IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# GENRE-CENTRIC BOOK RECOMMENDATION SYSTEM

**[1]Aleena, [2]Dr. Tamizharasi T**

[1]B.Tech Student, [2]Assistant Professor Sr. Grade 1

[1]School of Computer Science and Engineering (SCOPE),

[1]Vellore Institute of Technology, Vellore, India

***Abstract:*** This project delves into the world of literature, focusing on the dynamic relationship between authors, genres, and readers. By harnessing social and information network analysis techniques, the project seeks to uncover intricate connections within the literary landscape. In this project a comprehensive Book-dataset is analyzed to uncover the relationships between authors, genres, and reader preferences. It starts with data collection and preprocessing, followed by the construction of a bipartite network that links authors to genres based on their literary works. Through advanced analytics, the project identifies author-genre associations and genre-based author communities. The culmination of this effort is a robust recommendation system that suggests books to readers based on the communities formed. This project impacts both authors and genres by helping them reach their target audiences more effectively. Authors benefit from increased visibility within their specialized genres, while genres themselves thrive with enhanced readership. For readers, it means discovering books that align with their interests and exploring new authors within the same genre. In the future these well-defined author-genre communities have the potential to influence book searches on search engines, recommendations on e-commerce platforms, and advertisements on social media.

***Index Terms*** **- Literature networks, bipartite network analysis, author-genre associations, recommendation system, genre-based communities, reader preferences.**

## I. INTRODUCTION

In recent years, the explosive growth of data in the realm of literature and e-commerce, coupled with the increasing reliance on information technology and data science, has prompted a fundamental shift in how readers engage with books, authors, and genres. The intricate relationships between authors and genres play a pivotal role in shaping the literary landscape, yet understanding and harnessing these connections remain a complex challenge. In response to this, the "Genre-Centric Book Recommendation System" is proposed, a project situated at the intersection of information technology and data science.

Much like community detection algorithms in complex network analysis, my project seeks to uncover and leverage the inherent structure within the literary domain. While community detection has proven instrumental in deciphering relationships in social networks and Protein-Protein Interaction (PPI) networks, its application to the dynamic interplay of authors, genres, and readers in literature remains largely unexplored. The existing literature on community detection primarily focuses on social networks, with an emphasis on scalability issues as network sizes grow exponentially. Traditional algorithms, although effective, struggle to handle the vast scale of today's social networks. The need for efficient, scalable community detection methods has spurred research in parallel algorithms, local community discovery, and network scale reduction-based algorithms. In the context of literature and e-commerce, existing algorithms often overlook the nuances of author-genre associations and fail to capitalize on the unique characteristics of the literary landscape. This project aims to bridge this gap by introducing a novel Genre-Centric Book Recommendation System that goes beyond traditional community detection approaches. Social and information network analysis techniques are leveraged to delve into the relationships between authors, genres, and reader preferences.

The proposed architecture of my system involves a meticulous process, starting with data preprocessing to ensure the integrity of the book dataset. Subsequent steps include data analysis and network construction, author-genre association analysis, identification of genre-based author communities, and the development of a recommendation system. Each module is designed to contribute to the overarching goal of enhancing reader engagement, benefiting both authors and genres by optimizing visibility and readership. In contrast to conventional book recommendation systems that predominantly rely on user preferences and collaborative filtering, this study proposes a groundbreaking Genre-Centric Book Recommendation System that capitalizes on the intricate relationships between authors and genres within the literary landscape. Unlike existing algorithms that often overlook the nuanced associations between authors and genres, the methodology places a central emphasis on author-genre bipartite graphs. By comprehensively analyzing these graphs, we gain deeper insights into the interplay between literary creators and the genres they explore, allowing for a more nuanced understanding of reader preferences. This genre-centric approach ensures that the recommended books not only align with individual tastes but also resonate with the thematic elements and writing styles characteristic of specific genres. Moreover, the methodology incorporates a multi-faceted analysis, encompassing community detection, weighted edge computations, and joint probability estimations, offering a holistic perspective on the intricate web of

author-genre relationships. By considering factors such as author communities and the joint probability of author-genre occurrences, the system surpasses traditional recommendation models by providing a more refined and context-aware set of book suggestions. Through the integration of these advanced techniques, the Genre-Centric Book Recommendation System aims to address the limitations of conventional approaches, offering readers a more personalized and insightful literary experience based on the profound connections between authors and genres.

