February 28, 2025

C964: Computer Science Capstone

Task 2 parts A, B, C and D

Part A: Letter of Transmittal

February 28, 2025

Mr. Karston Trogstad

BookFluence

12265 E 1800 S

Salt Lake City, UT. 84102

Dear Mr. Trogstad,

I am pleased to present this proposal for developing and implementing a machine learning-based book recommendation system for BookFluence. This project aims to enhance user engagement, improve book discovery, and drive revenue growth by providing personalized recommendations based on sentiment analysis and content-based filtering.

BookFluence's current system primarily relies on popularity-based recommendations, favoring bestsellers and widely reviewed books. This approach fails to accommodate individual user preferences, leading to disengagement and limiting book diversity within the platform. We risk users turning to competitor platforms that offer more tailored suggestions by not offering personalized recommendations.

Our solution introduces a hybrid recommendation system integrating sentiment analysis with content-based filtering. Using natural language processing (NLP) and a support vector machine (SVM) model, the system will analyze user-generated book reviews to classify sentiment as positive or negative. Based on these classifications, the system will recommend books with similar attributes to those labeled with positive sentiment while avoiding those with negative reviews. This approach ensures that recommendations align with user preferences, leading to a more engaging experience. The system will continuously learn from new reviews and interactions, refining its recommendations.

There are four significant benefits of this proposed system for BookFluence. The enhanced personalization ensures that users receive book suggestions tailored to their preferences. This increases engagement and trust in our platform, further increasing platform retention. Next, users' options will be broadened beyond bestsellers because

lesser-known books with high positive sentiment will gain visibility. Higher user satisfaction can increase engagement, driving more significant revenue for our platform. Finally, the system is designed to quickly process large volumes of reviews, making it adaptable as our user base grows.

The project will be implemented over 25 days and divided into five key phases: data collection, model training, testing, prototype development, and deployment. The estimated project cost is $13,500, including hardware, software, and personnel costs. User-generated reviews from a dataset provided by Kaggle will be used for training (Bekheet, 2023). Data privacy will be ensured by anonymizing user information and securing stored data.

With my background in machine learning and data science, I am confident in the feasibility of this project. The proposed model is built on proven techniques, including sentiment analysis and content-based filtering, which have been successfully applied in recommendation systems across various industries.

I respectfully request your review and approval to proceed with this project. Thank you for your time and consideration. I look forward to your feedback.

Sincerely,

~~**********~~

******, Machine Learning Engineer

Part B: Project Proposal Plan

Project Summary

BookFluence, a decade-old book discovery platform, aims to improve its book recommendation system to provide users with more personalized and diverse book suggestions, ultimately increasing user engagement and driving revenue growth. The current system primarily relies on popularity-based recommendations, favoring bestsellers and widely reviewed books. This approach fails to cater to individual reader preferences. To address this issue, the proposed solution involves implementing a hybrid machine learning-based recommendation system that focuses on sentiment analysis and content-based filtering to generate more relevant book recommendations. This machine learning approach will improve BookFluence's book recommendations through personalization, discovery of lesser-known books, improving user engagement, and an improved user experience.

BookFluence's current recommendation system faces challenges that can be addressed through various machine-learning solutions. We will implement a hybrid recommendation system combining sentiment analysis with content-based filtering to enhance personalization. Sentiment analysis is a natural language processing (NLP) technique that classifies text as positive or negative based on the tone or emotion expressed (GeeksforGeeks, 2024). Meanwhile, content-based filtering recommends items by analyzing their characteristics to suggest similar ones to the user (Murel & Kavlakoglu, 2024).

Our system will assess user-generated book reviews within the app. The process begins by extracting and preprocessing the text. This includes cleaning the data and converting the text into numerical representations that a supervised learning model, the support vector machine (SVM), can process. The model then classifies each review as positive or negative. If the review is positive, the system recommends books with similar attributes, unlike when a review is negative, the system avoids recommending similar books. Through this approach, users receive a personalized list of book recommendations that align with their preferences based on the sentiment expressed in their past reviews.

