

Data-Driven Recommendation

with Amazon Products

Reet Nandy (rn2528)
Jacqueline Ji (xj235)
Yu-Yuan Chang (yc6549)
Tzu-Yi Chang (tc3930)
Zheng-Chen Yao (zy2876)

A. Introduction

In today's e-commerce landscape, consumers are often overwhelmed by the vast selection of products available on platforms like Amazon. To address this challenge, our project develops an **innovative Amazon product recommendation system** that not only simplifies the decision-making process but also enhances the shopping experience by delivering personalized product suggestions.

Our project aims to refine the approach to product recommendations by employing a **dual-analysis strategy that leverages both product metadata and user reviews**. By analyzing these two data streams, our system provides a more nuanced understanding of what products are truly relevant to the users, thereby increasing the accuracy and relevance of the recommendations.

Innovative Approach:

- **Big Data Challenge:**

Our recommendation system is built on a foundation of massive datasets collected by the McAuley Lab from UCSD, comprising approximately **571.54 million reviews across 48 million items in 33 different categories** (see Figure 1 and Figure 2). The sheer volume and variety of this data present a significant big data challenge, necessitating robust solutions for efficient data processing and analysis.














































 meta_All_Beauty.jsonl 	213 MB  LFS 
 meta_Amazon_Fashion.jsonl 	1.42 GB  LFS 
 meta_Appliances.jsonl 	285 MB  LFS 
 meta_Arts_Crafts_and_Sewing.jsonl 	2.17 GB  LFS 
 meta_Automotive.jsonl	5.35 GB  LFS 
 meta_Baby_Products.jsonl 	691 MB  LFS 
 meta_Beauty_and_Personal_Care.jsonl 	2.84 GB  LFS 
 meta_Books.jsonl	14.7 GB  LFS 
 meta_CDs_and_Vinyl.jsonl 	949 MB  LFS 
 meta_Cell_Phones_and_Accessories.jsonl 	4.02 GB  LFS 
 meta_Clothing_Shoes_and_Jewelry.jsonl	18 GB  LFS 
 meta_Digital_Music.jsonl 	67.1 MB  LFS 

Figure 1












































 All_Beauty.jsonl 	327 MB  LFS 
 Amazon_Fashion.jsonl 	1.05 GB  LFS 
 Appliances.jsonl 	929 MB  LFS 
 Arts_Crafts_and_Sewing.jsonl 	3.95 GB  LFS 
 Automotive.jsonl	8.73 GB  LFS 
 Baby_Products.jsonl 	2.95 GB  LFS 
 Beauty_and_Personal_Care.jsonl	11 GB  LFS 
 Books.jsonl	20.1 GB  LFS 
 CDs_and_Vinyl.jsonl 	3.29 GB  LFS 
 Cell_Phones_and_Accessories.jsonl	9.34 GB  LFS 
 Clothing_Shoes_and_Jewelry.jsonl	27.8 GB  LFS 
 Digital_Music.jsonl 	78.8 MB  LFS 

