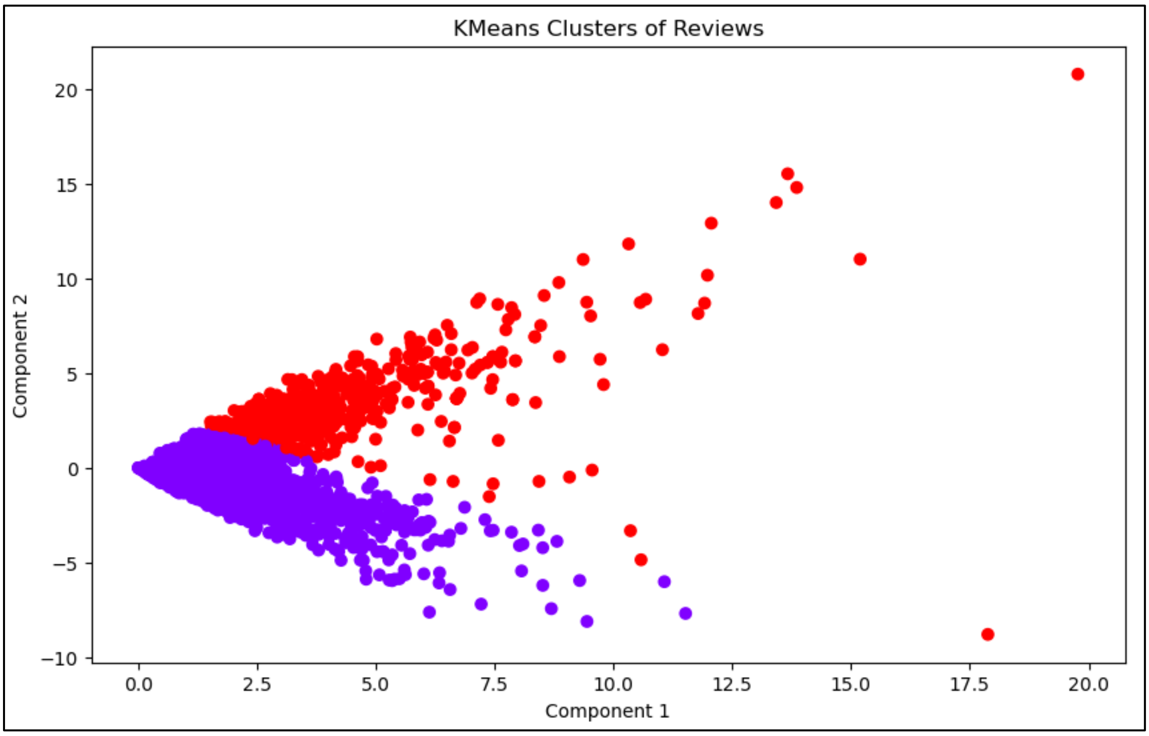


Figure A6: K-Means Clustering Scatter Plot

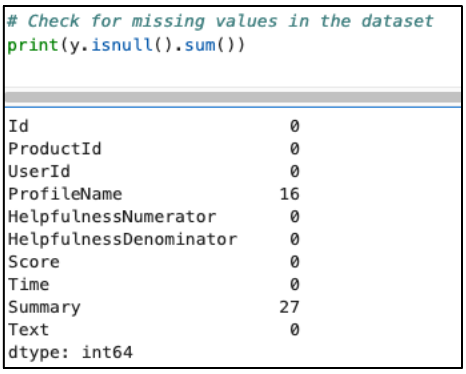


Appendix B: Table

Word	Count
Good	4800
Coffee	4700
Great	4000
Taste	3600
Product	3200
Flavor	3000
Love	2800
Really	2300
Amazon	2200
Food	2100
Tea	2100
Best	1900
Use	1800
Don	1800
Little	1800

Appendix C: Key Code Segments

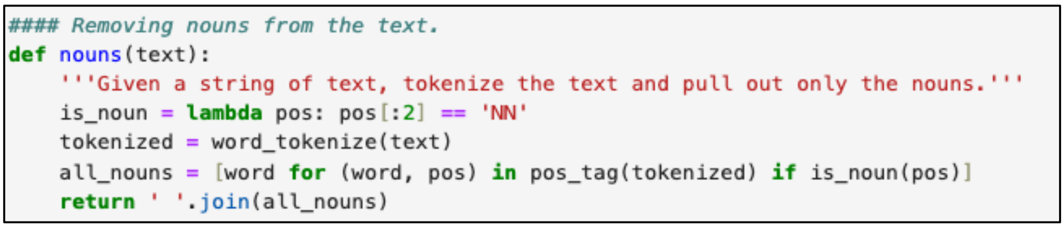
Code Segment C1: Checking for Missing Values



```
# Check for missing values in the dataset
print(y.isnull().sum())
```

