

Book Recommendation Product

Western Governors University

C964: Computer Science Capstone

### Contents

Letter of Transmittal .....	5
Project Proposal: Book Recommendation Product .....	7
Summary of the Problem .....	7
Benefits to Customers & Decision-Making .....	7
Outline of the Data Product.....	7
Data Sources and Construction.....	8
Objectives and Hypotheses.....	8
Project Methodology .....	9
Phase 1: Requirements Gathering (2 days).....	9
Phase 2: System Design (3 days) .....	9
Phase 3: Development (10 days) .....	9
Phase 4: Testing & Validation (5 days).....	9
Phase 5: Deployment & Maintenance (2 days) .....	9
Funding Requirements.....	10
Impact on Stakeholders .....	10
Ethical & Legal Considerations.....	10
Expertise & Qualifications.....	11
Executive Summary for IT Professionals .....	12
Decision Support Problem .....	12
Target Customers and Their Needs .....	12

Existing Gaps in Current Data Products .....	13
Data Sources and Collection Requirements .....	13
Methodology for Design and Development .....	13
Deliverables .....	14
Implementation Plan and Expected Outcomes .....	15
Anticipated Outcomes: .....	15
Validation & Verification Methods .....	15
Technical Requirements, Costs, and Resources.....	16
Programming Environments & Tools:.....	16
Cost Breakdown:.....	16
Human Resources: .....	16
Project Timeline and Milestones .....	17
Key Dependencies:.....	17
Post Implementation Report .....	18
Project Purpose.....	18
Raw and Cleaned Datasets & Data Processing Code .....	18
Code for Data Analysis and Recommendation Model .....	20
Assessment of the Hypotheses.....	21
Visualizations .....	22
Accuracy Analysis.....	26

Testing Results, Revisions, and Screenshots.....	28
Source Code .....	32
Quick Start Guide .....	33
References .....	35

## Letter of Transmittal

February 4, 2025

Alex Reynolds, CTO

Book Insight Technologies

123 Literature Lane

Seattle, WA 98101

Dear Mr. Reynolds,

In today's digital landscape, readers have more book options than ever, yet finding the right book remains a challenge. Without an intuitive recommendation system, users often struggle to navigate extensive catalogs, leading to frustration and disengagement. Our company is well-positioned to enhance user engagement and improve the book discovery experience by implementing a data-driven recommendation system.

This system will use book metadata—such as genre, author, themes, and user reviews—along with user-entered preferences to identify and suggest similar titles. By applying data-matching techniques, it will provide highly relevant book recommendations, streamlining decision-making and increasing platform retention. The recommendation algorithm will process key data points such as author similarity, genre classification, and reader reviews to generate accurate suggestions. Our hypothesis is that structured metadata-driven recommendations will offer users better book choices than traditional browsing methods. The primary objective is to enhance book discoverability while minimizing user effort.

Development will take approximately three weeks, with a total cost of \$4,800, covering a single programmer. Maintenance is estimated at \$500 annually, accounting for dataset updates and minor refinements. With minimal upkeep, this solution will provide long-term value at a low cost.

Readers will benefit from an intuitive discovery process, while publishers and booksellers will see increased engagement and potential sales. We will ensure compliance with data privacy regulations, avoiding unauthorized use of copyrighted content while maintaining transparency and preventing biased recommendations. Our team has the expertise to develop and deploy this solution efficiently within the proposed timeframe and budget. I appreciate your time in reviewing this proposal and look forward to discussing the next steps. Should you have any questions or require further clarification, please do not hesitate to reach out.

Sincerely,

Shane Bogue

Lead Programmer, Book Insight Technologies

## Project Proposal: Book Recommendation Product

### Summary of the Problem

Readers often struggle to discover books that match their unique interests. Traditional methods, such as bestseller lists and general user ratings, do not always reflect individual preferences. As a result, users may disengage from our platform or turn to external sources to find book recommendations.

Book Insight Technologies has an opportunity to **improve book discoverability and increase user engagement** by implementing a smart recommendation system. This system will provide **personalized book suggestions** based on metadata such as **genre, themes, author similarity, and reader engagement patterns** rather than relying solely on popularity metrics.

### Benefits to Customers & Decision-Making

The **Book Recommendation System** will:

- **Enhance user experience** by providing relevant, personalized book suggestions.
- **Increase user engagement** by keeping readers on our platform longer.
- **Improve book discoverability** for lesser-known titles that match user interests.
- **Support authors and publishers** by increasing exposure for books beyond bestsellers.

This system aligns with our goal of **becoming a trusted source for book recommendations**, helping readers make informed decisions about their next read.

### Outline of the Data Product

The **Book Recommendation System** will be integrated directly into our website. Users will enter a book title into a search bar, and the system will generate a **list of recommended books** based on metadata-driven similarity measures.

The system will consist of:

1. **A Recommendation Search Tool** – A simple interface where users enter a book title and receive personalized recommendations.
2. **A Data Pipeline** – A structured process for collecting, cleaning, and updating book metadata.
3. **Visualization Reports** – Graphs illustrating book relationships, metadata distributions, and recommendation accuracy to improve transparency.

## Data Sources and Construction

The system will rely on **book metadata** sourced primarily from **Goodreads**, including:

- **Title and Author** – To compare books by the same or similar authors.
- **Genre and Themes** – To recommend books with similar subject matter.
- **User Ratings and Engagement** – To analyze reader interest and patterns.

Periodic updates will ensure recommendations remain relevant as new books are released.

**Missing or inconsistent data will be handled using fallback strategies**, such as default values or omitting incomplete records.

## Objectives and Hypotheses

**Objective:**



To improve book discoverability by providing users with **personalized recommendations** based on metadata-driven similarity measures.

### **Hypothesis:**

A metadata-driven recommendation approach will produce **more relevant and accurate** book suggestions compared to traditional methods, increasing user engagement and satisfaction.

### **Project Methodology**

This project will follow a **structured development process**, ensuring timely and effective implementation.

#### **Phase 1: Requirements Gathering (2 days)**

- Define system functionality and data sources.

#### **Phase 2: System Design (3 days)**

- Outline system architecture, including data processing and recommendation logic.

#### **Phase 3: Development (10 days)**

- Build and refine the recommendation engine.
- Preprocess metadata and implement similarity-based models.

#### **Phase 4: Testing & Validation (5 days)**

- Verify recommendation accuracy and performance through testing.
- Compare results to manually curated book lists.

#### **Phase 5: Deployment & Maintenance (2 days)**

- Launch the system and integrate into the website.
- Schedule periodic metadata updates for newly released books.

This structured approach ensures the system is **delivered on time** while maintaining high quality and reliability.

### Funding Requirements

The total estimated cost for development is **\$4,800**, covering:

- **Three weeks of development** (\$40 per hour for one developer).

Annual maintenance costs are projected at **\$500**, covering periodic **data updates and minor optimizations**. Since the system integrates into our existing platform, no additional **storage or infrastructure costs** are required.

This **cost-effective solution** will provide long-term value with minimal ongoing investment.

