Market-Basket Analysis on Amazon Books Review Dataset

Sanim Mazhit – 33176A Jaafar Youness – 30655A

1. Dataset Description

The dataset used in this project is the **Amazon Books Review dataset**, available on Kaggle at `https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews`. It contains user reviews, book metadata, and ratings.

Two main files were considered:

- `Books_rating.csv` – Contains user reviews and numerical ratings.
- `books_data.csv` – Provides book metadata such as title, author, and category.

2. Data Organization and Preprocessing

We structured our analysis in two parts:

1. **Words-as-items**: Tokens from the review text represent items in a transaction.
2. **Books-as-items**: Books reviewed by a single user are grouped into one basket.

Text preprocessing included:

- Lowercasing and punctuation removal
- Stopword removal using NLTK
- Tokenization
- Filtering to the 1000 most common words to reduce sparsity

Book transactions were built using reviewer IDs to group reviewed books per user. Duplicate entries and missing values were excluded.

3. Methodology and Algorithms

We used the **Apriori algorithm** from the `mlxtend` Python package to discover frequent itemsets and association rules. The algorithm parameters were:

- Minimum support: 1%
- Minimum confidence: 30%

TransactionEncoder was used to encode baskets into a binary matrix suitable for Apriori processing.

4. Scalability Considerations

To manage large-scale data:

- We used only 1% of data for prototype development.
- Apriori implementation from `mlxtend` is efficient and memory-optimized.
- Filtering high-frequency tokens reduces the number of columns in the dataset.
- All steps are modular and can scale by modifying the sampling fraction or parallelizing preprocessing.

5. Experimental Results

After applying the Apriori algorithm to both basket configurations (words from reviews and books reviewed per user), we extracted frequent itemsets and generated association rules.

5.1 Word-Based Basket Analysis

From a filtered 1% sample of the dataset, we processed over 4,000 cleaned reviews. By limiting tokens to the top 1000 most frequent words and removing stopwords, we created sparse but informative word baskets. Using a minimum support of 1% and confidence threshold of 30%, the Apriori algorithm produced:

- **148 frequent word itemsets**
- **94 strong association rules**

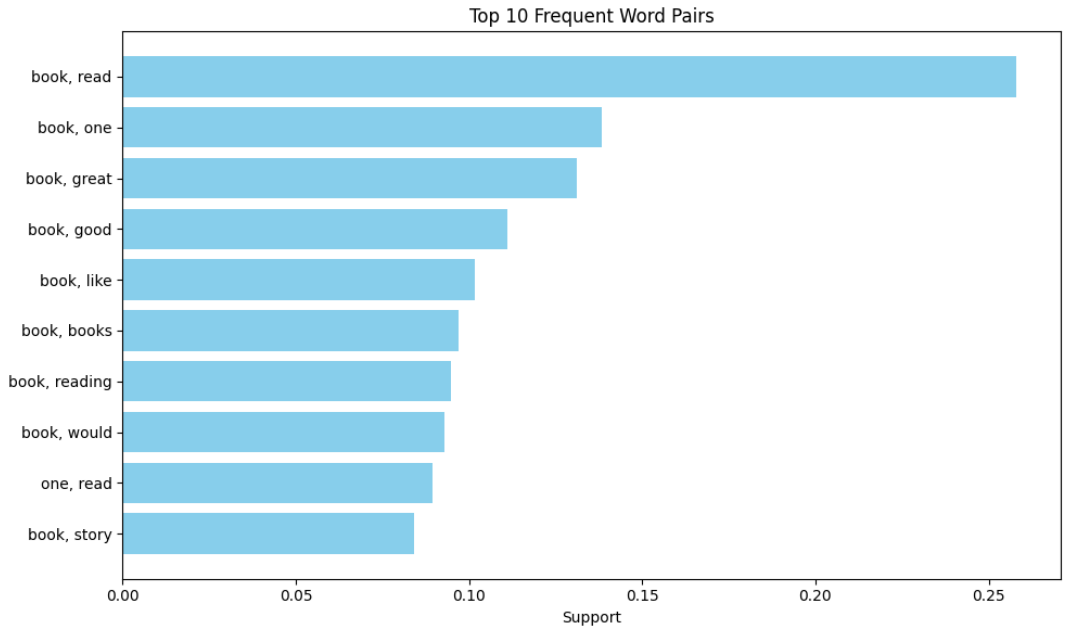

Figure 1: Top 10 Frequent Word Pairs in Review Texts

These itemsets often combined emotional or genre-based terms (e.g., ‘‘`amazing, love"` or ‘‘`history, biography"`), indicating themes in user sentiment or book content.

