

# **CSE508 Information Retrieval Winter 2024 Assignment-3 Report**

**AKASH KUMAR**

**MT23012**

**Date of Submission:-08-04-2024**

## **Table of Contents**

- 1. Introduction**
- 2. Objective:**
- 3. Data Acquisition and Preprocessing**
- 4. Descriptive Statistics**
- 5. Text Preprocessing**
- 6. Exploratory Data Analysis (EDA)**
- 7. Machine Learning Models**
- 8. Collaborative Filtering**
- 9. Results**
- 10. Conclusion**

## **1. Introduction**

- E-commerce Era: Abundance of choices for consumers.
- Product Recommendation Systems: Essential for guiding users.
- Personalized Suggestions: Tailored recommendations based on preferences.
- Enhanced User Experience: Improved shopping experience for users.
- Increased Customer Engagement: More engaged customers due to personalized recommendations.
- Customer Satisfaction: Higher satisfaction levels with personalized suggestions.

## 2. Objective

- Objective: Develop a product recommendation system.
- Methodology: Utilize collaborative filtering techniques.
- Data Source: Amazon review data.
- Collaborative Filtering: Analyze user interactions and preferences.
- Personalized Recommendations: Tailored suggestions for individual interests.
- Enhanced User Engagement: Increase user engagement through personalized recommendations.

## 3. Data Acquisition and Preprocessing

In this assignment, we have two datasets one is a 5-core dataset, and the other is meta\_data.

In this Assignment we used mouse

### Metadata Collection:

Extracted metadata from a gzip file containing information on electronics products.

Implemented a search function to filter metadata based on a user-defined word.

Preprocessed and stored the filtered metadata into a CSV file named 'meta\_data.csv.'

### Descriptive Statistics:

Analyzed a dataset containing customer reviews for a selected product (identified by ASIN).

Calculated the average rating score, total number of reviews, and counts of good and bad ratings (based on a 5-star rating system).

### Text Preprocessing:

Performed extensive text preprocessing on the 'title' column of the metadata and various text columns of the review data.

Preprocessing steps included converting text to lowercase, removing HTML tags, accents, and special characters, expanding acronyms, lemmatization, and normalization.

Handled missing values in the review data by filling NaN values with empty strings and dropping rows with empty strings or NaN values.

### Data Analysis:

Created a data frame for the review data after preprocessing.

Conducted text preprocessing on review text and reviewer information.

Provided insights into the length of the review data.

## 4. Descriptive Statistics

The information of the meta\_data is:-

```
Data columns (total 19 columns):
#   Column                Non-Null Count  Dtype
---  -
0   category               8698 non-null  object
1   tech1                  2364 non-null  object
2   description             8698 non-null  object
3   fit                     0 non-null     float64
4   title                   8698 non-null  object
5   also_buy                8698 non-null  object
6   tech2                   672 non-null   object
7   brand                   8632 non-null  object
8   feature                 8698 non-null  object
9   rank                    8698 non-null  object
10  also_view               8698 non-null  object
11  main_cat                8674 non-null  object
12  similar_item            2402 non-null  object
13  date                    8120 non-null  object
14  price                   2360 non-null  object
15  asin                    8698 non-null  object
16  imageURL                8698 non-null  object
17  imageURLHighRes         8698 non-null  object
18  details                 8697 non-null  object
dtypes: float64(1), object(18)
memory usage: 1.3+ MB
```

**category:** This column contains information about the category of the product.

**description:** Provides a description of the product.

**title:** Represents the title or name of the product.

**also\_buy:** Contains information about other products that customers also bought along with the main product.

**brand:** Indicates the brand of the product.

**feature:** Provides additional features or characteristics of the product.

**rank:** Gives the rank or position of the product in its category.  
**main\_cat:** Indicates the main category to which the product belongs.  
**similar\_item:** Contains information about similar items to the main product.  
**date:** Represents the date when the product information was recorded.  
**price:** Indicates the price of the product.  
**asin:** Represents the Amazon Standard Identification Number (ASIN) of the product.  
**imageURL:** Contains the URL of the product image.  
**imageURLHighRes:** Provides the URL of the high-resolution image of the product.  
**details:** Contains additional details or information about the product.