The deliverables that will be accomplished with this project include a hybrid book recommendation system that combines sentiment analysis with content-based filtering. Included in the delivery of this system, a user guide will be provided for using the application.

The proposed machine-learning solution offers several key benefits. First, it enhances user personalization by integrating sentiment analysis, allowing the system to go beyond the current recommendation system by considering the emotional tone of user reviews. This ensures that users receive book suggestions that are highly relevant and precisely tailored to their preferences and sentiments. Additionally, delivering more accurate recommendations fosters a more engaging and enjoyable user experience, increasing user trust and interaction. The next benefit is the system's ability to continuously learn and adapt based on new user reviews and interactions. This dynamic learning process ensures that recommendations remain fresh, relevant, and aligned with evolving user interests. Lastly, the system is designed for efficiency and scalability. By leveraging NLP and machine learning models like SVM, large volumes of user reviews can be processed quickly and efficiently. This ensures that the system will be scalable and accommodating to a growing user base. These combined benefits position the system as a powerful tool for delivering personalized, dynamic, and efficient book recommendations.

Data Summary

A dataset of user reviews from Kaggle.com called “Amazon Books Reviews” will be used as the source of raw data (Bekheet, 2023). This raw data will be loaded into the program through pandas. After it is loaded in, the data will be cleaned and preprocessed to ensure it is structured and ready for sentiment analysis.

The data will be prepared for sentiment analysis using an SVM and content-based filtering through the ETL (Extract, Transform, Load) Process. This ensures data integrity before it is used for training and predictions. First, the data, such as user reviews and book titles, will be extracted from the raw “Amazon Books Reviews” data using pandas.

The transformation phase ensures data is clean, reliable, and formatted correctly before use in machine learning models. This process includes data cleaning and standardization. Handling missing or incomplete data is crucial for accurate analysis. Reviews without text will be removed, as they provide no value for sentiment analysis. Stopwords, punctuation, and special characters will all be removed from the user reviews to improve the machine-learning model performance. Words will also go through lemmatization, and contractions will be expanded into their respective words to standardize the text. The text will then be converted to lowercase for consistency.

The final step of the ETL process is loading the cleaned data to be stored for later use. Processed data will be converted into numerical representations and saved for SVM classification, while processed book metadata will be prepared for content-based filtering recommendations.

The data used for this project meets the requirements for training the SVM properly for sentiment analysis. The user reviews within the data have a majority of positive reviews, but there are still negative reviews for the SVM to use as training. Data anomalies, such as incomplete data, will be filtered out since the SVM cannot classify the review's sentiment.

Data handling practices must prioritize security, integrity, and compliance while ensuring the ethical use of user-generated content. Since BookFluence collects book reviews, ratings, and metadata, special care must be taken to protect user anonymity, prevent data misuse, and maintain the dataset quality. BookFluence does not collect highly sensitive personal information, such as financial or medical data, but even user-generated reviews and associated metadata still require careful handling. Personally Identifiable Information (PII) will be stripped from the dataset before processing occurs.

Implementation

The development of the hybrid-based recommendation system will follow the SEMMA methodology. The project will systematically extract valuable insights from user reviews, improve sentiment classification accuracy, and generate personalized book recommendations.

Sample: Gather and prepare a representative dataset of book reviews and metadata for analysis.

During the sample stage, a dataset with various review ratings will be selected to ensure a diverse representation of sentiment. User-generated reviews will then be extracted from the dataset. Along with this, book metadata will be collected from Google Books API for content-based filtering.

Explore: Understanding the dataset's patterns, trends, and potential biases.

The explore stage handles performing exploratory data analysis (EDA) to identify word frequency, sentiment distribution, and data quality issues. This is where potential imbalances in sentiment labels can be identified to see if changes will need to be required to make the dataset more balanced.