Figure 2

- **Two-Stage Analysis:**

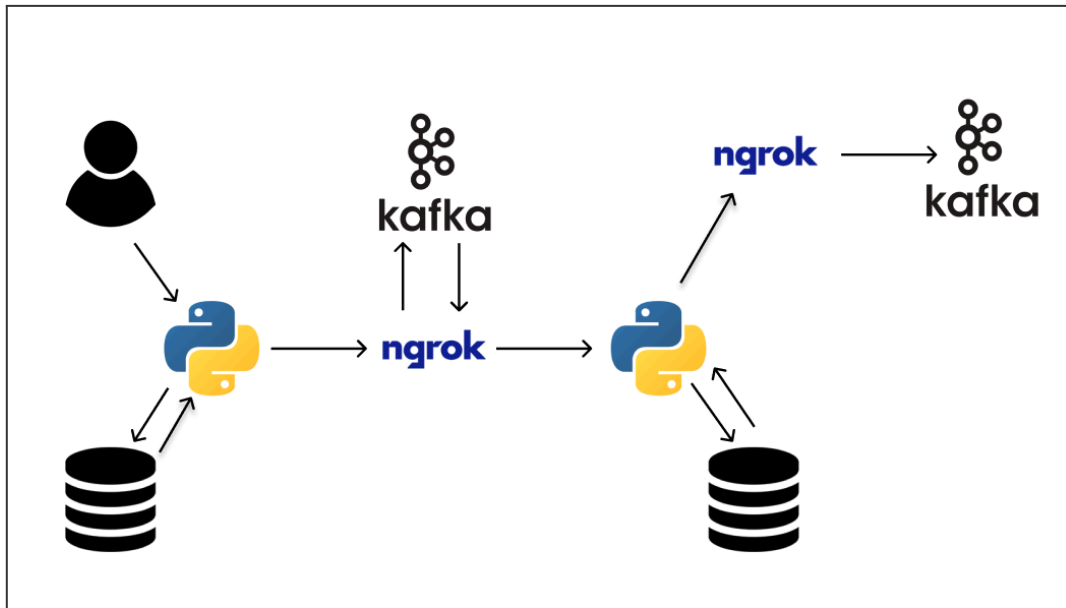


Figure 3

- **Metadata Analysis:** The first stage involves processing product metadata to identify features such as category, brand, price, and technical specifications. This stage aims to quickly narrow down the vast array of products to the top 50 items most similar to the user's search query or interest.
- **Review Analysis:** The second stage focuses on analyzing user reviews and ratings for these 50 products. It assesses sentiment, user satisfaction, and popularity to rank these products, ultimately selecting the top 5 that best meet the user criteria.
- **Real-Time Data Handling with Kafka:**
Middleware Integration: Apache Kafka is integrated as a core component of our architecture to manage real-time data flow between the metadata analysis and review analysis stages. Kafka ensures that the data pipeline is not only robust and fault-tolerant but also capable of handling high throughput and data volumes inherent to a major e-commerce platform like Amazon.
- **Kafka Topics for Segmented Data Processing:**
 We utilize separate Kafka topics for each stage of the analysis to ensure that the system remains organized and efficient. This separation allows for independent scaling and optimization of each part of the data processing workflow, enhancing the system's overall performance.

- **Benefits of the Dual-Stage Recommendation System:**

- **Enhanced Accuracy:** By synthesizing insights from both product metadata and user reviews, the recommendations are grounded in both technical product details and actual user experiences.
- **Increased Relevance:** Real-time analysis ensures that the recommendations reflect the latest trends and consumer preferences, providing users with the most relevant and timely information.
- **Improved User Experience:** A more targeted approach reduces the cognitive load on users, making their shopping experience more enjoyable and efficient.

B. Similarity Analysis

To generate the similarity out of user input, we did a couple steps of similarity analysis. We utilized tokenization, stop word removal, hashing, normalization, LSH, and similarity calculations to find the cosine similarity between user input and metadata. The metadata comprises various features with different data structures and types. We extracted five columns from a total of fourteen that contain useful information for building our model: title, features, description, categories, and details (see Figure 4).

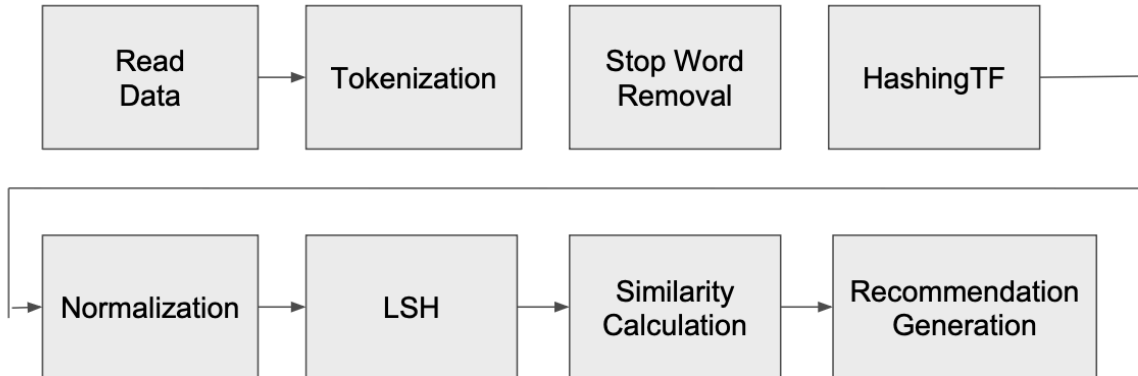


Figure 4

After extracting features, we tokenized the data and removed unnecessary stop words from the metadata. We then performed a word count for each dataset and mapped it to an index using a hashing function. Figures 5 and 6 demonstrate our workflow.