This project presents a Genre-Centric Book Recommendation System at the intersection of information technology and data science. Leveraging community detection algorithms, the system explores intricate relationships between authors and genres, offering a nuanced understanding of reader preferences. Through meticulous data preprocessing, network construction, and author-genre association analysis, my approach surpasses conventional models by considering weighted averages of book ratings, joint probabilities, and community structures. The methodology aims to enhance reader engagement and benefit authors and genres by optimizing visibility and readership. The Genre-Centric Book Recommendation System represents a refined and context-aware solution, addressing the limitations of traditional recommendation models.

This paper is organized as follows: Section II reviews existing literature on community detection in social networks, providing context for the unique challenges posed by the literary domain. Section III presents the detailed architecture and methodology of the proposed Genre-Centric Book Recommendation System. Section IV discusses the results of implementation of the proposed system. Finally, Section V concludes the paper, highlighting key findings and outlining avenues for future research.

## II. LITERATURE REVIEW

1. *A community detection algorithm based on graph compression for large-scale social networks [1]*
   As social networks grow in size, traditional algorithms to uncover the community structure become less effective due to their time and spatial complexity. The proposed algorithm employs a technique called graph compression. It iteratively merges vertices with a low degree (1 or 2) into their neighbors with higher degrees, creating a compressed graph. After identifying communities in the compressed graph, the community structure is expanded and propagated back to the original social network. The proposed algorithm is suitable for undirected networks but may not work well for attribute networks or multilayer networks.
2. *Personalized Book Recommendation System using Machine Learning Algorithm [2]*
   The concept discussed here is about improving the recommendation of books to online users. Traditional recommendation systems are often based on user ratings, which can be problematic when users unsubscribe from the service or stop rating books. To address this, the paper proposes a novel book recommendation system that relies on clustering methods. It clusters books based on user ratings and preferences and then finds similarities between these clusters to suggest new books. The system uses k-means clustering and cosine distance measures to group books into clusters.
3. *Leveraging genre classification with RNN for Book recommendation [3]*
   This paper discusses ways to improve the recommendation of books and movies by utilizing user-generated content in the form of reviews. Traditional recommendation algorithms face limitations when they do not consider the combination of reviews and genre for books. In this context, the paper proposes the use of Recurrent Neural Networks (RNNs) as a deep learning approach to enhance the classification of book plots and reviews into various categories and provide more accurate recommendations to users. RNNs offer an advantage over traditional models because they allow each neuron to utilize its internal memory to retain information about previous inputs. This capability enables the model to maintain context between reviews, leading to more accurate classification and, consequently, better recommendations.
4. *Content-Based Movie Recommendation System Using Genre Correlation [4]*
   The paper presents a content-based movie recommendation system that focuses on genre correlation to provide personalized movie suggestions to users. The paper categorizes recommendation systems into three main types: collaborative filtering, content-based filtering, and hybrid systems. It then delves into content-based filtering, which revolves around analyzing a user's past behavior and recommending items with similar attributes. In this case, the focus is on recommending movies based on their genre. To create the recommendation system, they first construct a matrix that combines movie genres and user ratings, converting both into binary values for consistency. The recommendation algorithm calculates the dot product of this matrix to determine user preferences for specific genres and movies. Then Euclidean distance is used to find the similarity between users' preferences and recommends movies with the least deviation from the current user's choices.
5. *Advanced Graph Analytics Algorithms On Genre Based Recommending System [5]*
   The paper titled "Advanced Graph Analytics Algorithms on Genre-Based Recommending System" discusses the development of a novel ranking method to enhance the accuracy of recommendations in genre-based movie or content recommendation systems, particularly in the context of online streaming platforms like Netflix. The primary challenge addressed in the paper is the time bias present in traditional recommendation systems, which tend to favour older content over recent releases, even when newer content may have better quality. The proposed ranking method applies an "Equality Rebalance Methodology" to minimize the time bias in bipartite graphs, where users and content are represented as nodes. The researchers conducted experiments using real-world datasets as benchmarks and found that their proposed methodology improved predictive performance or accuracy by at least 20% and up to 80% while reducing time bias in ranking scores.
6. *Link prediction in Social Network Analysis: Steps and Algorithms [6]*
   This project explores the significance of link prediction in social networks and its broader applications across diverse fields. The research aims to assess various methods and algorithms employed for link prediction, emphasizing the prediction of missing links in current networks and the formation or dissolution of ties in future networks. The methodology involves a thorough examination of link prediction in social networks through comprehensive testing, analysis, and discussions, encompassing both statistical and machine learning approaches. The project utilizes social network datasets from platforms like social media, citation networks, email communication networks, and online communities. It compiles a diverse range of methods for link prediction, including topology-based metrics and node-based algorithms, with a focus on binary networks. However, it excludes networks with weighted or directed edges from consideration. One key finding suggests that existing evaluation methods might be insufficient for assessing the performance of link prediction algorithms. Despite this limitation, the study offers valuable insights into the evolution of social networks, presents a wide array of prediction methods, and outlines potential challenges and future directions in the field.