Modify: Preprocess and transform the data to improve machine learning model performance.

The modification stage is one of the most essential stages for machine learning model performance. Stopwords, punctuation, and any special characters will be removed. Next, contractions will be expanded, and lemmatization will be applied to standardize words. Text will also be converted to lowercase for consistency. Finally, book reviews will be converted into numerical representations, and the sentiment polarity scores will be assigned.

Model: Train and deploy machine learning models to analyze sentiment and recommend books.

In the model stage, the support vector machine (SVM) model will begin training to classify reviews as positive or negative to implement the sentiment analysis model. In the content-based filtering model, cosine similarity will be used to compare book metadata to suggest similar books based on positive sentiment. Both models will then be combined to form a hybrid recommendation model.

Assess: Evaluate the system's performance and refine the models for better results.

The sentiment model performance will be measured through a Classification Report, which assesses Precision, Recall, F1 Score, and Support. These four assessments are essential in determining how well the sentiment model is learning from the dataset. The recommendation quality will be assessed by comparing the recommendation precision.

Timeline

Milestone or Deliverable	Dependency	Duration (Days)	Projected Start Date	Anticipated End Date	Resource Assigned
▪ Acceptance of the proposal by senior leadership.	-	1 Day	03/10/2025	03/11/2025	Machine Learning Engineer
▪ Identify data sources and technology stack.	Acceptance of proposal	2 Days	03/11/2025	03/13/2025	Machine Learning Engineer
▪ Collect and clean user-generated book reviews and metadata. ▪ Perform exploratory data analysis (EDA) to identify patterns and trends. ▪ Apply text preprocessing. ▪ Prepare dataset for sentiment classification and content-based filtering.	Identification of data source	4 Days	03/13/2025	03/17/2025	Machine Learning Engineer
▪ Train the support vector machine (SVM) for classifying reviews. ▪ Develop and optimize the content-based filtering model for recommendations. ▪ Test and fine-tune models for accuracy and efficiency.	Preparation and cleaning of the dataset	3 Days	03/17/2025	03/20/2025	Machine Learning Engineer
▪ Build a basic functional prototype demonstrating sentiment-based recommendations. ▪ Conduct internal testing and refine models based on initial results. ▪ Gather feedback from senior leadership to improve the system.	Development, training, and testing of SVM and content-based filtering model	4 Days	03/20/2025	03/24/2025	Machine Learning Engineer
▪ Integrate sentiment analysis and content-based filtering into a single recommendation engine. ▪ Develop a simple user interface to display book recommendations.	Development of prototype and gathering of feedback	2 Days	03/24/2025	03/26/2025	Machine Learning Engineer

▪ Conduct system testing to ensure performance, accuracy, and scalability.					
▪ Submit the project for review. ▪ Collect senior leadership feedback and implement necessary improvements.	Models are combined and UI development	2 Days	03/26/2025	03/28/2025	Machine Learning Engineer
▪ Finalize machine learning models. ▪ Prepare and submit performance metric reports. ▪ Conduct final testing and validation before deployment.	Submission of project and feedback	3 Days	03/28/2025	03/31/2025	Machine Learning Engineer
▪ Deploy the working prototype of the recommendation system. ▪ Monitor system performance and user engagement with the deployed prototype.	Finalization of models and final testing is complete	4 Days	03/31/2025	04/04/2025	Machine Learning Engineer

Evaluation Plan

The verification methods that will be used include histograms, word clouds, word correlation heat maps, and classification reports. Histograms will determine sentiment distribution before and after balancing the data. Word clouds will be used to see the most common words used within the reviews in the dataset. Word correlation heat maps will be used to see how the words within reviews correlate. These two data visualizations will be used to see how the data is affected before and after data cleaning. Finally, classification reports will be used to determine the accuracy of the sentiment analysis during model training, and a cosine similarity score will be used to measure the similarities of the recommendations.