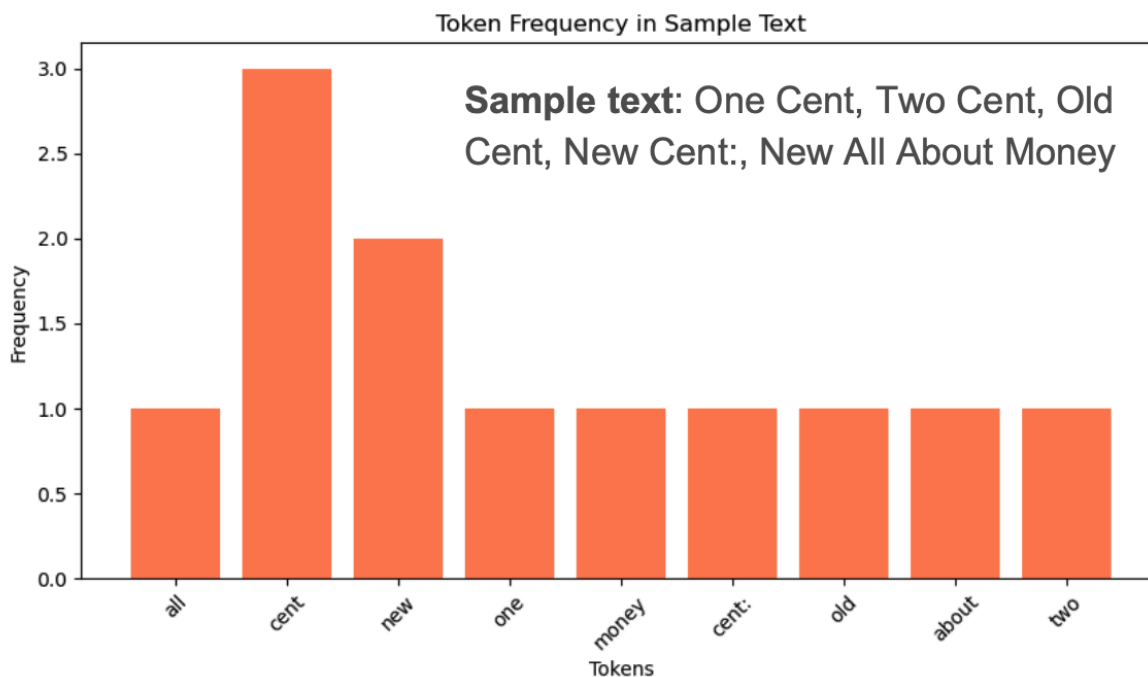


Figure 5

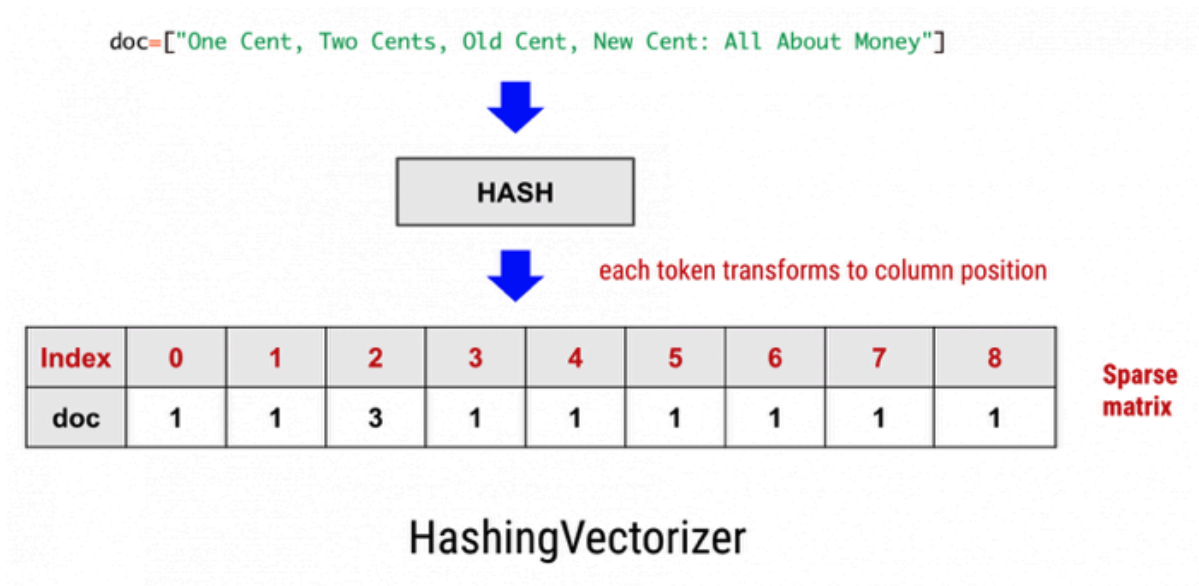


Figure 6

After normalization, we introduce the locality-sensitive hashing instead of naively comparing all pairs of items in the set. With the LSH approach, the numbers of comparison can be significantly reduced. All the items will be hashed into the buckets and the similar items are likely to be hashed into the same bucket (Figure 7). The distances between the candidate pairs in the high-dimensional space represent the similarity between items. This approach is particularly useful in the large dataset. It provides the approximation of similarity between items in an efficient way. The accuracy and computational efficiency are related to the number of hash functions. We can set the number of hash functions using in approximation. If

not set, Apache Spark's MLib has numHashTable = 1 as its default value. The more hash functions could be more accurate at the more cost of computational resources. The fewer hash functions could be more efficient but less accurate. Another important factor could affect the behavior of LSH is the bucketLength. The shorter bucket results in higher sensitivity to small differences between two items but potentially