7. *Building a predictor for movie ratings [7]*
   The paper investigates the correlation between movie popularity and its cast and genre, proposing that a movie's popularity is significantly influenced by these factors. Network analysis techniques are employed to reveal the underlying structure of relationships among movies, actors, directors, and genres. Utilizing the open-source IMDb dataset, the authors construct bipartite graphs connecting movies to actors and genres. The HITS algorithm is then applied to assign hub and authority scores to actors and directors, predicting their future project's popularity. A similar technique, incorporating movie ratings, is employed to forecast the popularity of specific genres. PageRank is also utilized on the bipartite graphs, offering additional insights into their structural characteristics. The IMDb dataset is randomly split into a training set (75% of movies) and a test set (25% of movies). The study demonstrates that the HITS algorithm serves as an accurate predictor of movie popularity, revealing an average absolute difference of 0.8668 between predicted and true ratings. Although the predictor exhibits a slight positive bias, with a predicted average rating of 6.57 compared to the true average of 6.41 in the test set, the results highlight its effectiveness. Notably, the study identifies a tendency for over-rating low-rated movies in the prediction outcomes.
8. *Predicting IMDB Movie Ratings Using Social Media [8]*
   The paper delves into cross-channel prediction tasks within information retrieval systems, particularly focusing on predicting movie ratings through signals obtained from various social media channels. The authors employ textual feature extraction, comparing the log-likelihood of terms in tweets and YouTube comments related to top- and bottom-rated movies on IMDb. The dataset comprises 70 films with IMDb ratings as of April 4, 2011. Textual features are extracted from 10 selected films, leaving 60 for testing. The best-performing model, integrating textual data, exhibits high predictive performance and alignment with observed ratings. Quantitative and qualitative indicators from social media channels, such as Twitter and YouTube, are analyzed to derive surface features and assess the meaning of activity around movie titles. Regression analysis, utilizing WEKA toolkit's linear regression implementation, is employed for predicting movie ratings, validated through ten-fold cross-validation on the set of 60 movies. Additional social media applications like Reddit, Instagram, and Facebook are considered for optimal performance. Positive and negative textual features, as well as user demographics, are explored to enhance predictive methods, aiming to forecast ratings well into the future.
9. *Movies Recommendation Networks as Bipartite Graphs [9]*
   The author collected data to construct three distinct networks for analysis: IMDb recommendations (IMDb), User-driven bipartite network (UD-BP), and One-mode projection of user-driven network (UD-OM). These networks, built from information on over 43,000 movies and 350,000 consumers, exhibited high clustering coefficients and small-world properties. Despite some variations, the author observed universal degree distributions for both types of networks. The study aimed to generalize the presented concepts by exploring the underlying natural causes for the identified network traits. Utilizing collaborative filtering and IMDb data analysis methodologies, the author examined user clustering, community structures, and theoretical modeling. The research suggested that power-law distributions in networks were not solely due to favoring more popular movies but might involve self-organizing mechanisms. The analysis also indicated variations in behavior for different network sizes, emphasizing the robustness of the observed degree distributions.
10. *Book Genre Classification Based on Reviews of Portuguese-Language Literature [10]*
    The paper explores the application of machine learning techniques in classifying Portuguese-language literature into various genres based on reviews. The methodology involves data collection, pre-processing to eliminate unnecessary information, feature extraction (e.g., word frequency, sentiment analysis), model training with algorithms like Naive Bayes or Support Vector Machines, model evaluation, and classification of new reviews. The classification can offer valuable insights for readers, publishers, and authors, aiding them in understanding audience preferences. The success depends on data quality, accurate feature extraction, and the algorithm's effectiveness. A relevant study on Bengali text classification is mentioned, employing similar methods for genre classification in Bengali texts. However, potential challenges include overfitting, language limitations (specific to Portuguese), and the need for careful consideration of evaluation metrics to accurately assess model performance. The methodology's effectiveness may vary based on language and genre-specific characteristics.