**Another dataset that is the Review Data set is:-**

```
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  -
0   overall                151378 non-null float64
1   verified                151378 non-null bool
2   reviewTime             151378 non-null datetime64[ns]
3   reviewerID             151378 non-null object
4   asin                   151378 non-null object
5   reviewerName           151378 non-null object
6   reviewText             151378 non-null object
7   summary                151378 non-null object
8   unixReviewTime         151378 non-null int64
9   vote                   12142 non-null object
10  style                   77574 non-null object
11  image                   1884 non-null object
12  year                   151378 non-null int32
dtypes: bool(1), datetime64[ns](1), float64(1), int32(1), int64(1), object(8)
memory usage: 14.6+ MB
```

The DataFrame contains 151,378 entries and 13 columns. It includes information about reviews, such as overall rating, verification status, reviewer details, review text, product ASIN, review timestamp, and metadata like review votes and associated images.

## 5. Text Preprocessing

**Removing HTML Tags:**

Eliminates HTML tags from text data.

Enhances text analysis by removing irrelevant HTML markup.

**Removing Accented Characters:**

Substitute accented characters with their non-accented equivalents.

Standardizes text, ensuring consistency during analysis.

**Expanding Acronyms:**

Replace acronyms with their full phrases.

Enhances text clarity and maintains consistency.

**Removing Special Characters:**

Deletes punctuation marks, symbols, and non-alphanumeric characters.

Focuses on meaningful content by eliminating non-semantic characters.

**Lemmatization:**

Reduces words to their base forms (lemmas).

Identifies canonical word forms, considering context and grammar.

**Text Normalization:**

Standardizes text by converting it to a consistent format.

Lowercase all text remove extra spaces, and handle contractions uniformly.

Ensures consistency across writing styles and improves NLP accuracy.

## 6. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial preliminary step in understanding the characteristics and structure of a dataset. In the context of building a product recommendation system based on Amazon review data, EDA provides valuable insights into various aspects of the dataset related to the chosen product category, which in this case is '**mouse**'. Here's a detailed explanation of each component of EDA:

**Descriptive Statistics:**

EDA involves calculating descriptive statistics such as count, mean, median, standard deviation, minimum, and maximum values for relevant variables associated with the 'Headphones' category. These statistics summarize key metrics, including the number of reviews, average rating score, and other pertinent information.

```

Total number of reviews: 151378
Average Rating Score: 4.172719946095206
Number of Good Ratings: 117762
Number of Bad Ratings: 33616
Number of ratings corresponding to 5: 91104
Number of ratings corresponding to 4: 26658
Number of ratings corresponding to 3: 13445
Number of ratings corresponding to 2: 9000
Number of ratings corresponding to 1: 11171

```

### Top-Reviewed Brands:

By analyzing the dataset, EDA identifies the top-reviewed brands within the 'mouse' category. This analysis involves determining which brands have received the highest number of reviews or have the most significant presence in the dataset. Understanding the popularity of different brands helps in assessing brand reputation and customer preferences.

#### Top-Review 20 brand

```

brand
Logitech      548
Microsoft    283
TOP CASE      250
HP            189
Dell          155
Generic       149
Kensington    137
Road Mice     91
Gear Head     85
Targus        85
Cooper Cases  76
Rapoo         74
Amsahr        70
iHome         63
Best Deal     63
Lenovo        61
Micro Innovations 57
Belkin        56
Sony          52
Perman        46
Name: count, dtype: int64

```

#### Least 20 review brand

```

brand
Linkskey      1
Suppion       1
AplusElek     1
As Seen On TV 1
kayond        1
Advanta - Mousemats 1
Hayand        1
DroidBOX      1
UNISEN LIMITED 1
shenzhen vership Co. LTD 1
mele          1
Epower Mall   1
Jastore       1
Work Smart    1
Acer Gateway  1
Meridian Point 1
MAXAH         1
635 STARS UNITED 1
Office Depot  1
by\n      \n      Taonology 1
Name: count, dtype: int64

```

[illegible]

**This shows the distribution of the ratings**

A pie chart showing the distribution of responses for the statement 'The government should do more to protect the environment'. The chart is divided into five segments, each labeled with a number and a percentage. The segments are: 5 (60.2%, blue), 4 (17.6%, orange), 3 (8.9%, green), 1 (7.4%, red), and 2 (5.9%, purple).

Response	Percentage
5	60.2%
4	17.6%
3	8.9%
1	7.4%
2	5.9%

The classification report reveals that the Logistic Regression and Linear SVC models exhibit strong performance across all sentiment categories, with high precision, recall, and F1-score values. While the SGD Classifier and Passive Aggressive Classifier also show promising results, the Naive Bayes model demonstrates comparatively lower precision and recall for negative sentiment classification.