The validation method that will be used upon project completion is the Holdout Validation Method. In this method, a portion of the data is set aside to train the model, while the rest is used to test the model (GeeksforGeeks, 2020). This method is beneficial because it prevents the model from memorizing and allowing it to learn from the data.

Resources and Costs

Resource	Description	Cost
Hardware	Computer	$1,500
Software	Google Colab, Google Books API	$0
Data Acquisition	Datasets from Kaggle (Bekheet, 2023).	$0
Personnel	Machine Learning Engineer $60/hour, 8 hours/ day for 25 days	$12,000
	Total:	$13,500

Part C: Application

Required Files:

- **Book_Recommendation_System.ipynb:** Contains the source code for the machine-learning solution
- **Books_rating_shortened.csv:** The raw dataset used to train the SVM

Part D: Post-implementation Report

Solution Summary

BookFluence needed to improve its book recommendation system to serve users better. Their old system favored recommending bestsellers and widely reviewed books due to using popularity-based recommendations. The project proposed to solve this issue was a hybrid machine learning-based recommendation system focused on sentiment analysis and content-based filtering. By leveraging the system, users have received more relevant recommendations based on the reviews that are left. The system has also provided higher user engagement, generating more revenue for BookFluence.

Data Summary

A dataset of user reviews from Kaggle.com called “Amazon Books Reviews” was used as a raw data source to train the hybrid model (Bekheet, 2023). This raw data was loaded into the program through pandas, which were then cleaned and preprocessed to ensure the sentiment analysis classification functions smoothly.

The ETL (Extract, Transform, Load) Process prepared the data for sentiment analysis. This ensured data integrity before it was used for training and predictions. User reviews and book titles from the raw dataset were extracted using pandas. Next, data was cleaned and formatted in the transform phase before use in the machine learning models. Reviews without text were removed since they provided no value for sentiment analysis. When the reviews were preprocessed, stopwords, punctuations, and special characters were removed to improve the models' learning performance. Contractions were expanded, and then words were lemmatized to standardize the text. Finally, all text was converted to lowercase for consistency.

The transformation phase ensures data is clean, reliable, and formatted correctly before use in machine learning models. This process includes data cleaning and standardization. Handling missing or incomplete data is crucial for accurate analysis. Reviews without text will be removed, as they provide no value for sentiment analysis. Stopwords, punctuation, and special characters will all be removed from the user reviews to improve the machine-learning model performance. Words will also go through lemmatization, and conjunctions will be separated into their respective words to standardize the text. The text will then be converted to lowercase for consistency.

In the final step, load, processed data was converted into numerical representations and saved for SVM classification. Processed book metadata was prepared and saved for content-based filtering.

Machine Learning

The machine learning methods used were a support vector machine (SVM) for sentiment analysis and content-based filtering for recommendations.

A support vector machine is a supervised machine learning algorithm used mainly for classification tasks. This project used the SVM to classify book reviews as either positive or negative based on their textual context. The data was first cleaned and preprocessed by removing punctuation, stopwords, and special characters. Next, the processed text was converted into a numerical format using TF-IDF (Term Frequency-Inverse Document Frequency) to capture the word importance. The dataset was split into training and testing subsets to train the model, and the SVM model was trained on the TF-IDF transformed data. Once trained, the model was used to classify new user-submitted reviews as positive or negative. The SVM was selected for text classification because it is very effective in small dataset classification. In addition to this, SVMs are designed to maximize the margin between classifications, leading to better performance on unseen data.

Content-based filtering is a recommendation system technique that suggests books based on their similarity to a given book. Each book's metadata was extracted and converted into numerical format to develop this method. Next, cosine similarity was used to compare books based on their vector representations. Finally, the system would select the most similar books to recommend. This method was chosen because it doesn't require other user reviews and ratings like collaborative filtering. It works even if a book is newly added. Next, users can understand why a book was recommended because of the shared attributes. The final reason why this method was selected is because it is computationally efficient and doesn't require large datasets of user interactions.

