INFORMATION RETRIEVAL

Assignment -3 Report

Anshuman Uniyal (MT 23018)

Multimodal Retrieval System

Libraries Used:

Pandas: a powerful Python library for data manipulation and analysis, facilitating tasks such as data cleaning, exploration, and transformation through its intuitive DataFrame structure.

Gzip: a compression utility in Python that enables efficient compression and decompression of files using the gzip format, commonly used for reducing file sizes during data transmission or storage.

Pickle: a Python module used for serializing and deserializing Python objects, facilitating data storage and retrieval in a compact binary format, commonly employed for object persistence and inter-process communication.

NLTK (Natural Language Toolkit): a Python library for natural language processing (NLP) tasks, offering tools for tokenization, stemming, tagging, parsing, and more, aiding in the analysis and understanding of human language.

Scikit-learn: a Python library for machine learning, providing a wide range of supervised and unsupervised learning algorithms, along with tools for model selection, evaluation, and data preprocessing, enabling developers to build predictive models and perform data analysis tasks efficiently.

Matplotlib: a Python library for creating static, interactive, and animated visualizations, offering a wide range of plotting functions and customization options, empowering users to generate publication-quality graphs and charts for data analysis and presentation purposes.

1. **Data Acquisition and Preparation:**
   i) Obtained the Amazon Reviews dataset, focusing on Electronics, specifically utilizing the 5-core dataset for a smaller subset to facilitate experimentation.
   ii) Employed Python to load the dataset into a pandas DataFrame, while keeping product metadata separate for potential use in subsequent analyses.
   iii) Carried out preprocessing tasks, including handling missing data, removing duplicates, and other data cleaning procedures to ensure the reliability and integrity of the data.

2. **Product Selection:**

   For the purpose of analysis, the product 'Headphones' was selected from Amazon Electronics dataset

3. **Analysis of 'Headphones' Product:**
   Filtered the dataset to isolate rows pertaining exclusively to the 'Headphones' product, facilitating targeted analysis.
   Determined the total count of rows associated with the 'Headphones' product to gauge dataset representation.

4. **Descriptive Statistics of the “Headphone’s Product” :**

   a. **Total Number of Reviews:** 412390
   b. **Average Rating Score:** 4.11266034578918
   c. **Number of Unique Headphones:** 8064
   d. **Number of Good Ratings:** 3554483
   e. **Number of Bad Ratings:** 57907

Text Preprocessing for Reviews:

a. Eliminating HTML Tags:

● Utilizing appropriate parsing techniques or libraries like BeautifulSoup, any HTML tags within the review text are eliminated**.**

b. Removing Diacritics:

● Diacritics are substituted with their non-diacritic equivalents to standardize text and ensure uniformity.

c. Expanding Abbreviations:

● Abbreviations and acronyms commonly found in reviews are expanded to their full forms to enhance comprehension and readability.

d. Eliminating Special Characters:

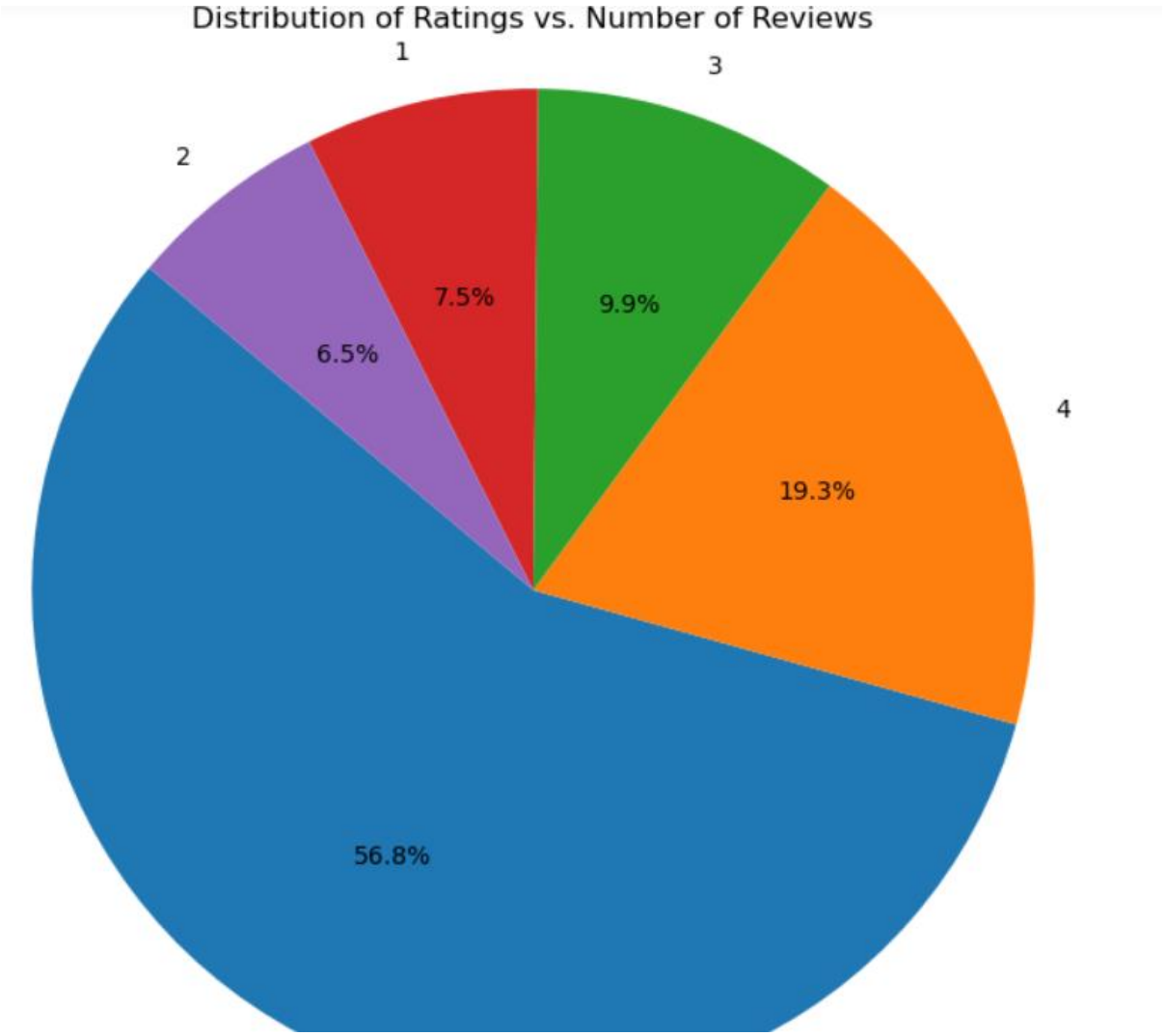
● Special characters, symbols, and punctuation marks are eradicated from the review text to emphasize the core content.

e. Lemmatization:

● Employing lemmatization, words within the review text are transformed into their base or dictionary forms, simplifying the text for analysis and interpretation.

f. Text Normalization:

- Text normalization methods like converting to lowercase, eliminating stopwords, and tokenization can be employed to further refine and standardize the review text, preparing it for analysis and modeling purposes.



Feature Engineering using Hashed Vector :

```
pd.DataFrame(hashed_matrix.toarray())
```