*** Linear SVC ***					
	precision	recall	f1-score	support	
1.0	0.52	0.58	0.55	2855	
2.0	0.38	0.08	0.13	2270	
3.0	0.41	0.12	0.18	3472	
4.0	0.43	0.23	0.30	6624	
5.0	0.73	0.95	0.82	22579	
accuracy			0.67	37800	
macro avg	0.50	0.39	0.40	37800	
weighted avg	0.61	0.67	0.61	37800	

*** SGD Classifier ***					
	precision	recall	f1-score	support	
1.0	0.50	0.53	0.51	2855	
2.0	0.29	0.06	0.10	2270	
3.0	0.42	0.07	0.12	3472	
4.0	0.44	0.09	0.15	6624	
5.0	0.69	0.98	0.81	22579	
accuracy			0.65	37800	
macro avg	0.47	0.35	0.34	37800	
weighted avg	0.58	0.65	0.56	37800	

*** Logistic Regression ***					
	precision	recall	f1-score	support	
1.0	0.54	0.55	0.55	2855	
2.0	0.35	0.16	0.22	2270	
3.0	0.40	0.21	0.27	3472	
4.0	0.46	0.25	0.33	6624	
5.0	0.75	0.94	0.83	22579	
accuracy			0.68	37800	
macro avg	0.50	0.42	0.44	37800	
weighted avg	0.62	0.68	0.63	37800	


*** Linear SVC ***					
	precision	recall	f1-score	support	
1.0	0.52	0.58	0.55	2855	
2.0	0.38	0.08	0.13	2270	
3.0	0.41	0.12	0.18	3472	
4.0	0.43	0.23	0.30	6624	
5.0	0.73	0.95	0.82	22579	
accuracy			0.67	37800	
macro avg	0.50	0.39	0.40	37800	
weighted avg	0.61	0.67	0.61	37800	

*** Naive Bayes ***					
	precision	recall	f1-score	support	
1.0	0.63	0.28	0.39	2855	
2.0	0.32	0.01	0.02	2270	
3.0	0.36	0.04	0.08	3472	
4.0	0.40	0.09	0.15	6624	
5.0	0.64	0.99	0.78	22579	
accuracy			0.63	37800	
macro avg	0.47	0.28	0.28	37800	
weighted avg	0.55	0.63	0.53	37800	

## 8. Collaborative Filtering :

- User-item rating matrix was created.
- Ratings were normalized using min-max scaling.
- User-user and item-item recommender systems were implemented using cosine similarity.
- Mean Absolute Error (MAE) was calculated for different values of N (number of similar users or items) and K (number of folds in k-fold validation).

## 9. Result:

The result of collaborative filtering is on the basis of some user ID and term ID is 



productId	9803751263	B00000J1V7	B00000K4LF	B00002JXBI	B00004S9AK
userId					
A100UD67AHF0DS	0.016116	0.000000	0.0	0.0	0.000000
A10LWFKVC21F82	0.000000	0.121039	0.0	0.0	0.000000
A1007THJ2020AG	0.015980	0.000000	0.0	0.0	0.000000
A10SE0U42ABS9S	0.000000	0.000000	0.0	0.0	0.000000
A10Y058K7B96C6	0.000000	0.000000	0.0	0.0	0.051663
productId	B00004W3YK	B0000511E5	B0000511L1	B000052WM4	B00005853X
userId					
A100UD67AHF0DS	0.0	0.0	0.0	0.000000	0.002227
A10LWFKVC21F82	0.0	0.0	0.0	0.034372	0.000000
A1007THJ2020AG	0.0	0.0	0.0	0.000000	0.000000
A10SE0U42ABS9S	0.0	0.0	0.0	0.022762	0.000000
A10Y058K7B96C6	0.0	0.0	0.0	0.048658	0.000000
productId	...	B01G230V4S	B01GDCZJPE	B01GE5T59G	B01GI93F76 \
userId	...				
A100UD67AHF0DS	...	0.009251	0.008313	0.0	0.662242
A10LWFKVC21F82	...	0.000000	0.000000	0.0	0.186062
A1007THJ2020AG	...	0.009173	0.008242	0.0	0.697930
A10SE0U42ABS9S	...	0.022004	0.000000	0.0	0.448542

## 10. Conclusion:

- The product recommendation system was successfully implemented using collaborative filtering techniques.
- Performance metrics for machine learning models and recommender systems were evaluated.
- The top 10 products by user sum ratings were reported.