increases the false negatives; the longer bucket leads to less sensitivity to small differences and potentially increases the false positives. In the practical setting in our project, we set the bucketLength as 0.5 to compute the similarity. We could set bucketLength as 0.25 to improve the accuracy with the longer computation. We also set the threshold of 5 for distance metric to ensure the items pushed to users are under a certain level of similarity. In summary, through fine-tuning above parameters, we can control the accuracy, efficiency and sensitivity of LSH to fit our business needs.

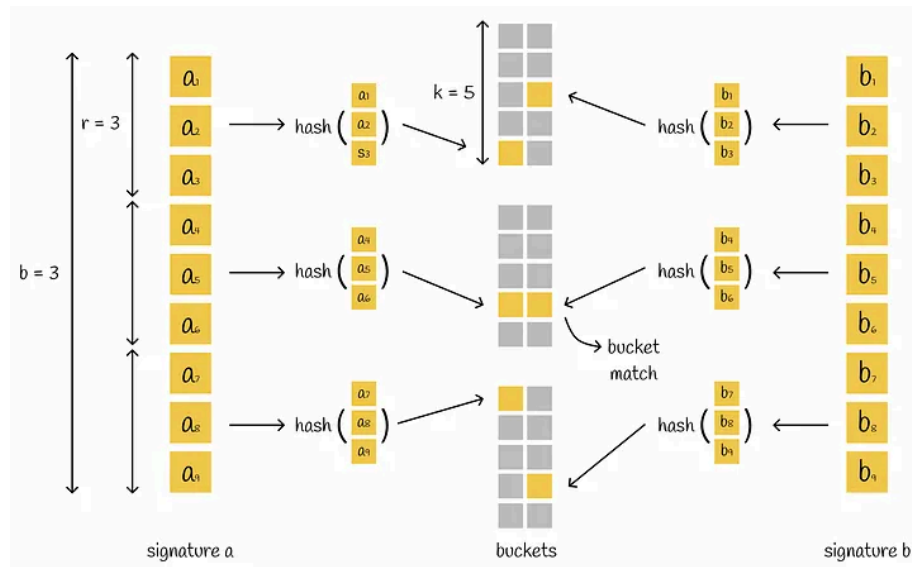


Figure 7

After all the preprocessing, we prepared a dataset to calculate cosine similarity. We used the cosine similarity model to measure the similarity between a user's desired item and other items. Cosine similarity is a metric that determines how similar documents are, irrespective of their size. (See Figures 8 and 9). The higher the cosine similarity score, the greater the similarity. We can observe that the first row has similarity 1, which indicates that this is the user input item. From the second row, the similarity scores start at 0.601 and decrease to 0.502 for the last item. (See Figure 9).