## III. METHODOLOGY

The proposed "Genre-Centric Book Recommendation System" is driven by a comprehensive methodology that navigates the intricate relationships between authors, genres, and readers within the literary realm. The project unfolds through a systematic process encompassing data collection, preprocessing, network analysis, community detection, recommendation system development, and thorough testing and validation.

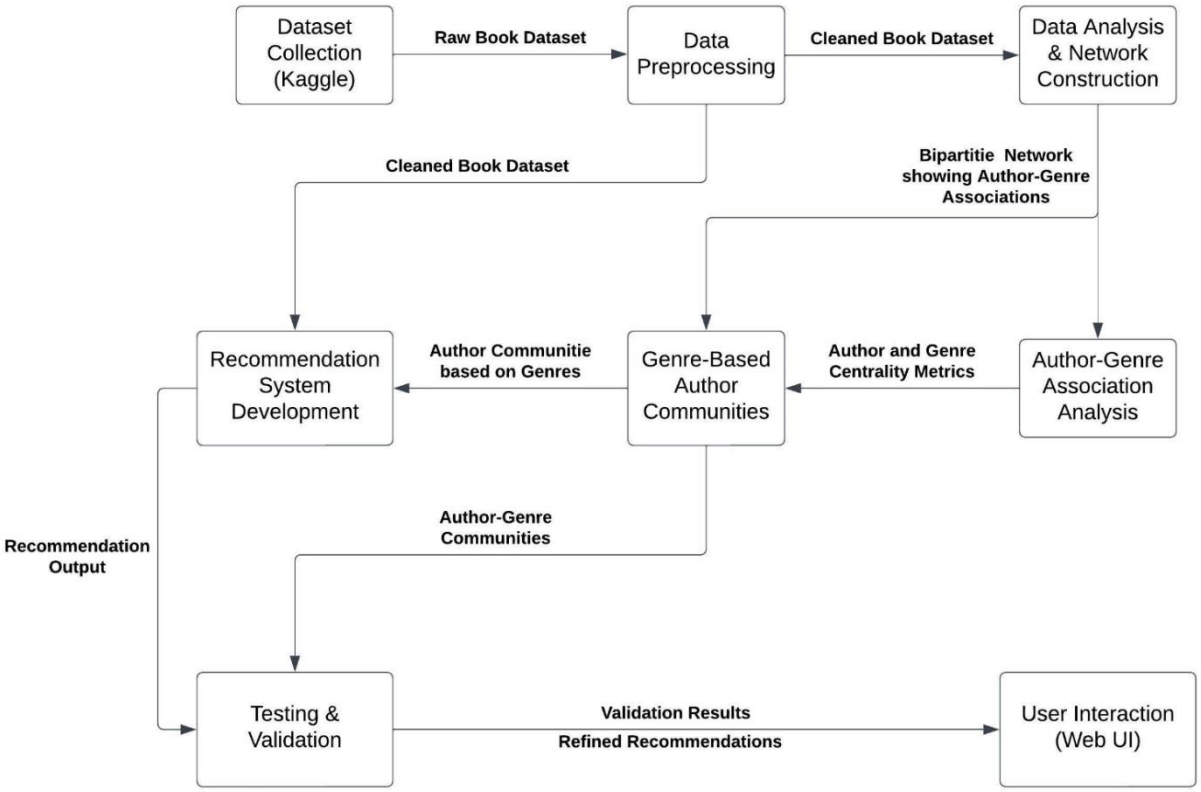



Figure 1 Data Flow Diagram of the Genre-Centric Book Recommendation System

The methodology commences with a detailed exploration of Dataset Collection, a pivotal phase that sets the stage for subsequent analyses. The dataset, a treasure trove of literary information, was meticulously sourced from Kaggle, encompassing 54,301 books with a diverse array of attributes. These include critical details such as book_authors, book_description, book_edition, book_format, book_isbn, book_pages, book_rating, book_rating_count, book_review_count, book_title, genres, and image_url. This rich dataset, originating from a reputable platform like Kaggle, ensures the project's foundation is built upon high-quality and real-world literary data. Moving seamlessly from Dataset Collection to Data Preprocessing, the Pandas library is employed to load and cleanse the dataset systematically. This module ensures data consistency by handling duplicates, missing values, and inconsistencies, laying a robust foundation for subsequent analyses. The cleaned dataset becomes the cornerstone for the ensuing exploration into the dynamic relationships between authors, genres, and readers.