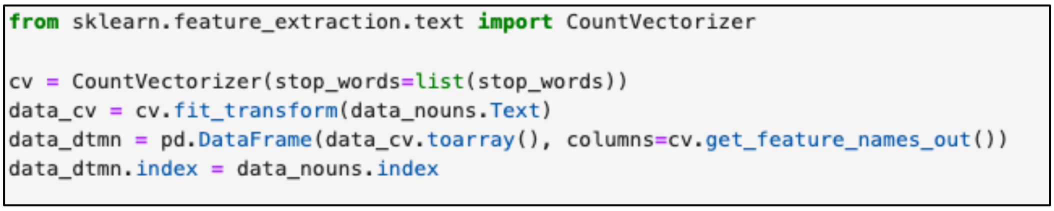
```
Id                         0
ProductId                  0
UserId                     0
ProfileName               16
HelpfulnessNumerator       0
HelpfulnessDenominator     0
Score                      0
Time                       0
Summary                   27
Text                       0
dtype: int64
```

Code Segment C2: Tokenization and Noun Extraction



```
#### Removing nouns from the text.
def nouns(text):
    '''Given a string of text, tokenize the text and pull out only the nouns.'''
    is_noun = lambda pos: pos[:2] == 'NN'
    tokenized = word_tokenize(text)
    all_nouns = [word for (word, pos) in pos_tag(tokenized) if is_noun(pos)]
    return ' '.join(all_nouns)
```

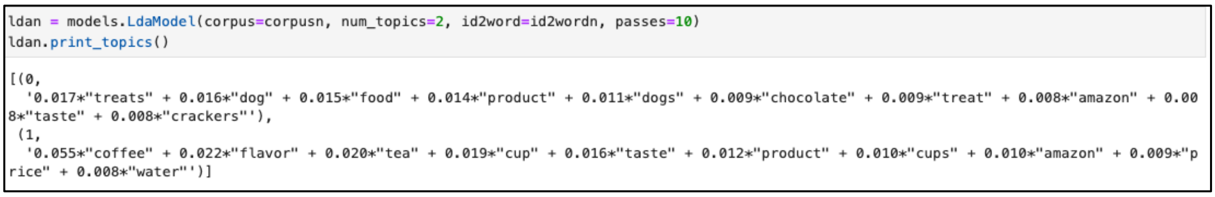
Code Segment C3: Document-Term Matrix Creation



```
from sklearn.feature_extraction.text import CountVectorizer

cv = CountVectorizer(stop_words=list(stop_words))
data_cv = cv.fit_transform(data_nouns.Text)
data_dtmn = pd.DataFrame(data_cv.toarray(), columns=cv.get_feature_names_out())
data_dtmn.index = data_nouns.index
```