Validation

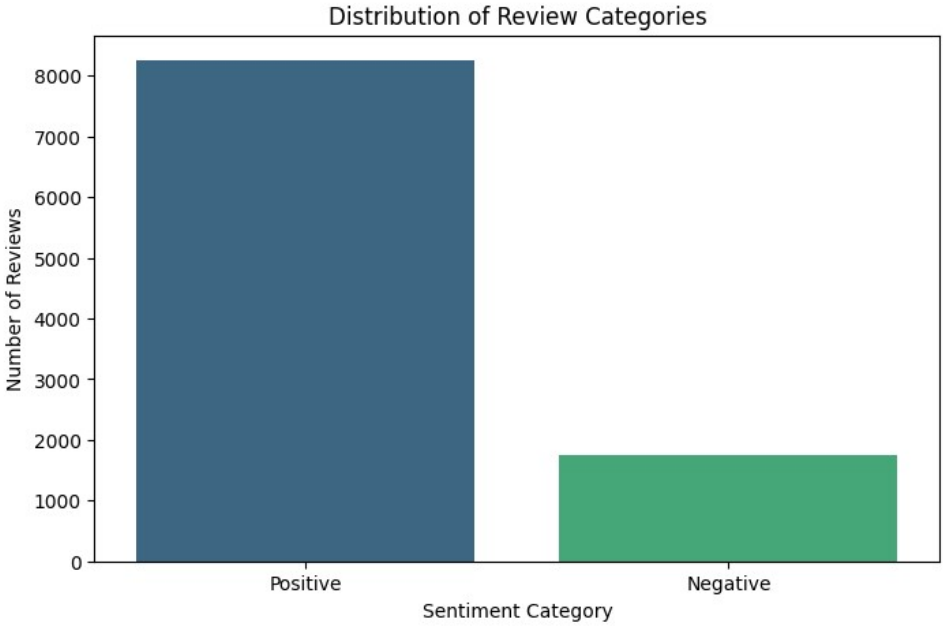
The validation method used for the SVM was a Classification Report. This report measures several metrics, such as Precision, Recall, F1 Score, and Support, to provide the proportions of correctly classified reviews. The result of this validation method was an SVM Accuracy of 91.38%.

```
SVM Accuracy: 91.38%
Classification Report:
              precision    recall  f1-score   support

    Negative       0.88      0.92      0.90      1264
    Positive       0.94      0.91      0.92      1659

    accuracy                           0.91      2923
   macro avg       0.91      0.92      0.91      2923
weighted avg       0.91      0.91      0.91      2923
```

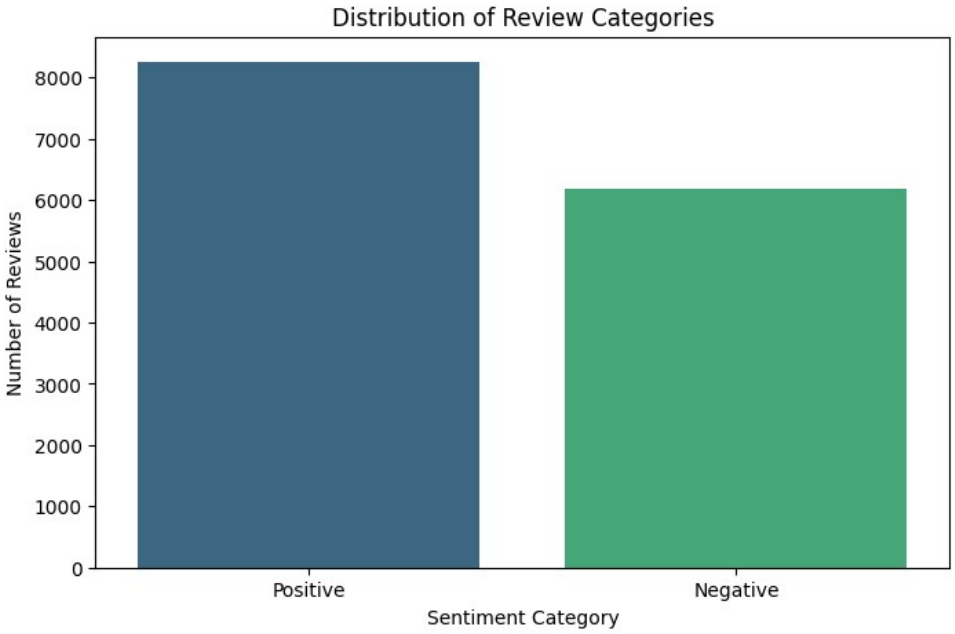
Next, a cosine similarity score is a validation method to check the content-based filtering recommendations. This score measures how similar the two books are. Values closer to 1 mean the system recommends a book with high similarity.