Figure 8

recommended_item	recommended_item_title	similarity
B08R39MRDW	DouBCQ Women's Pa...	1.0
B08R39X39B	DouBCQ Women's Pa...	0.6019333249445532
B08R39Z847	DouBCQ Women's Pa...	0.5480119797307619
B085CFKBCK	DouBCQ Women's Su...	0.5305934868767245
B085CF5G6K	DouBCQ Women's Su...	0.5270542710570822
B08R395V5M	DouBCQ Women's Pa...	0.5208776500283604
B08R3CNH89	DouBCQ Women's Pa...	0.5173643032421159
B08S7D6333	TICOSA 2021 New W...	0.5162870604758762
B085CFP9MH	DouBCQ Women's Ca...	0.5158101700422401
B08R3C12K7	DouBCQ Women's Pa...	0.5052786525618881

Figure 9

From the cosine similarity matrix, we can see that the cosine similarity model has accurately identified a related item based on user input. The similar item may be the same item in different sizes, colors, or styles, which makes sense for promoting this item to the user. (See Figure 10).

User Input:

asin: B08R39MRDW

**Similar Items:**

asin: B08R39X39B

asin: B08R39Z847

Similar pants with different style, color and size.



Figure 10

C. Middleware - Kafka and ngrok Integration

In our recommendation system, Apache Kafka plays a crucial role as the backbone for data streaming between the metadata and review analysis processes. To securely expose our Kafka broker to external systems, we use ngrok, a reverse proxy tool that creates a secure tunnel from a public endpoint to a locally running service

- **Kafka's Role and Configuration:**

- **Middleware for Streaming Data:**

Apache Kafka serves as the middleware that efficiently manages the streaming of data between different components of our system. It ensures that there is a continuous, real-time flow of data from the metadata analysis phase to the review analysis phase.

- **Topics and Partitions:**

- We use two separate Kafka topics to streamline the data flow:
- Topic-1 for streaming the top 50 similar products identified from the metadata.
- Topic-2 for the final top 5 recommendations derived from the review analysis.
- Each topic is partitioned to distribute the load evenly across the Kafka cluster, enhancing the system's scalability and reliability.

- **Producer-Consumer Model:**

- The metadata analysis script acts as the producer, sending data to Topic-1.
- The review analysis script acts as the consumer, receiving data from Topic-1 and upon processing, sends the refined data to Topic-2.

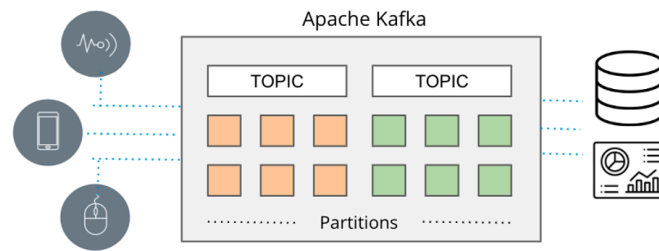


Figure 11

- **Integration of ngrok:**
 - **Secure Remote Access:**
ngrok is configured to create a secure TCP tunnel to our Kafka broker, which is typically only accessible within a private network. This tunnel allows external consumers, such as the review analysis script running on different servers or in different environments, to access the Kafka topics securely.
 - **Setup and Benefits:**
 - Setting up ngrok involves running it on the same machine as the Kafka broker and specifying the port that Kafka listens on. ngrok then provides a public URL (e.g., `tcp://2.tcp.ngrok.io:17851`), which can be used to access the broker from anywhere.
 - This setup not only enhances the security of our data transactions by encrypting the tunnel but also adds flexibility in how and where our analysis scripts can be deployed.
- **Benefits of Using Kafka with ngrok:**
 - **Scalability:** Kafka's ability to handle large volumes of data and ngrok's capability to expose these services to a wider network mean that our system can scale up as the amount of data or the number of users increases without significant reconfiguration.
 - **Reliability:** Kafka's built-in fault tolerance, coupled with ngrok's stable tunneling service, ensures that our data pipelines are robust against system failures or network issues.
 - **Real-Time Data Processing:** The combination of Kafka and ngrok allows our system to process and transmit data in real-time, which is crucial for maintaining the relevance and accuracy of the product recommendations provided to the users.

By integrating Kafka and ngrok, our Amazon product recommendation system achieves high levels of efficiency, security, and scalability. This setup not only supports the real-time processing requirements of our dual-stage analysis but also ensures that our system can be accessed and managed from diverse geographic and network environments.