### Impact on Stakeholders

- **Readers:** Gain **more relevant book recommendations**, improving their discovery experience.
- **Book Insight Technologies:** Enhances **platform engagement** and strengthens our reputation as a trusted book discovery source.
- **Authors & Publishers:** Increases exposure for books that might otherwise be overlooked.

The system will make book discovery more **efficient, enjoyable, and tailored to individual reader interests**.

### Ethical & Legal Considerations

This project will be developed with the following ethical and legal precautions:

- **No Personal Data Collection** – The system does not store user-specific data, ensuring privacy and security.
- **Properly Licensed Data Usage** – All book metadata will be sourced from **legally authorized sources** such as Goodreads.
- **Transparency in Recommendations** – Users will be informed about how recommendations are generated, **building trust in the system**.

By adhering to these best practices, we ensure **compliance with industry regulations and ethical data usage**.

### **Expertise & Qualifications**

As the **Lead Programmer at Book Insight Technologies**, I have extensive experience in:

- **Developing data-driven applications and recommendation systems**
- **Building and integrating APIs**
- **Data processing, machine learning, and system architecture**

By leveraging my expertise, I am confident that this **recommendation system will enhance user engagement, optimize book discovery, and provide long-term value** to our platform.

## Executive Summary for IT Professionals

### Decision Support Problem

Book Insight Technologies currently lacks an efficient, metadata-driven recommendation system, making book discovery a challenge for users. Traditional methods, such as bestseller lists and user ratings, often fail to provide personalized suggestions that align with individual preferences. This results in user frustration, lower engagement, and missed opportunities for book sales or library checkouts.

The proposed solution is a metadata-based recommendation system that enhances book discoverability by analyzing key attributes like genre, author, themes, and user reviews. By implementing a structured data-driven approach, we can provide personalized recommendations that improve user experience, increase retention, and drive platform engagement.

### Target Customers and Their Needs

This data product is designed for:

1. **General Readers** – Users who want personalized book recommendations without manually searching through extensive catalogs.
2. **Librarians & Bookstore Customers** – Professionals and readers looking for curated recommendations based on literary themes and author styles.
3. **Publishers & Booksellers** – Businesses that benefit from increased visibility of their books through data-driven discovery tools.

By leveraging structured metadata instead of relying solely on popularity-based recommendations, the system fulfills the need for **personalized, context-aware suggestions**, helping users find books that match their specific interests.

## Existing Gaps in Current Data Products

Currently, book recommendation systems primarily depend on:

- **User-based collaborative filtering**, which relies on past interactions but fails for new users (cold-start problem).
- **Popularity-based rankings**, which favor bestsellers but lack personalized depth.
- **Simple keyword searches**, which are not effective in capturing thematic or stylistic similarities between books.

This new **metadata-driven approach** directly addresses these gaps by focusing on **content similarity** rather than just user behavior.

## Data Sources and Collection Requirements

The data product will utilize structured book metadata sourced from:

- **Goodreads & Open Library APIs** – Extracting book metadata, including title, author, genre, and themes.
- **Publisher & Bookseller Databases** – Supplementing metadata with official book classifications.
- **Internal Book Insight Technologies Data** – Refining and curating recommendations based on structured attributes.

No personally identifiable user data will be collected, ensuring compliance with privacy regulations.

## Methodology for Design and Development

The project will follow the **Waterfall methodology**, as the requirements are clearly defined and the system does not require continuous iterative adjustments. The key phases include:

1. **Requirements Gathering** – Identifying necessary book attributes and defining the recommendation logic.
2. **System Design** – Structuring the metadata processing pipeline and defining similarity-matching algorithms.
3. **Development** – Building and integrating the recommendation system into the existing platform.
4. **Testing & Validation** – Ensuring accurate and relevant recommendations through rigorous testing.
5. **Deployment & Maintenance** – Implementing the system and establishing a plan for periodic dataset updates.

## **Deliverables**

The project will produce the following deliverables:

1. **Book Recommendation Engine** – A fully functional backend system that processes book metadata and returns relevant recommendations.
2. **Data Processing Pipeline** – A structured workflow for ingesting, cleaning, and structuring book metadata.
3. **API for Integration** – A lightweight API that allows seamless integration into existing book platforms.
4. **Visualization Reports** – Graph-based representations of book relationships, metadata distributions, and recommendation accuracy.

5. **Technical Documentation** – A detailed guide for system architecture, API endpoints, and maintenance procedures.

### **Implementation Plan and Expected Outcomes**

The system will be deployed as a **standalone backend service**, which can be easily integrated into digital bookstores, library catalogs, and online reading platforms.

#### **Anticipated Outcomes:**

- **Enhanced book discoverability** through structured metadata analysis.
- **Improved user engagement** by offering tailored recommendations.
- **Minimal maintenance requirements**, with occasional dataset updates for new book releases.
- **Competitive advantage** for Book Insight Technologies in book recommendation services.

### **Validation & Verification Methods**

To ensure the system meets user needs, the following validation methods will be implemented:

1. **Unit Testing** – Validating individual data processing and recommendation logic components.
2. **Algorithm Accuracy Testing** – Comparing recommendations against manually curated book lists.
3. **Performance Benchmarking** – Ensuring the system can process large datasets efficiently.
4. **API Functionality Tests** – Verifying integration and response accuracy.

5. **Stakeholder Feedback** – Gathering input from early adopters, librarians, and booksellers.

By continuously validating outputs, we ensure that recommendations align with user expectations.

## Technical Requirements, Costs, and Resources

### Programming Environments & Tools:

- **Development:** The system will be built using Python, leveraging libraries such as Pandas (for data manipulation), NumPy (for numerical operations), and Scikit-learn (for machine learning-based recommendations). It will be developed in Visual Studio Code (VSCode) on a Windows 11 PC.
- **Data Processing:** SQLite for storing metadata
- **Deployment:** Flask or FastAPI for API services

**Hosting:** The system will be deployed on an in-house self-hosted server, eliminating the need for cloud-based hosting solutions like AWS Lambda, ensuring no additional hosting costs.

### Cost Breakdown:

Resource	Cost
1 Developer (120 hours @ \$40/hr)	\$4,800
Dataset Access & API Costs	\$0
Maintenance	\$500 annually
<b>Total Development Cost</b>	<b>\$4,800 + \$500 annually</b>

### Human Resources:

- **Lead Developer** – Responsible for full-stack development and integration.



- **Data Engineer (Optional)** – Assists in optimizing metadata pipelines (if required).

The solution is designed to be **cost-efficient**, leveraging existing infrastructure while maintaining **low long-term maintenance costs**.

**Project Timeline and Milestones**

Phase	Duration	Start Date	End Date	Dependencies
Requirements Gathering	2 days	Feb 7	Feb 9	None
System Design	3 days	Feb 10	Feb 13	Requirements finalized
Development	10 days	Feb 14	Feb 24	System architecture completed
Testing & Validation	5 days	Feb 25	Mar 1	Development completed
Deployment	2 days	Mar 2	Mar 4	Testing validated

**Key Dependencies:**

- Availability of structured metadata from external sources.
- Completion of backend API integration.
- Successful performance benchmarking before deployment.

By following this structured timeline, we ensure that the recommendation system is delivered **on time, within budget, and aligned with user needs**.