Visualizations

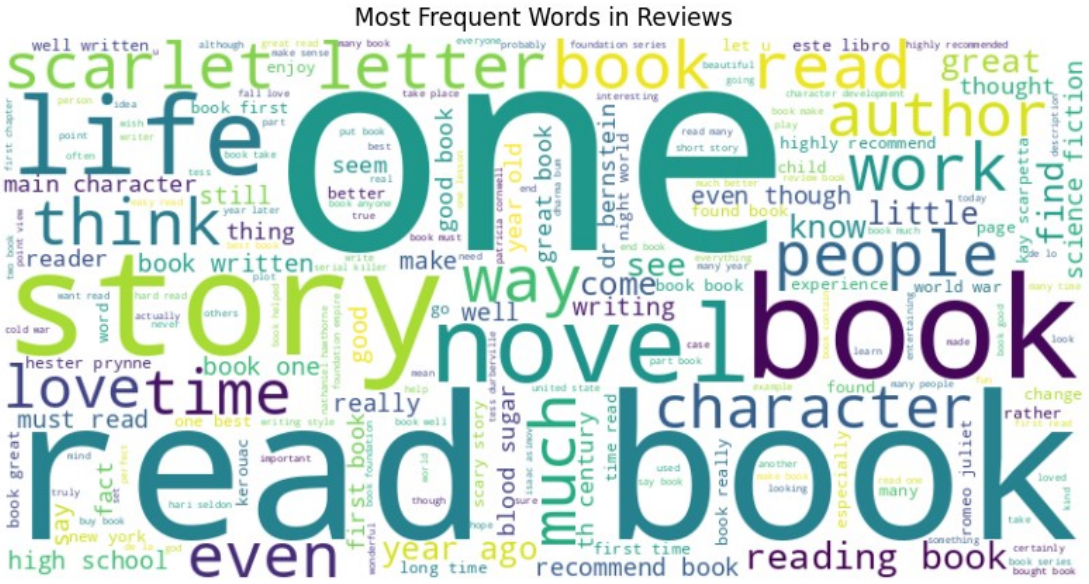
This visualization can be found in the Google Colab program. It displays the distribution of sentiments for the raw dataset before balancing. The histogram shows an imbalance of positive and negative reviews, which could train the SVM to favor the positive as it is learning to classify.