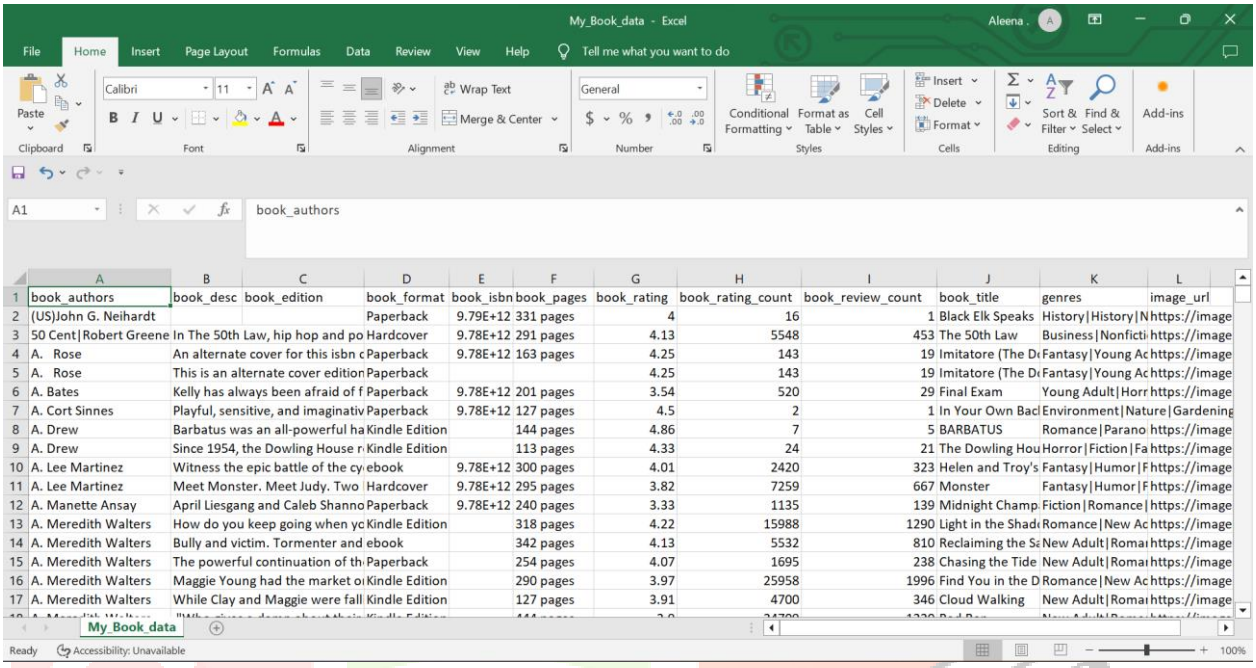
Following Data Preprocessing, the journey unfolds into the Data Analysis and Network Construction module. Leveraging the NetworkX library, the pre-processed data undergoes rigorous analysis to uncover intricate author-genre associations. The construction of a bipartite network, linking authors and genres based on book authorship, is visualized using the Matplotlib library. This visual representation provides insights into the complex web of connections within the literary landscape. Transitioning seamlessly, the Author-Genre Association Analysis module employs the NetworkX library to manipulate the bipartite network. Calculating degree centrality metrics of authors in the genre layer is pivotal in identifying genres commonly associated with specific authors. This analysis serves as a foundation for understanding the key players in the literary ecosystem. The Genre-Based Author Communities module takes center stage, transforming the bipartite network into an author-author projected network using the NetworkX library. Community detection algorithms such as Louvain and Girvan-Newman are applied to identify genre-based author communities. Matplotlib and Seaborn libraries are then utilized for visualization, creating heatmaps or dendrogram plots that vividly illustrate the relationships within these author communities.