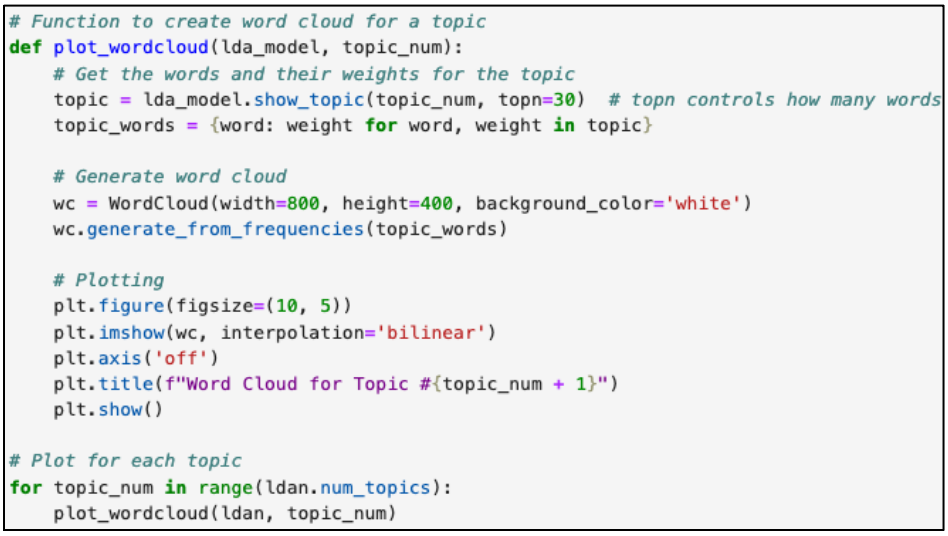
Code Segment C4: Topic Modeling (LDA)



```
ldan = models.LdaModel(corpus=corpusn, num_topics=2, id2word=id2wordn, passes=10)
ldan.print_topics()
```

```
[(0,
  '0.017*"treats" + 0.016*"dog" + 0.015*"food" + 0.014*"product" + 0.011*"dogs" + 0.009*"chocolate" + 0.009*"treat" + 0.008*"amazon" + 0.00
8*"taste" + 0.008*"crackers"'),
 (1,
  '0.055*"coffee" + 0.022*"flavor" + 0.020*"tea" + 0.019*"cup" + 0.016*"taste" + 0.012*"product" + 0.010*"cups" + 0.010*"amazon" + 0.009*"p
rice" + 0.008*"water"')]
```

Code Segment C5: Word Cloud Generation



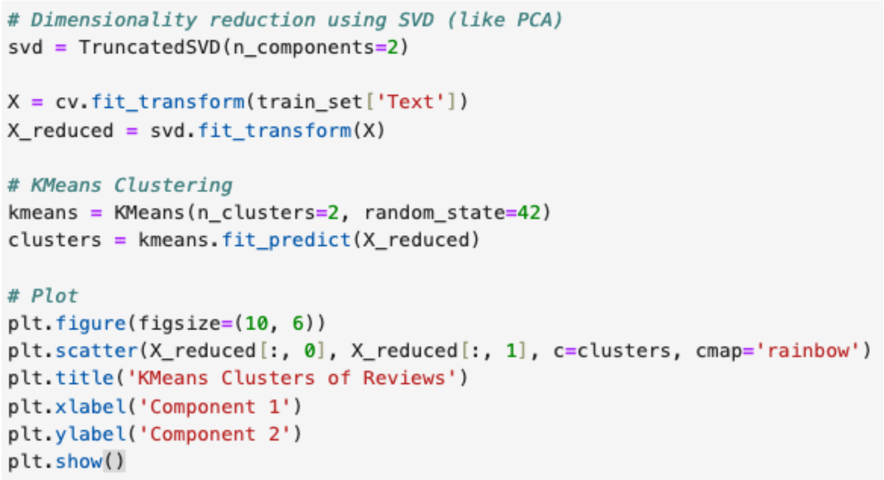
```python
# Function to create word cloud for a topic
def plot_wordcloud(lda_model, topic_num):
    # Get the words and their weights for the topic
    topic = lda_model.show_topic(topic_num, topn=30)  # topn controls how many words
    topic_words = {word: weight for word, weight in topic}

    # Generate word cloud
    wc = WordCloud(width=800, height=400, background_color='white')
    wc.generate_from_frequencies(topic_words)

    # Plotting
    plt.figure(figsize=(10, 5))
    plt.imshow(wc, interpolation='bilinear')
    plt.axis('off')
    plt.title(f"Word Cloud for Topic #{topic_num + 1}")
    plt.show()

# Plot for each topic
for topic_num in range(ldan.num_topics):
    plot_wordcloud(ldan, topic_num)
```

Code Segment C6: K-Means Clustering



```python
# Dimensionality reduction using SVD (like PCA)
svd = TruncatedSVD(n_components=2)

X = cv.fit_transform(train_set['Text'])
X_reduced = svd.fit_transform(X)

# KMeans Clustering
kmeans = KMeans(n_clusters=2, random_state=42)
clusters = kmeans.fit_predict(X_reduced)

# Plot
plt.figure(figsize=(10, 6))
plt.scatter(X_reduced[:, 0], X_reduced[:, 1], c=clusters, cmap='rainbow')
plt.title('KMeans Clusters of Reviews')
plt.xlabel('Component 1')
plt.ylabel('Component 2')
plt.show()
```