D. Data Analysis

In this part of our project, we focus on utilizing the Amazon review dataset to derive meaningful insights and generate practical recommendations. The initial step in our analysis involved a thorough cleaning of the dataset. We implemented two main filters:

1. Removal of Redundant Reviews: We made sure our dataset contains only unique entries by eliminating any repetitions.
2. Elimination of Duplicate Reviews: We identified and removed reviews that not only shared identical text and timestamps but also featured user IDs with similar prefixes, differing only in their suffixes. An example of how similar user IDs look can be seen in the following figure (see Figure 12).

asin	text	count_reviews	distinct_users	helpful_vote	images	parent_asin	rating	timestamp	title	user_id
C9N9OM	Works great!	2	2	0	[]	B001C9N9OM	4.0	1441456721000	Four Stars	AF7R3PGUQF4BQYJUA3UCVMRN4AUQ
C9N9OM	Works great!	2	2	0	[]	B001C9N9OM	5.0	1454211680000	Five Stars	AELFACQZMFDNPMHLH2QZQWBC23LQA
'5M6FM8	This is a good quality butcher's coat. It's we...	2	2	1	[]	B00O97232I	4.0	1561065579310	Good quality butcher's coat	AEIIRIHLIYKQGI7ZOCIJTRDF5NPQ
'5M6FM8	This is a good quality butcher's coat. It's we...	2	2	1	[]	B00O97232I	4.0	1561065579310	Good quality butcher's coat	AEIIRIHLIYKQGI7ZOCIJTRDF5NPQ_2_2_1

Figure 12

After this thorough cleaning process, we ended up with a robust dataset comprising 2,475,451 reviews.

A critical aspect of our dataset is the distinction between verified and unverified reviews. Unverified reviews, which constitute about 6.5% of the dataset (see Figure 13), are characterized by a 'False' value in the 'verified_purchase' column (see Figure 14). Such reviews potentially originate from users who did not purchase the product through Amazon or whose purchases could not be independently verified by Amazon. While these reviews can offer valuable insights, they may also be less reliable.

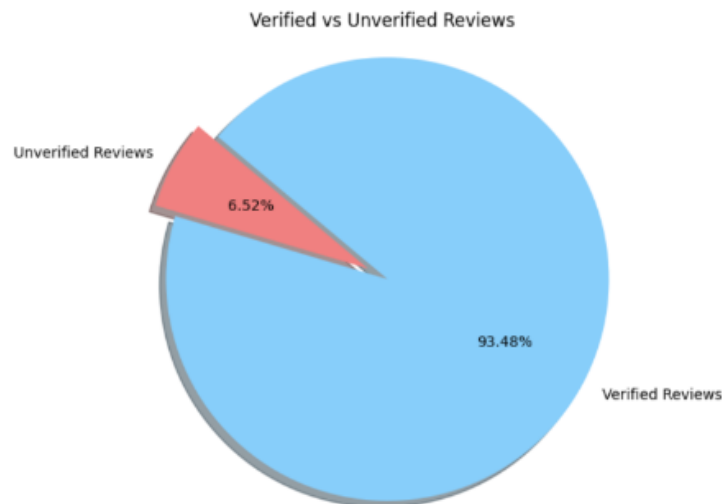


Figure 13

	asin	helpful_vote	images	parent_asin	rating	text	timestamp	title	user_id	verified_purchase
0	B01H75WEU4	10	[(IMAGE, https://images-na.ssl-images-amazon.c...	B01H75WEU4	5.0		1474928024000	buy it	AHIOOQUVUTIDSEZT6PQWPFZ3Q4CQ	True
1	B0007YDCK4	8	[]	B0007YDCK4	1.0	(...) This is the worst toy I've ever had...	1137262973000	The worst toy	AGCOE3TE34NGP43AXUQG3KZ7TAOA	False

Figure 14

The question arises: should we incorporate these unverified reviews in our subsequent analysis? To address this question, we conducted a thorough investigation into both the 'unverified' and 'verified' reviews.

We've determined that the unverified reviews should indeed be considered for several compelling reasons.

Firstly, as the following two histograms showed, the rating distributions of both verified and unverified reviews are remarkably similar (see Figure 15). If unverified reviews were predominantly malicious or biased, we would anticipate a noticeable difference in these distributions. For instance, extreme ratings might be more prevalent.