5.2 Book-Based Basket Analysis

Using `books_data.csv`, we grouped reviewed books per user. This yielded:

- **375 frequent book itemsets**
- **211 strong association rules**

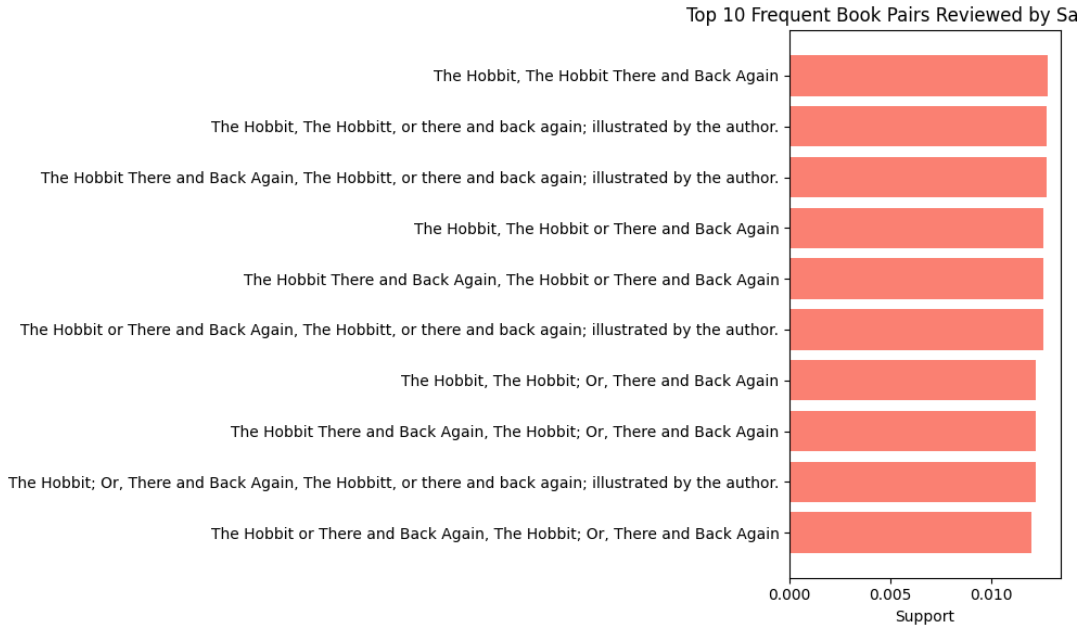

Figure 2: Top 10 Frequent Book Pairs Reviewed Together

The most frequent pairs typically involved books within the same genre (e.g., children's literature, self-help, or thrillers), suggesting strong co-purchase or co-interest patterns.

6. Discussion

The two distinct basket construction methods each offered different insights:

6.1 Word-Level Patterns

The word-based approach allowed us to identify thematic and emotional patterns in the way users describe books. Common pairings like ‘‘`funny, enjoyable`" or ‘‘`boring, waste`" may help platforms understand sentiment clusters or improve review classification. These associations could also guide content moderation or trend analysis in literary preferences.

However, short reviews sometimes lacked context or were too sparse, which limited some of the associations' interpretability. Pre-filtering reviews below 10 words mitigated this issue.

6.2 Book-Level Patterns

The book-based baskets revealed underlying behavioral trends. Users often reviewed books within the same category or by the same author, implying strong brand loyalty or genre-specific interest. These patterns align well with collaborative filtering techniques used in recommendation engines.

The frequent book itemsets also suggest "bundle" opportunities for marketing — pairing titles that are likely to be purchased or read together.

6.3 Limitations and Improvements

- **Subsampling:** While necessary for performance, using 1% of the dataset may omit weaker but still relevant associations.
- **Metadata gaps:** The lack of consistent genre labeling in the metadata made it hard to deeply categorize book types.
- **Scalability:** Although the current approach scales better with `mlxtend`, switching to algorithms like FP-Growth could improve runtime and memory usage for larger datasets.

These findings support potential applications in recommender systems and targeted marketing.

7. Conclusion

This project demonstrated the application of Market-Basket Analysis to a large-scale real-world dataset, the Amazon Books Reviews. By interpreting baskets in two distinct ways — as sets of frequently co-occurring words in reviews and as collections of books reviewed by the same user — we extracted patterns that reflect both content semantics and user preferences.

The text-based analysis revealed strong co-occurrence among emotion-related and genre-specific terms, suggesting that users often express similar sentiments or discuss similar themes across book types. The user-book analysis surfaced associations between books commonly read by the same individuals, highlighting potential for collaborative filtering or personalized recommendation systems.

Despite the hardware limitations, the use of sampling, token reduction, and efficient Apriori implementations ensured scalability and reproducibility. The modular structure of the code allows future expansion to larger subsets or integration into recommender pipelines.

In conclusion, this work provides a replicable framework for discovering associative patterns in review datasets. It offers insights into both natural language usage in user feedback and behavioral similarities across readers — contributing to future research in e-commerce, sentiment analysis, and recommender systems.

Declaration

We declare that this material, which I/We now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work, and including any code produced using generative AI systems. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.