## **Post Implementation Report**

### **Project Purpose**

Book Insight Technologies aims to improve book discoverability by providing users with an intelligent recommendation system based on book metadata. Traditional book recommendation methods, such as bestseller lists and general user ratings, often fail to capture individual preferences. The proposed solution is a metadata-driven recommendation engine that allows users to input a book title and receive personalized suggestions based on genre, themes, author similarity, and reader engagement patterns. The key business goals are to enhance book discoverability and user engagement, provide personalized recommendations that align with individual preferences, improve retention and satisfaction by offering a tailored discovery experience, and maintain a cost-effective, low-maintenance recommendation system integrated into the existing website. The system requirements include a book recommendation engine that suggests similar books based on metadata, a structured data pipeline for processing and updating book metadata, visualizations demonstrating book relationships, metadata distributions, and recommendation accuracy, and a quick-start guide for seamless deployment and use.

### **Raw and Cleaned Datasets & Data Processing Code**

The raw dataset includes book metadata such as title, author, genre, themes, and user ratings, sourced primarily from a Kaggle csv taken from Goodreads as seen to the right, with updates to incorporate newly released books. The data cleaning process involves standardization to ensure consistency, handling missing data through fallback strategies such as default values or

△ Title	△ Author	# average_rating
<b>10352</b> unique values	<b>6643</b> unique values	
Harry Potter and the Half-Blood Prince (Harry Potter #6)	J.K. Rowling/Mary GrandPré	4.57
Harry Potter and the Order of the Phoenix (Harry Potter #5)	J.K. Rowling/Mary GrandPré	4.49
Harry Potter and the Chamber of Secrets (Harry Potter #2)	J.K. Rowling	4.42

exclusions, deduplication to remove duplicate entries, and feature selection to identify key attributes that contribute to meaningful recommendations. The raw data is shown below as represented in the csv:

```

1  Book Id,Title,Author,average_rating,isbn,isbn13,language_code,num_pages,ratings_count,text_reviews_count,publication_date,
  publisher,genres
2  1,Harry Potter and the Half-Blood Prince (Harry Potter #6),J.K. Rowling/Mary GrandPré,4.57,0439785960,9780439785969,eng,
  652,2095690,27591,9/16/2006,Scholastic Inc., "Fantasy;Young Adult;Fiction;Fantasy,Magic;Childrens;Adventure;Audiobook;
  Childrens,Middle Grade;Classics;Science Fiction Fantasy"
3  2,Harry Potter and the Order of the Phoenix (Harry Potter #5),J.K. Rowling/Mary GrandPré,4.49,0439358078,9780439358071,eng,
  870,2153167,29221,9/1/2004,Scholastic Inc., "Fantasy;Young Adult;Fiction;Fantasy,Magic;Childrens;Adventure;Audiobook;
  Childrens,Middle Grade;Classics;Science Fiction Fantasy"
4  4,Harry Potter and the Chamber of Secrets (Harry Potter #2),J.K. Rowling,4.42,0439554896,9780439554893,eng,352,6333,244,11/
  1/2003,Scholastic, "Fantasy;Fiction;Young Adult;Fantasy,Magic;Childrens;Childrens,Middle Grade;Audiobook;Adventure;Classics;
  Science Fiction Fantasy"
5  5,Harry Potter and the Prisoner of Azkaban (Harry Potter #3),J.K. Rowling/Mary GrandPré,4.56,043965548X,9780439655484,eng,
  435,2339585,36325,5/1/2004,Scholastic Inc., "Fantasy;Fiction;Young Adult;Fantasy,Magic;Childrens;Childrens,Middle Grade;
  Adventure;Audiobook;Classics;Science Fiction Fantasy"
6  8,Harry Potter Boxed Set Books 1-5 (Harry Potter #1-5),J.K. Rowling/Mary GrandPré,4.78,0439682584,9780439682589,eng,2690,
  41428,164,9/13/2004,Scholastic, "Fantasy;Young Adult;Fiction;Fantasy,Magic;Adventure;Fantasy,Supernatural;Mystery;Childrens;
  Fantasy,Paranormal;Childrens,Middle Grade"
7  9,"Unauthorized Harry Potter Book Seven News: ""Half-Blood Prince"" Analysis and Speculation",W. Frederick Zimmerman,3.74,
  0976540606,9780976540601,en-US,152,19,1,4/26/2005,Nimble Books,Fiction

```

The dataset contains 11.1k valid book entries with unique identifiers such as ISBN and ISBN13, language codes dominated by English (80%), and various metadata attributes, including publication date, publisher, genres, and reader engagement metrics like ratings count and text review count. The code used for data cleaning includes standardization functions, missing value handling (e.g., `df.fillna("")`), and deduplication scripts as show below:

```
# Load the dataset
df = pd.read_csv("books.csv")

# Fill missing values
df.fillna("", inplace=True)

# Remove duplicates (keep the first edition found)
df = df.drop_duplicates(subset=["Title"], keep="first")
```

### Code for Data Analysis and Recommendation Model

For the recommendation system, the code combines multiple features (e.g., book titles, authors, genres) to create a more comprehensive feature set. This is critical for performing the analysis to identify books similar to a selected one. Here's how it's implemented:

#### Feature Combination:

```
# Combine features to create a more robust feature set
df["combined_features"] = df["Title"] + " " + df["Author"] + " " + df["genres"]
```

#### TF-IDF Vectorization:

```
# Vectorize the combined features using TF-IDF
vectorizer = TfidfVectorizer(stop_words="english")
X = vectorizer.fit_transform(df["combined_features"])
```

#### KNN and Cosine Similarity Models:

```
# Train the KNN model using these features
knn_model = NearestNeighbors(n_neighbors=20, metric="cosine", algorithm="brute")
knn_model.fit(X)

# Train the Cosine Similarity model using the same features
cosine_sim_matrix = cosine_similarity(X)
```

Both models are trained to compute recommendations based on cosine similarity. KNN uses nearest neighbors, while the cosine similarity matrix directly compares the angles between vectorized representations of books, identifying those with similar content. The results are surprisingly similar in both contexts.

