**CSE508 Information Retrieval**

**Winter 2024**

**Assignment – 3**

**Name-Ankur Tiwari**

**Rollno-MT23017**

**Problem Statement:**

**Pandas** is a powerful library for data manipulation and analysis. In the script, it’s used to create dataframes from JSON files, filter data, and perform various operations like merging dataframes.

**Pickle** is used for serializing and deserializing Python object structures. In the script, it’s employed to store dataframes in a binary format, which can be quickly loaded later.

**BeautifulSoup** is a library for pulling data out of HTML and XML files. It’s used in the script to check for and remove HTML tags from review texts, ensuring clean data for analysis.

Text Processing with NLTK

**NLTK (Natural Language Toolkit)** is a leading platform for building Python programs to work with human language data. It’s used for text preprocessing tasks like tokenization and lemmatization.

Machine Learning with Scikit-learn

**Scikit-learn** is a machine learning library that provides simple and efficient tools for data mining and data analysis. It’s used for feature extraction with TF-IDF, model training, and evaluation.

Visualization with Matplotlib:

**Matplotlib** is a plotting library for creating static, interactive, and animated visualizations in Python. The script uses it to display data distributions and results.

# Dataset Loading and Preprocessing:

# The Amazon Reviews dataset for the Electronics category was acquired, specifically the 5-core dataset for a small subset, to facilitate experimentation.

# The dataset was loaded into a pandas DataFrame using Python, and the product metadata was kept in a separate DataFrame for further analysis.

# Product Selection:

# Choose “headphones” as a product from metadata\_electronic using the title column

# Analysis of 'Headphones' Product:

# The dataset was refined to isolate entries specifically associated with the product category 'Headphones', in order to conduct a targeted analysis.

# The overall count of rows pertaining to the 'Headphones' product category was calculated.

# We used the inner join on asin to get the all information about product headphones

# Descriptive Statistics of the 'Headphones' Product:

# a. Number of Reviews:

# ● The total number of reviews available for the 'Headphones' product in the dataset is 411201.

# b. Average Rating Score:

# ● The average rating score for the 'Headphones' product is 4.112 out of 5.

# c. Number of Unique Products:

# ● There are 8064 unique 'Headphones' products represented in the dataset.

# d. Number of Good Ratings:

# ● Based on the threshold of >=3 for a 'Good' rating, the number of reviews considered as 'Good' for the 'Headphones' product is 353401.

# e. Number of Bad Ratings:

# ● Reviews with a rating score below 3 are considered as 'Bad'. The number of 'Bad' ratings for the 'Headphones' product is 57800

# f. Number of Reviews Corresponding to Each Rating

# 

# 5. Text Preprocessing for Reviews:

# a. Removing HTML Tags:

# ● HTML tags, if present in the review text, are removed using appropriate parsing techniques or libraries such as BeautifulSoup.

# b. Removing Accented Characters:

# ● Accented characters are replaced with their non-accented counterparts to standardize text and ensure consistency.

# c. Expanding Acronyms:

# ● Acronyms and abbreviations commonly used in reviews are expanded to their full forms for better understanding and readability.

# d. Removing Special Characters:

# ● Special characters, symbols, and punctuation marks are removed from the review text to focus on the essential content.

# e. Lemmatization:

# ● Words in the review text are lemmatized to reduce inflected forms to their base or dictionary form, thereby simplifying the text for analysis and interpretation.

# f. Text Normalization:

# ● Text normalization techniques such as lowercasing, removing stopwords, and tokenization may be applied to further clean and standardize the review text, making it suitable for analysis and modeling.

# 6.To extract relevant statistics, perform the following EDA .

# 

# Top 20 most reviewed brands in the category that you have chosen.

# 

# Top 20 least reviewed brands in the category you have chosen.



# c. Which is the most positively reviewed ‘Headphone’ ( Or for any other

# electronic product you have selected)

Sony

# d.Show the count of ratings for the product over 5 consecutive years.

# 

# e. Form a Word Cloud for ‘Good’ and ‘Bad’ ratings. Report the most

# commonly used words for positive and negative reviews by observing the

# good and bad word clouds.

# 

# f. Plot a pie chart for Distribution of Ratings vs. the No. of Reviews.

# 

# g. Report in which year the product got maximum reviews.

The product received the maximum reviews in the year 2016 with a

count of 98376.

# h. Which year has the highest number of Customers?

The year with the highest number of customers is 2016 with 75483

custmers.

# 7. Use a relevant feature engineering technique to model review text as Bag Words model, TF-IDF,Hashing Vectorizer or Word2Vec

# I used the hashing vectorizer for feature engineering

# 

# 8. add rating class

# 

# 9. we divide data in train and test in the ratio of 75:25.

# 10. Compare the performance of 5 Machine Learning based models on the basis of

# Precision, Recall, F-1 Score and Support for each of the 3 target classes distinctly.

# 

# 

# 11. c

# 

# 

**d.** MAE for N = 10: 0.8430072602540921

MAE for N = 20: 0.7649284900395303

MAE for N = 30: 0.8430072602540921

MAE for N = 40: 0.7649284900395303

MAE for N = 50: 0.8429590926994692

12. Top 10 products by user sum ratings:

1321. Product Name: Sony, Product ID: B004WODP20, Sum Rating: 13295

2464. Product Name: Sony, Product ID: B00BN0N0LW, Sum Rating: 13242

587. Product Name: iNassen, Product ID: B000WL6YY8, Sum Rating: 11456

3920. Product Name: Toysdone, Product ID: B00LP6CFEC, Sum Rating: 10454

4787. Product Name: XBRN, Product ID: B00STP86CW, Sum Rating: 10289

1843. Product Name: iNassen, Product ID: B007FHX9OK, Sum Rating: 10004

2872. Product Name: Fourcase, Product ID: B00EEHNNNG, Sum Rating: 9999

3572. Product Name: Etre Jeune, Product ID: B00JJ2C0S0, Sum Rating: 9780

69. Product Name: Belkin, Product ID: B000067RC4, Sum Rating: 8246

29. Product Name: Panasonic, Product ID: B00004T8R2, Sum Rating: 7392