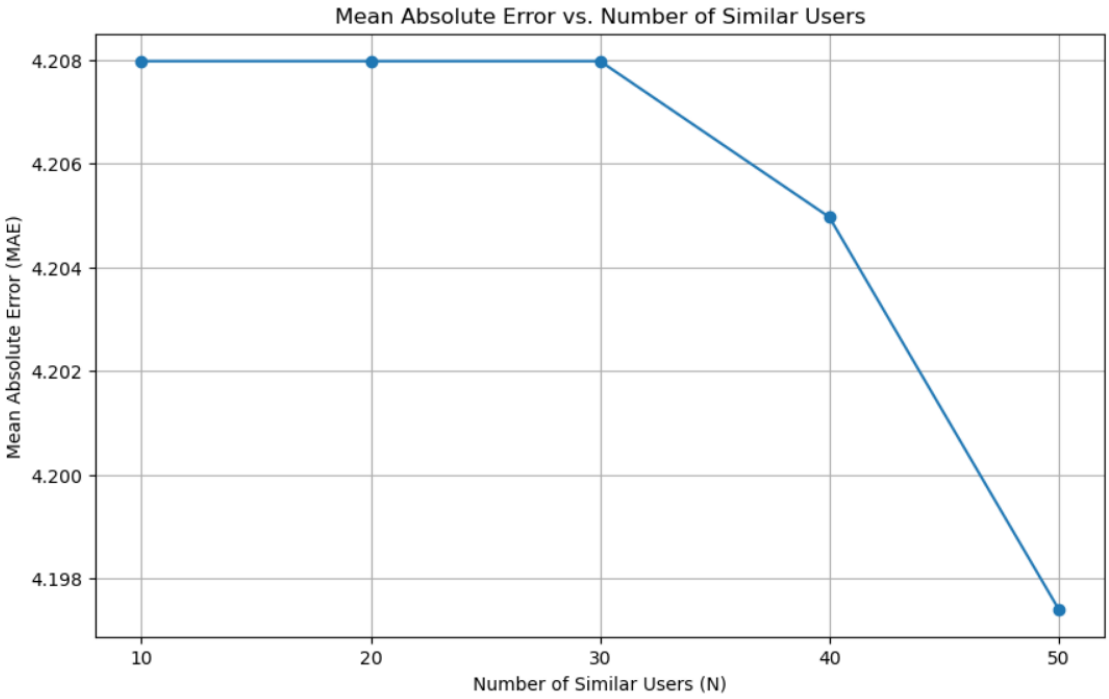
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	-0.333333	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	-0.149071	0.000000	-0.149071	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	-0.301511	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
412385	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	-0.064150	-0.064150	-0.064150	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.128300	0.0	0.0	0.0	0.0
412386	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	-0.162221	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.081111	0.0	0.0	0.0	0.0
412387	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	-0.036637	-0.036637	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
412388	0.0	0.0	0.0	0.069007	0.0	0.0	0.0	0.000000	0.276026	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.069007	0.0	0.0	0.0	0.0
412389	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.120386	-0.240772	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

412390 rows × 1024 columns

	Classifier	Class	Precision	Recall	F1 Score	Support
0	Random Forest	Class 1	0.756048	0.308803	0.438503	14472.0
1	Random Forest	Class 0	0.631751	0.037684	0.071126	10243.0
2	Random Forest	Class -1	0.803471	0.989960	0.887019	78383.0
3	Gradient Boosting	Class 1	0.729420	0.298784	0.423922	14472.0
4	Gradient Boosting	Class 0	0.485981	0.040613	0.074962	10243.0
5	Gradient Boosting	Class -1	0.803424	0.987217	0.885888	78383.0
6	Naive Bayes	Class 1	0.337753	0.499862	0.403121	14472.0
7	Naive Bayes	Class 0	0.149333	0.522601	0.232290	10243.0
8	Naive Bayes	Class -1	0.905441	0.529452	0.668186	78383.0
9	Logistic Regression	Class 1	0.673621	0.495301	0.570860	14472.0
10	Logistic Regression	Class 0	0.422528	0.080933	0.135846	10243.0
11	Logistic Regression	Class -1	0.838179	0.967697	0.898293	78383.0
12	Decision Tree	Class 1	0.407214	0.421227	0.414102	14472.0
13	Decision Tree	Class 0	0.196716	0.177780	0.186769	10243.0
14	Decision Tree	Class -1	0.839244	0.844469	0.841848	78383.0

Performance of 5 Machine learning Model

Collaborative Filtering:



User-item matrix

Top 10 products by User Sum Rating

```
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
```