The journey culminates in the Recommendation System Development module, where logic is implemented to suggest books to users based on identified author-genre communities. A web-based user interface, crafted using HTML, CSS, and JavaScript, facilitates user interaction and provides tailored book recommendations aligned with individual preferences. To ensure the effectiveness of the developed recommendation system, a comprehensive Testing and Validation module is introduced. A mathematical algorithm is devised to compare the output recommendations with the detected author-communities, providing a quantitative measure of the system's impact on enhancing readership for authors and genres.

In summary, the methodology intricately navigates the literary landscape, utilizing advanced analytics and systematic processing to unravel the complex relationships between authors and genres. Each module serves as a building block, contributing to the development of a robust recommendation system poised to impact authors, genres, and readers within the literary realm.

### 3.1 Data Preprocessing

This module cleans and preprocesses a dataset of book data, handling missing values, converting data types, filtering out non-English content, and processing genres before saving the processed data to a new CSV file. The code begins by importing necessary libraries, such as pandas and numpy. It loads a dataset from a CSV file ("My_Book_data.csv") into a Pandas DataFrame (df). Initial exploratory data analysis is done using `head()`, `shape`, `size`, `info()`, and `columns` to understand the structure and content of the dataset. The code defines a list of columns to keep (`filter_cols`) and removes the columns not in this list from the DataFrame. Unimportant columns such as book_description, book_edition, book_format, book_isbn, image_url are filtered out. The resulting DataFrame (`df_filtered`) is displayed, and its information is printed. Rows with missing values are removed from `df_filtered` using `dropna()`. The 'book_pages' column, initially of object type, is converted to integers after removing the "pages" suffix from each element. Some cells in the 'book_authors' column have multiple authors separated by "|". The code defines a function (`removeMultipleAuthors`) to remove rows with multiple authors. Two functions (`is_english` and `removeOtherLangs`) are defined to filter out rows where 'book_authors' or 'book_title' are not in English. Rows with non-English content are removed from the DataFrame. The 'genres' column, initially containing pipe-separated strings, is converted to lists using the `changeDelimiter` function. Duplicate genres within each row are removed, and the lists are sorted. The processed lists are converted back to strings and assigned to the 'genres' column. The processed DataFrame is saved to a new CSV file ("Processed_BookData.csv") using `to_csv()`.