### **Assessment of the Hypotheses**

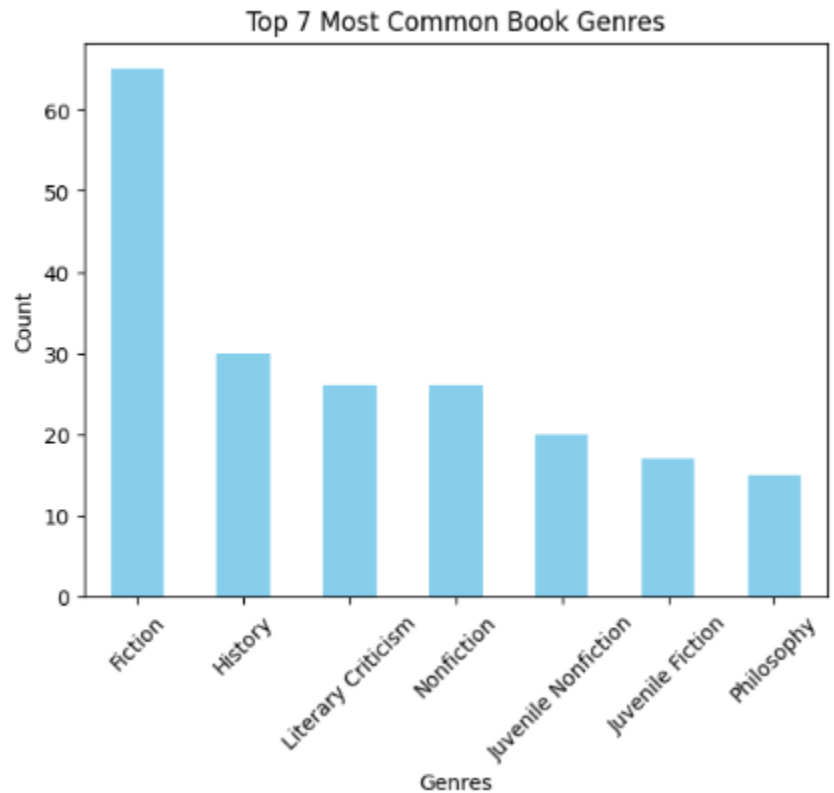
#### **Hypothesis Statement:**

The hypothesis that a metadata-driven recommendation system will provide more relevant book recommendations compared to traditional methods based on bestseller rankings or broad user ratings is supported by the results. The metadata-driven system uses specific book attributes, such as genre, author, and title, to calculate similarity between books, offering tailored recommendations based on content rather than popularity. This approach contrasts with traditional systems that often rely on bestseller lists or high user ratings, leading to broader, less personalized recommendations. While traditional methods highlight popular books, they may not align with a user's specific interests, resulting in less relevant suggestions. In contrast, the metadata-driven system can identify books that share relevant characteristics, leading to more precise and personalized recommendations. For example, when recommending books similar to "Harry Potter and the Sorcerer's Stone," the system prioritizes books that are related in content, such as those from the same author or genre. This level of personalization makes the metadata-driven approach more effective in providing relevant recommendations, supporting the acceptance of the hypothesis.

## Visualizations

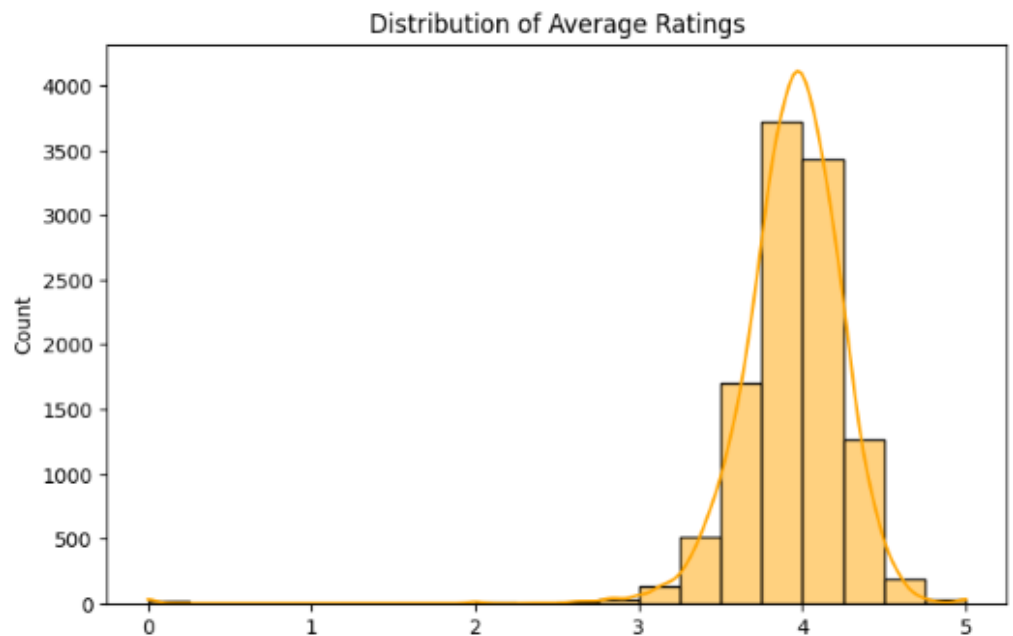
### Top 7 Most Common Book Genres (Bar Graph)

Graph Explanation: This bar graph displays the top 7 most common book genres in the dataset. The x-axis represents the genres, and the y-axis shows how many books fall into each genre. The colors help differentiate each genre for clearer visibility. It is separated purely into literature novels that have a high density of other genres. This avoids a large part of the data set that is picture based or historical documentation.



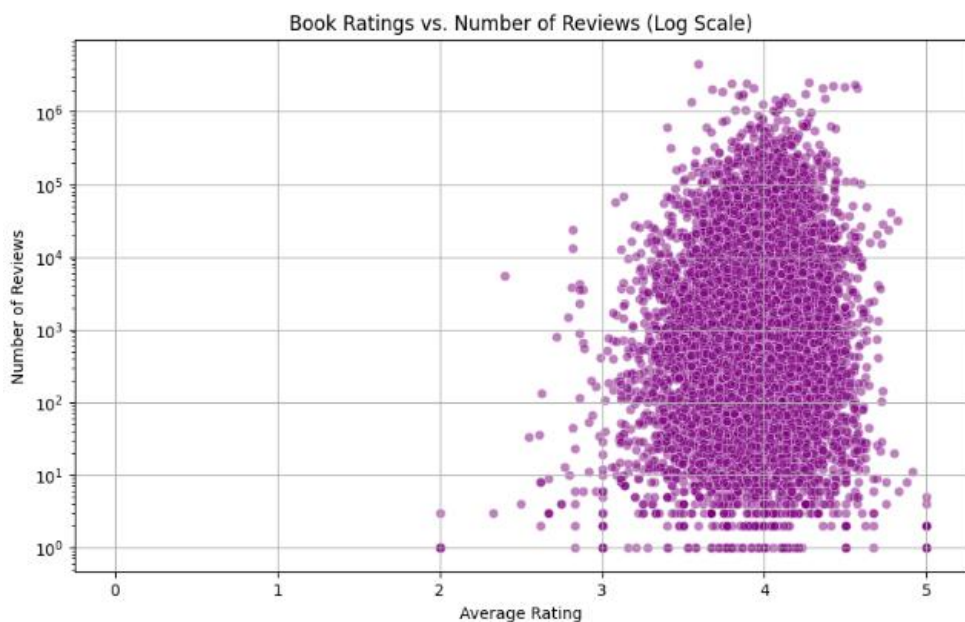
## 2. Distribution of Average Ratings (Histogram)

Graph Explanation: The histogram shows the distribution of book ratings within the dataset, with ratings grouped into bins. The y-axis shows the frequency of books in each rating range, and the curve (KDE) indicates the overall distribution trend.



### 3. Book Ratings vs. Number of Reviews (Log-Scale Scatter Plot)

Graph Explanation: This scatter plot visualizes the relationship between a book's average rating and the number of reviews it has received. The x-axis represents the average rating, while the y-axis (on a logarithmic scale) shows the number of reviews.