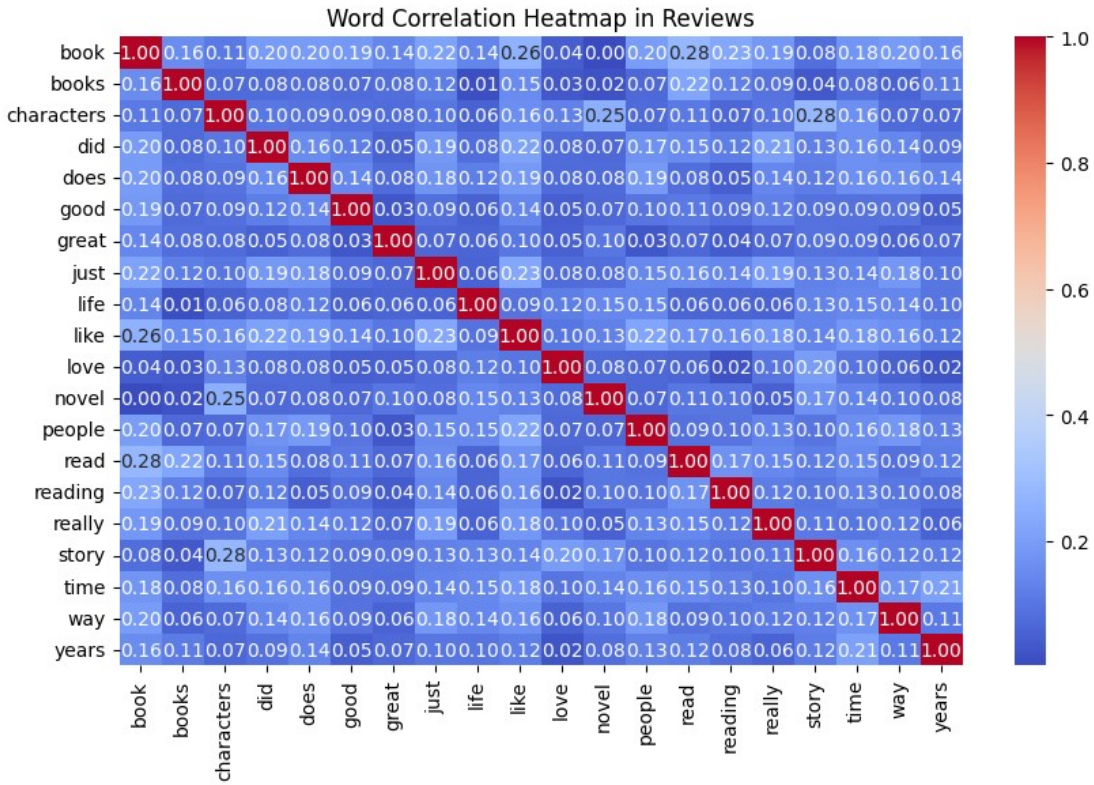
Here is the histogram of the same sentiment categories after the balanced dataset. After balancing the data, the SVM was more accurate in its classifications.



Next, the word cloud below shows the most common words throughout the reviews in the dataset.



The word correlation heat map below shows the correlation between words across reviews.



User Guide

1. Navigate to the following Google Colab link:
    - https://colab.research.google.com/
2. Click “Upload,” then “Browse,” and find the file named **Book_Recommendation_System.ipynb** and click “Open.”
3. Select the folder icon on the left-hand side of the page.
4. Upload the provided dataset “Books_ratings_shortened.csv”
5. At the top of the screen, click the “Runtime” tab and select “Run All.”
6. This will run through each code block automatically.
7. Follow the running process until prompted to enter a book title.
8. Enter “The Great Gatsby”
9. Next, a prompt for the review will appear; enter “So good! I loved every second of the book!”
10. The system will then output five books that it recommends.

Reference Page

GeeksforGeeks (2024, January 24). *What is Sentiment Analysis?* Geeksforgeeks.org. Retrieved February 22, 2025, from https://www.geeksforgeeks.org/what-is-sentiment-analysis/?ref=header_outind

Murel, J., & Kavlakoglu, E. (2024, March 21). *What is content-based filtering?* IBM. Retrieved February 22, 2025, from https://www.ibm.com/think/topics/content-based-filtering

Bekheet, Mohamed (2023). *Amazon Books Reviews.* Kaggle.com. Retrieved February 25, 2025, from https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

GeeksforGeeks (2020, August 26). *Introduction of Holdout Method* Geeksforgeeks.org. Retrieved February 28, 2025, from https://www.geeksforgeeks.org/introduction-of-holdout-method/