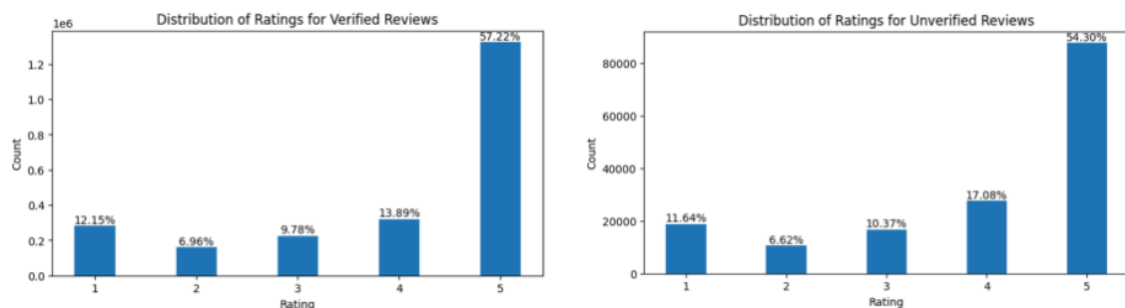


Figure 15

Secondly, our examination of the 'helpful_vote' column (see Figure 16) reveals that people find unverified reviews to be equally helpful as verified reviews (see Figure 17). This suggests that unverified reviews are valued by users and hold significance in guiding purchasing decisions, further supporting their inclusion in our analysis.

	asin	helpful_vote	images	parent_asin	rating	text	timestamp	title	user_id	verified_purchase
0	B01H75WEU4	10	[(IMAGE, https://images-na.ssl-images-amazon.c...	B01H75WEU4	5.0		1474928024000	buy it	AHIOOQUVUTIDSEZT6PQWPFZ3Q4CQ	True
1	B0007YDCK4	8	[]	B0007YDCK4	1.0	(...) This is the worst toy I've ever had...	1137262973000	The worst toy	AGCOE3TE34NGP43AXUQG3KZ7TAOA	False

Figure 16

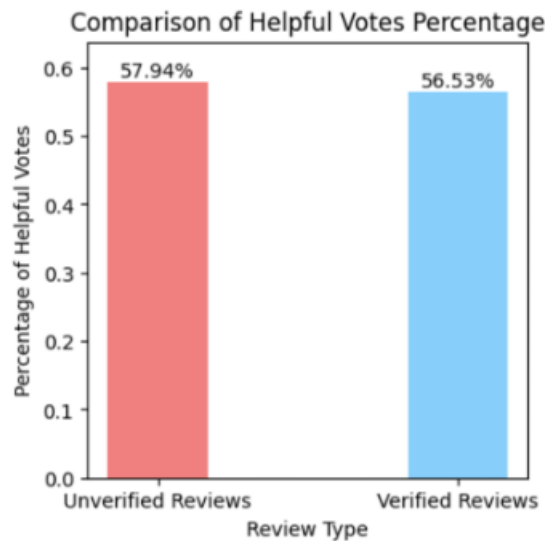


Figure 17

Additionally, we compared the number of products reviewed in verified and unverified reviews (see Figure 18). We discovered that 51,694 products were exclusively reviewed in unverified reviews. Removing unverified reviews would result in losing valuable information about these products.

Moreover, we found around 54,513 common products that received feedback from both verified and unverified reviews.

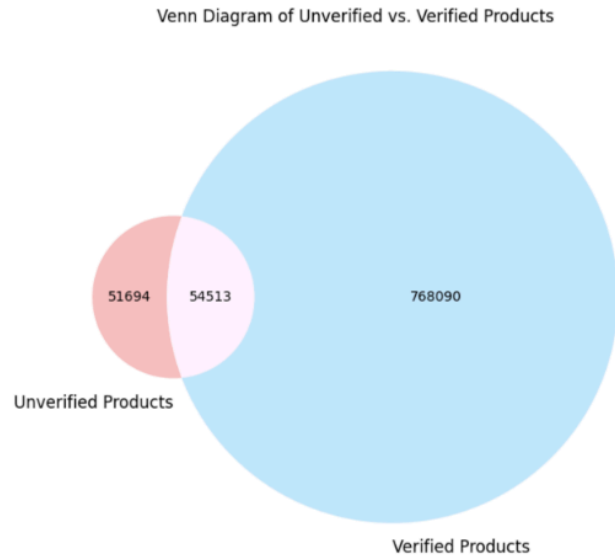


Figure 18

In our analysis, we initially observed that the rating distributions between verified and unverified reviews are remarkably similar. This similarity might lead one to assume that both types of reviews offer consistent insights for common products. Essentially, one could speculate that unverified reviews might be redundant when verified ones are available.