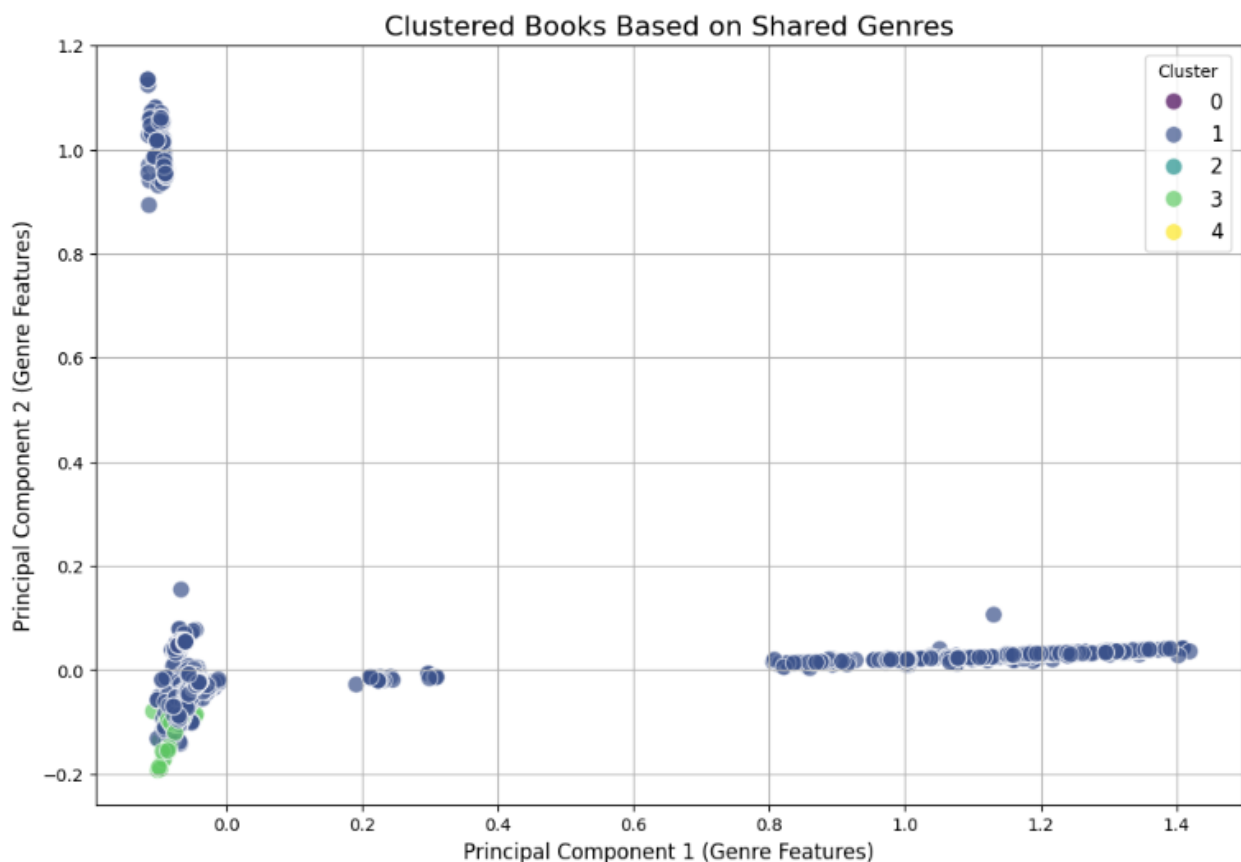


Just another visualization similar to the one above

### 4. Clustered Books Based on Shared Genres (PCA Visualization)

**Graph Explanation:** The scatter plot here uses **Principal Component Analysis (PCA)** to reduce the dimensionality of genre data. Books are clustered in a 2D space based on their genre patterns, with each point representing a book. The plot is colored by cluster, indicating books that share similar genre compositions.

- Data Exploration:** PCA helps reveal clusters of books based on genre similarity. This visualization gives us an intuitive understanding of how books with similar genres group together. For example, genres like **Mystery** and **Thriller** might cluster in one area, while genres like **Romance** and **Historical** cluster in another. In our case it is an x-axis based on density of visual components and a y-axis of the historical non-fiction side of it. The bottom of the graph is heavily fiction; the less the density of the fictional genre. The further right it goes the more pictures it has (manga or kids comics) and the left is, in return, purely literature.



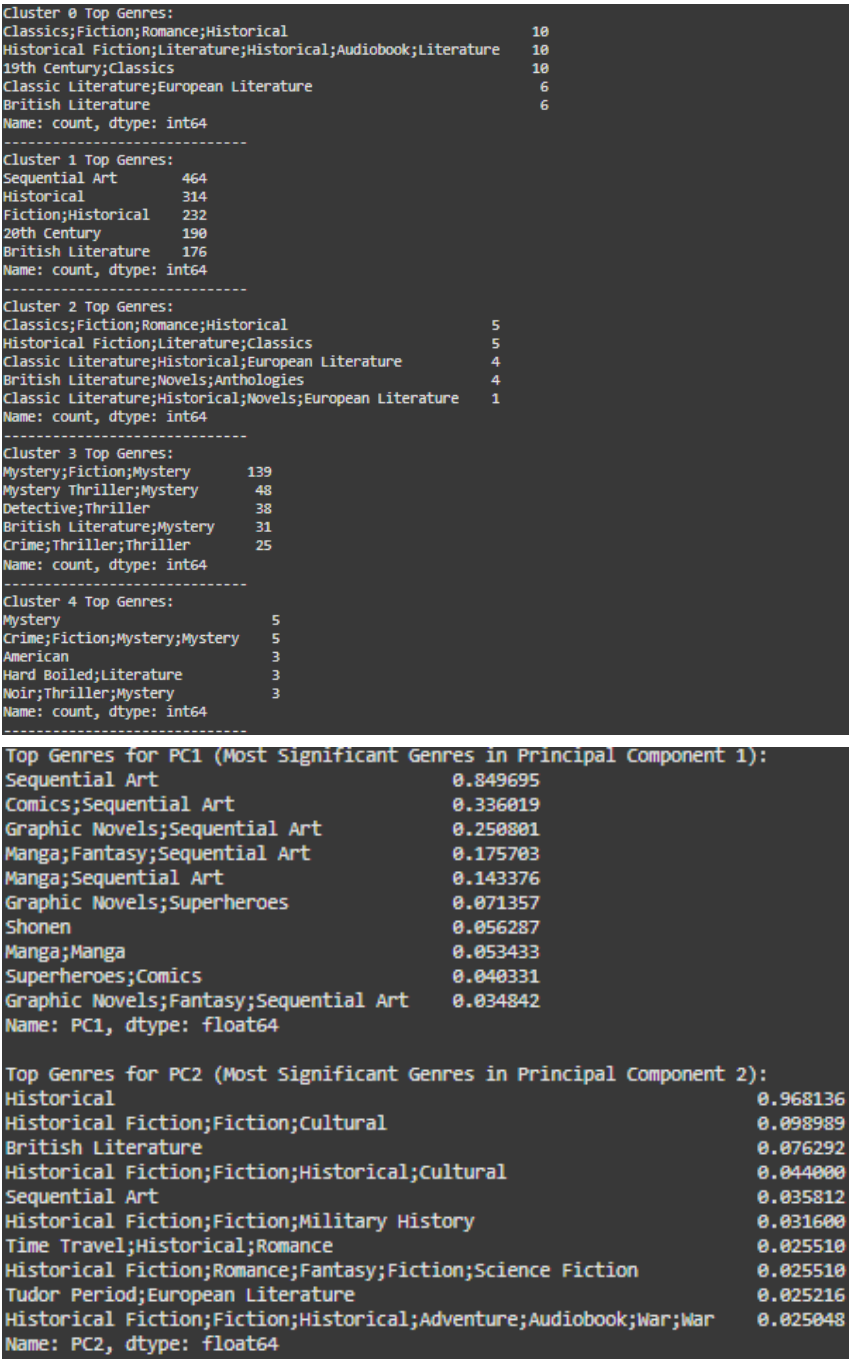


**Graph Explanation:** This is a breakdown of the **Principal Component Analysis (PCA)** **loadings**, showing the most significant genres in **PC1** and **PC2**. For each cluster, the top genres are shown to indicate the types of books in each cluster. The significance of genres in each principal component gives insight into what each cluster represents in terms of genre composition.

**Storytelling Relevance:**

- **Data Exploration:** The loadings help to interpret what each principal component represents. For example, **PC1** may capture the distinction between

**Sequential Art** genres (such as **Comics**, **Manga**) and more traditional genres like **Historical Fiction** and **Literature**. Similarly, **PC2** could separate genres like **Historical** from **Fantasy** or **Science Fiction**.



- **Phenomenon Detection:** Understanding the loadings for each component allows us to detect genre-based patterns that drive the clustering, helping to explain why certain books end up in particular clusters. For example, if **Sequential Art** and **Comics** contribute heavily to **PC1**, it indicates that the component is largely capturing the visual and graphic novel genres.

## Accuracy Analysis

The core hypothesis is that books with similar titles, authors, and genres will be more relevant to users, and the use of cosine similarity or KNN will produce a good recommendation. Though this is purely a subjective matter of similarity, so it can only be accurate to others ratings. This hypothesis is tested and validated by the results of a similarity computation just using taking the standard KNN and Cosine findings and adding weight to their author and genre:

```
# Check if rec_idx is within valid bounds of the cosine similarity matrix
if rec_idx < max_idx:
    # Get the cosine similarity score
    cosine_sim_score = cosine_sim_matrix[selected_idx, rec_idx]

    # Boost similarity based on the same author
    if df.iloc[selected_idx]["Author"] == df.iloc[rec_idx]["Author"]:
        cosine_sim_score *= author_weight # Increase weight if the authors are the same

    # Boost similarity based on the same genres (split genres into lists and compare)
    selected_genres = set(df.iloc[selected_idx]["genres"].split(','))
    rec_genres = set(df.iloc[rec_idx]["genres"].split(','))
    common_genres = selected_genres & rec_genres # Intersection of genres

    cosine_sim_score += genre_weight * len(common_genres) # Add weight for shared genres

    # Ensure cosine_sim_score is between 0 and 1
    cosine_sim_score = min(1, max(0, cosine_sim_score))
```

It works wonders on series as the overlap is boosts the weights to extremey high relevance. Books not contained within a series still have very relevant results, the similarity score just appears to be low as it is scaled in comparisson to that of books in the same series.

## SERIES:


## Books Similar to 'Harry Potter and the Chamber of Secrets (Harry Potter #2)'

---

**Harry Potter and the Prisoner of Azkaban (Harry Potter #3)** by J.K. Rowling/Mary GrandPré

Publisher: Scholastic Inc.

Average Rating: 4.57

 **Similarity Score: 1.00**


Genres: Fantasy, Young Adult, Fiction, Fantasy, Magic, Childrens, Adventure, Audiobook, Childrens, Middle Grade, Classics, Science Fiction Fantasy

---

**Harry Potter Collection (Harry Potter #1-6)** by J.K. Rowling/Mary GrandPré

Publisher: Scholastic Inc.

Average Rating: 4.49

 **Similarity Score: 1.00**


Genres: Fantasy, Young Adult, Fiction, Fantasy, Magic, Childrens, Adventure, Audiobook, Childrens, Middle Grade, Classics, Science Fiction Fantasy

---

**Harry Potter and the Order of the Phoenix (Harry Potter #5)** by J.K. Rowling/Mary GrandPré

Publisher: Scholastic Inc.

Average Rating: 4.56

 **Similarity Score: 0.86**


Genres: Fantasy, Fiction, Young Adult, Fantasy, Magic, Childrens, Childrens, Middle Grade, Adventure, Audiobook, Classics, Science Fiction Fantasy