To test this assumption, we randomly selected 100 common products. We found a significant difference between the ratings from unverified and verified reviews (see Figure 19). This indicates that unverified reviews provide unique and valuable perspectives on these products.

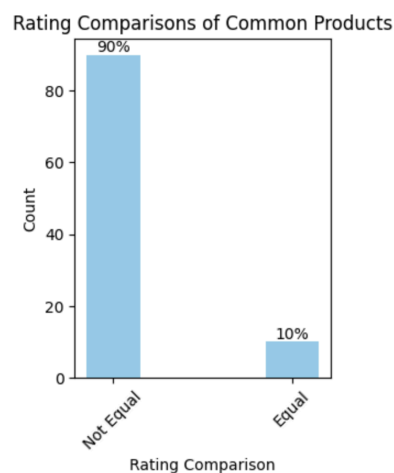


Figure 19

In conclusion, our results showed that it's necessary to include unverified reviews in the further analysis and recommendation generation.

E. Integration

Integrating the three parts of our project, we cultivated a recommendation strategy that delivered accurate product recommendations. Taking a specific product as the input (see Figure 20), we generated a recommendation list of top rated products (see Figure 21).

	parent_asin	title	features	description	price	store	categories	details	bought_together	subtitle
0	B08R39MRDW	DouBCQ Women's Palazzo Lounge Wide Leg Casual Flowy Pants(Flowr Mix Blue, XL)	["Drawstring closure","Machine Wash"]		NaN	DouBCQ		{"Package Dimensions":"15 x 10.2 x 0.4 inches; 9.59 Ounces","Item model number":"Drop Crotch","Date First Available":"February 5, 2021"}	None	None

Figure 20

	parent_asin	title	features	description	price	store	categories	details	bought_together	subtitle
0	B08R3CTLR	DouBCQ Women's Palazzo Lounge Wide Leg Casual Flowy Pants(Green, L)	["Drawstring closure","Machine Wash"]		NaN	DouBCQ		{"Package Dimensions":"15 x 10.2 x 0.4 inches; 9.59 Ounces","Date First Available":"February 5, 2021"}	None	None
1	B085CFKCK	DouBCQ Women's Summer Palazzo Comfy Lounge Flowy Wide Leg Casual Beach Pants(Brown Floral Black L)	["Drawstring closure","Machine Wash"]		NaN	DouBCQ		{"Package Dimensions":"15 x 10.2 x 0.4 inches; 9.59 Ounces","Item model number":"Drop Crotch","Date First Available":"May 8, 2020"}	None	None
2	B08R39X39B	DouBCQ Women's Palazzo Lounge Wide Leg Casual Flowy Pants(Flowr White, XL)	["Tie closure","Machine Wash"]		NaN	DouBCQ		{"Package Dimensions":"15 x 10.2 x 0.4 inches; 9.59 Ounces","Item model number":"Drop Crotch","Date First Available":"February 5, 2021"}	None	None
3	B08R3BZK7	DouBCQ Women's Palazzo Lounge Wide Leg Casual Flowy Pants(Flowr White,L)	["Drawstring closure","Machine Wash"]		NaN	DouBCQ		{"Package Dimensions":"15 x 10.2 x 0.4 inches; 9.59 Ounces","Item model number":"Drop Crotch","Date First Available":"February 5, 2021"}	None	None
4	B08R39RB23	DouBCQ Women's Palazzo Lounge Wide Leg Casual Flowy Pants(Camouflage, XL)	["Drawstring closure","Machine Wash"]		NaN	DouBCQ		{"Package Dimensions":"15 x 10.2 x 0.4 inches; 9.59 Ounces","Item model number":"Drop Crotch","Date First Available":"February 5, 2021"}	None	None

Figure 21

In summary, we offered a fine recommendation system by generating products of high resemblance with the input, passing down the results through middleware, and ultimately created a list of recommendations based on massive data analysis. Through this project, we aimed to improve customer experience, empower businesses with insightful marketing decisions, and optimize the recommendation system for Amazon products.