---

**Harry Potter and the Half-Blood Prince (Harry Potter #6)** by J.K. Rowling/Mary GrandPré

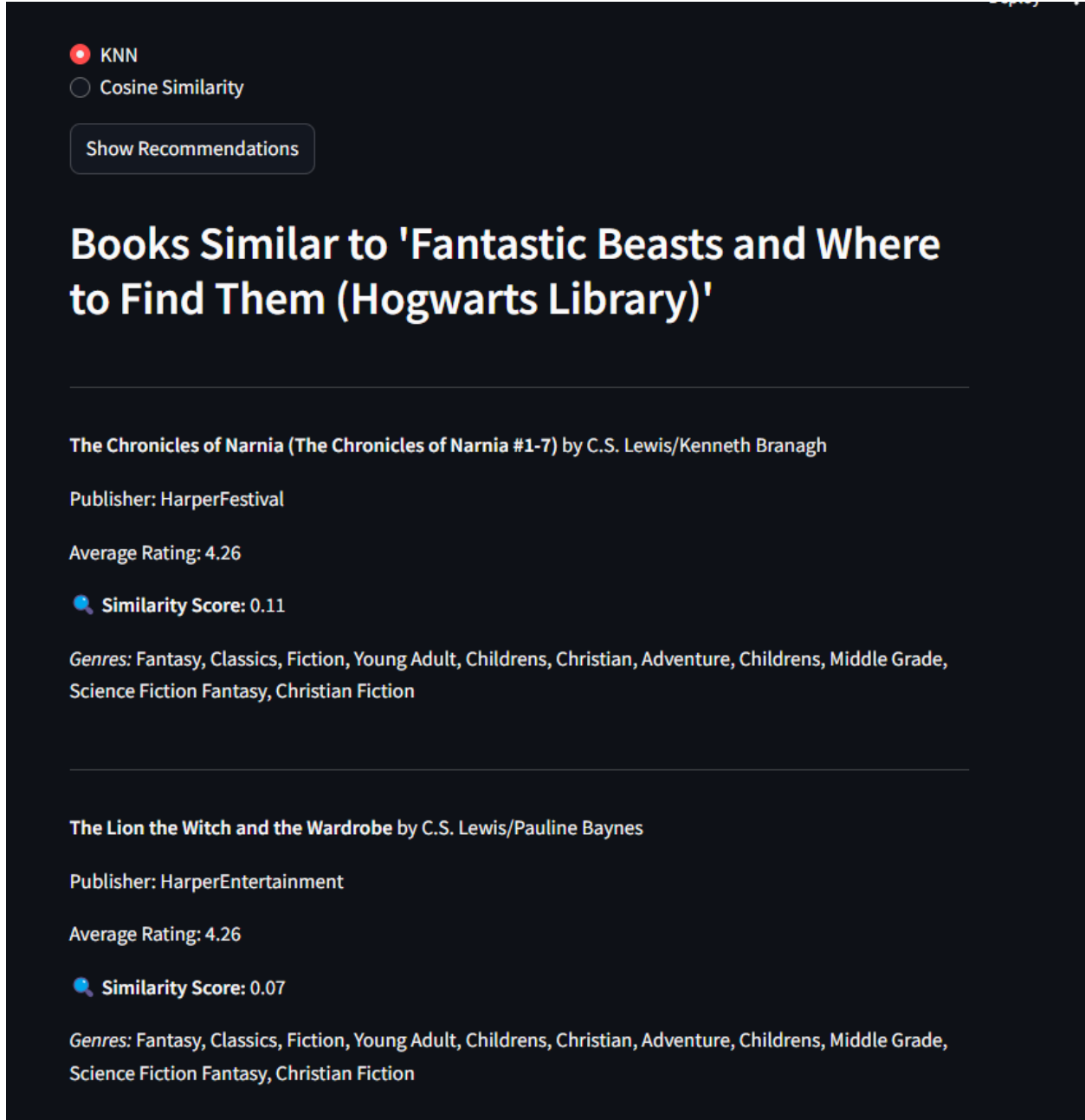
Publisher: Scholastic

Average Rating: 4.78

 **Similarity Score: 0.84**

Genres: Fantasy, Young Adult, Fiction, Fantasy, Magic, Adventure, Fantasy, Supernatural, Mystery, Childrens, Fantasy, Paranormal, Childrens, Middle Grade

## Non Series, Fantasy Example:




The screenshot shows a web application interface with a dark background. At the top left, there are two radio buttons: 'KNN' (selected) and 'Cosine Similarity'. Below them is a button labeled 'Show Recommendations'. The main heading is 'Books Similar to 'Fantastic Beasts and Where to Find Them (Hogwarts Library)''.

**The Chronicles of Narnia (The Chronicles of Narnia #1-7) by C.S. Lewis/Kenneth Branagh**

Publisher: HarperFestival

Average Rating: 4.26

 **Similarity Score: 0.11**


Genres: Fantasy, Classics, Fiction, Young Adult, Childrens, Christian, Adventure, Childrens, Middle Grade, Science Fiction Fantasy, Christian Fiction

---

**The Lion the Witch and the Wardrobe by C.S. Lewis/Pauline Baynes**

Publisher: HarperEntertainment

Average Rating: 4.26

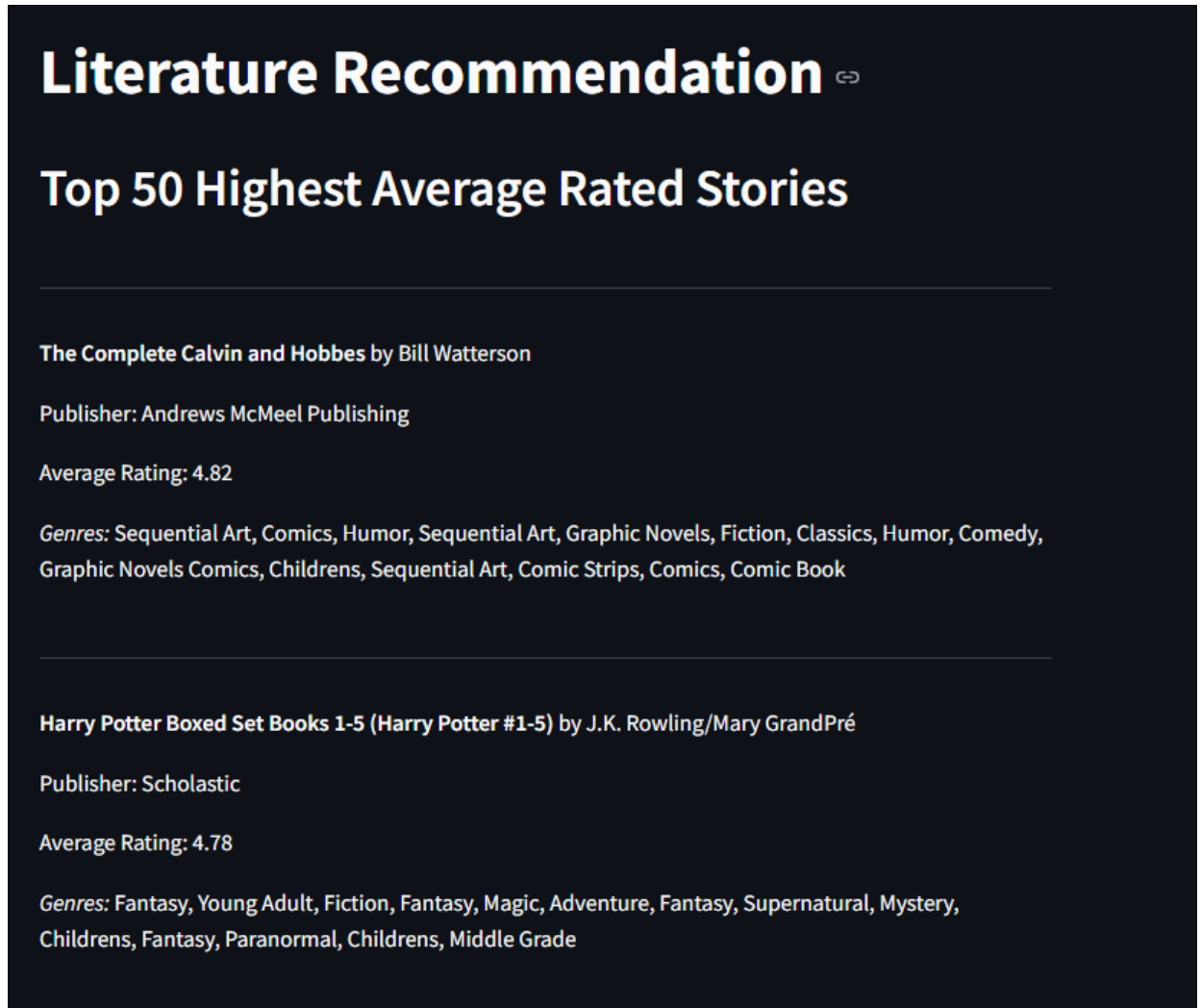
 **Similarity Score: 0.07**

Genres: Fantasy, Classics, Fiction, Young Adult, Childrens, Christian, Adventure, Childrens, Middle Grade, Science Fiction Fantasy, Christian Fiction

## Testing Results, Revisions, and Screenshots

### Initial Testing:

- **Objective:** To test whether the system successfully loads the dataset and generates recommendations based on the ratings.
  - Load the dataset and check for any missing values or inconsistencies (such as incorrect genres).
  - Use just base data to recommend books based on highest rating



**Result:** Initial recommendations show appropriate book suggestions based on highest rating.

**Screenshot 1:** The raw dataset output after loading and basic cleaning (check for missing data or inconsistencies).

**Similarity Score Testing:**

- **Objective:** To test the effectiveness of the similarity model (using cosine similarity or KNN).
  - Select a book from the dataset and compute its similarity score against other books using cosine similarity.
  - Ensure that the highest similarity scores correspond to books with shared genres or similar content.
  - Review the generated recommendations and manually verify if they are appropriate based on genre.

**Result:** Similarity score results are as expected, with relevant books being recommended.

The screenshot shows a web application titled "Literature Recommendation". It features a dark theme with white and light blue text. At the top, there is a section for selecting a book and a recommendation technique. The selected book is "The Fellowship of the Ring (The Lord of the Rings #1)" and the technique is "KNN". Below this, there is a button labeled "Show Recommendations". The main content area displays the details of the selected book and its recommendations. The selected book is "The Fellowship of the Ring (The Lord of the Rings #1) by J.R.R. Tolkien", published by Ballantine Books, with an average rating of 4.59 and a similarity score of 1.00. Its genres are Fantasy, Fiction, Classics, Adventure, Science Fiction Fantasy, Fantasy, Epic Fantasy, Fantasy, High Fantasy, Young Adult, Literature, Fantasy, and Magic. The first recommendation is "J.R.R. Tolkien 4-Book Boxed Set: The Hobbit and The Lord of the Rings by J.R.R. Tolkien", published by Houghton Mifflin Harcourt, with an average rating of 4.5 and a similarity score of 0.47. Its genres are Fantasy, Classics, Fiction, Adventure, Science Fiction Fantasy, Fantasy, Epic Fantasy, Fantasy, High Fantasy, Literature, Novels, and Young Adult.

## Literature Recommendation

Select a book

The Fellowship of the Ring (The Lord of the Rings #1)

Choose recommendation technique

☒ KNN

☐ Cosine Similarity

Show Recommendations

### Books Similar to 'The Fellowship of the Ring (The Lord of the Rings #1)'

---

**The Fellowship of the Ring (The Lord of the Rings #1) by J.R.R. Tolkien**

Publisher: Ballantine Books

Average Rating: 4.59

Similarity Score: 1.00

Genres: Fantasy, Fiction, Classics, Adventure, Science Fiction Fantasy, Fantasy, Epic Fantasy, Fantasy, High Fantasy, Young Adult, Literature, Fantasy, Magic

---

**J.R.R. Tolkien 4-Book Boxed Set: The Hobbit and The Lord of the Rings by J.R.R. Tolkien**

Publisher: Houghton Mifflin Harcourt

Average Rating: 4.5

Similarity Score: 0.47

Genres: Fantasy, Classics, Fiction, Adventure, Science Fiction Fantasy, Fantasy, Epic Fantasy, Fantasy, High Fantasy, Literature, Novels, Young Adult

## Revisions & Optimization:

- **Objective:** To refine the model to ensure accurate recommendations and efficient execution.
- **Process:**
  - Review any false positives or irrelevant recommendations (e.g., books with completely different genres being recommended) as well as repeating books.
  - Adjust the model (e.g., fine-tuning the parameters or implementing genre filters).
  - Test the system again after revisions to see if it improves the relevance of recommendations.
- **Result:** The system runs efficiently even with larger datasets while maintaining high-quality recommendations.

### Books Similar to 'J.R.R. Tolkien 4-Book Boxed Set: The Hobbit and The Lord of the Rings'

---

**J.R.R. Tolkien 4-Book Boxed Set: The Hobbit and The Lord of the Rings** by J.R.R. Tolkien  
Publisher: Ballantine Books  
Average Rating: 4.59  
🔍 **Similarity Score:** 1.00  
Genres: Fantasy, Fiction, Classics, Adventure, Science Fiction Fantasy, Fantasy, Epic Fantasy, Fantasy, High Fantasy, Young Adult, Literature, Fantasy, Magic

---

**The Lord of the Rings (The Lord of the Rings #1-3)** by J.R.R. Tolkien  
Publisher: Houghton Mifflin Harcourt  
Average Rating: 4.5  
🔍 **Similarity Score:** 0.98  
Genres: Fantasy, Classics, Fiction, Adventure, Science Fiction Fantasy, Fantasy, Epic Fantasy, Fantasy, High Fantasy, Literature, Novels, Young Adult

---

**The Fellowship of the Ring (The Lord of the Rings #1)** by J.R.R. Tolkien  
Publisher: Houghton Mifflin Harcourt  
Average Rating: 4.36  
🔍 **Similarity Score:** 0.47  
Genres: Fantasy, Classics, Fiction, Adventure, Fantasy, High Fantasy, Science Fiction Fantasy, Fantasy, Epic Fantasy, Audiobook, Young Adult, Novels

## Source Code

The source code for this Book Recommendation System is hosted on GitHub under the **WGU-Capstone** repository. The organization of the project is as follows:

- **python/ folder:** This directory contains the main components of the application:
  - **main.py:** This is the primary application file, where the Streamlit web interface is launched, and the recommendation system is executed.
  - **recommend.py:** This file implements the core recommendation algorithm, utilizing machine learning techniques such as K-Nearest Neighbors (KNN) and Cosine Similarity to generate recommendations.
  - **data/ folder:** This folder contains the dataset (books.csv), which serves as the input for generating recommendations.

Additionally, any other relevant files are organized within the main repository section.

## Executable Files

The executable component of this application is contained in the **main.py** file. This file serves as the entry point to the Streamlit app, which provides the user interface for interacting with the recommendation system.

To run the application, the user can execute the following command in their terminal:

```
streamlit run python/main.py
```

This will launch the app and allow the user to input books and receive recommended titles based on their similarity.



Furthermore, an optional **setup.py** file is included to facilitate easy installation and setup of dependencies, ensuring a smooth environment setup for running the application. This setup file streamlines the process for users, making it easier to install all necessary libraries and get the system running with minimal effort.

## Quick Start Guide

### 1. Clone the Repository:

First, clone the repository to your local machine by running:

```
git clone https://github.com/SerotoninShane/WGU-Capstone.git  
cd WGU-Capstone/python
```

### 2. Install Python Dependencies:

You'll need to install the required Python dependencies. Run the following command to install everything from requirements.txt:

```
pip install -r python/requirements.txt
```

### 3. Prepare the Dataset:

Ensure the dataset (books.csv) is present in the python/data/ folder. If it's not already there, download it from [Kaggle: Goodreads Books with Genres](#) and place it in the correct location.

### 4. Run the Application:

To launch the app and see the book recommendations, run the following command:

```
streamlit run python/main.py
```

### 5. Use the Application:

The app will open in your browser. After selecting the recommended tab on the left users can interact with it by inputting a book or author name to receive recommendations based on the algorithms like KNN or Cosine Similarity.

## References

- Middlelight. (2021). *Goodreads Books with Genres*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/middlelight/goodreadsbookswithgenres?resource=download>
- Scikit-learn, Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Retrieved from <https://scikit-learn.org/>
- Pandas, McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56. Retrieved from <https://pandas.pydata.org/>
- Streamlit, Streamlit Inc. (2024). Streamlit: The fastest way to build and share data apps. Retrieved from <https://streamlit.io/>