| | book_authors | book_desc | book_edition | book_format | book_isbn | book_pages | book_rating | book_rating_count | book_review_count | book_title | genres | image_url |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | (US)John G. Neihardt | | | Paperback | 9.79E+12 | 331 pages | 4 | 16 | 1 | Black Elk Speaks | History\|History\|N | https://image |
| 3 | 50 Cent\|Robert Greene | In The 50th Law, hip hop and po | | Hardcover | 9.78E+12 | 291 pages | 4.13 | 5548 | 453 | The 50th Law | Business\|Nonficti | https://image |
| 4 | A. Rose | An alternate cover for this isbn c | | Paperback | 9.78E+12 | 163 pages | 4.25 | 143 | 19 | Imitatore (The D | Fantasy\|Young Ac | https://image |
| 5 | A. Rose | This is an alternate cover edition | | Paperback | | | 4.25 | 143 | 19 | Imitatore (The D | Fantasy\|Young Ac | https://image |
| 6 | A. Bates | Kelly has always been afraid of f | | Paperback | 9.78E+12 | 201 pages | 3.54 | 520 | 29 | Final Exam | Young Adult\|Horr | https://image |
| 7 | A. Cort Sinnes | Playful, sensitive, and imaginativ | | Paperback | 9.78E+12 | 127 pages | 4.5 | 2 | 1 | In Your Own Bac | Environment\|Nature\|Gardening | |
| 8 | A. Drew | Barbatus was an all-powerful ha | | Kindle Edition | | 144 pages | 4.86 | 7 | 5 | BARBATUS | Romance\|Parano | https://image |
| 9 | A. Drew | Since 1954, the Dowling House r | | Kindle Edition | | 113 pages | 4.33 | 24 | 21 | The Dowling Hou | Horror\|Fiction\|Fa | https://image |
| 10 | A. Lee Martinez | Witness the epic battle of the cy | | ebook | 9.78E+12 | 300 pages | 4.01 | 2420 | 323 | Helen and Troy's | Fantasy\|Humor\|F | https://image |
| 11 | A. Lee Martinez | Meet Monster. Meet Judy. Two | | Hardcover | 9.78E+12 | 295 pages | 3.82 | 7259 | 667 | Monster | Fantasy\|Humor\|F | https://image |
| 12 | A. Manette Ansay | April Liesgang and Caleb Shanno | | Paperback | 9.78E+12 | 240 pages | 3.33 | 1135 | 139 | Midnight Champ | Fiction\|Romance\| | https://image |
| 13 | A. Meredith Walters | How do you keep going when yc | | Kindle Edition | | 318 pages | 4.22 | 15988 | 1290 | Light in the Shad | Romance\|New Ac | https://image |
| 14 | A. Meredith Walters | Bully and victim. Tormenter and | | ebook | | 342 pages | 4.13 | 5532 | 810 | Reclaiming the Se | New Adult\|Romar | https://image |
| 15 | A. Meredith Walters | The powerful continuation of th | | Paperback | | 254 pages | 4.07 | 1695 | 238 | Chasing the Tide | New Adult\|Romar | https://image |
| 16 | A. Meredith Walters | Maggie Young had the market o | | Kindle Edition | | 290 pages | 3.97 | 25958 | 1996 | Find You in the D | Romance\|New Ac | https://image |
| 17 | A. Meredith Walters | While Clay and Maggie were fall | | Kindle Edition | | 127 pages | 3.91 | 4700 | 346 | Cloud Walking | New Adult\|Romar | https://image |

Figure 2 Raw Book Dataset from Kaggle

### 3.2 Network Construction

This module focuses on creating and analyzing bipartite graphs for 1% and 10% samples of the processed book dataset. It calculates weights based on various metrics and visualizes the resulting graphs, providing insights into the relationships between authors and genres in the sampled data. The code begins by importing necessary libraries: pandas, networkx, matplotlib, random, and pickle. It loads a previously processed dataset from a CSV file ("Processed_BookData.csv") into a Pandas DataFrame (df). The code extracts unique genres from the 'genres' column in the DataFrame, creating a set (`genreList`). The total number of unique genres is printed. A 1% sample of the dataset (`df_1_sample`) is created for testing purposes using random sampling. The genres column is transformed into a list of genres. A bipartite graph (`B_1_sample`) is created using NetworkX. Nodes are added for authors and genres. Edges are added with weights calculated based on specified metrics, including ratings, rating counts, review counts, and probabilities. The resulting graph is saved using the `writeGraph` function. The code visualizes the bipartite graph using matplotlib, with nodes representing authors and genres, and edges representing connections between them. Weights are displayed on the edges. A 10% sample of the dataset (`df_10`) is created for further analysis using random sampling. Like the 1% sample, the genres column is transformed into a list of genres. A new bipartite graph (`B_10`) is created for the 10% sample. Nodes, edges, and weights are added similar to the process for the 1% sample. The resulting graph is saved using the `writeGraph` function. The code visualizes the bipartite graph for the 10% sample using matplotlib, with nodes representing authors and genres, and edges representing connections between them. Weights are displayed on the edges.

***Author-Genre Weight Computation***

Say there is an author A1 who has written books B111, B112, and B113 of genre G1. The bipartite graph will therefore include an edge <A1,G1>. In the dataset each book has a book rating, book rating count and book review count. In order to estimate the overall rating that people would give for the A1-G1 pair, the weighted average of the book ratings for B111, B112 and B113 will be computed, where the weights will be the averages of the corresponding book rating counts and book review counts.

*Table 1 Author Genre Edge Weight Computation Example*

| Book | Book rating | Book Rating Count | Book Review Count | Book weight |
|---|---|---|---|---|
| B111 | R111 | RC111 | ReC111 | BW111 = (RC111+ReC111)/2 |
| B112 | R112 | RC112 | ReC112 | BW112 = (RC112+ReC112)/2 |
| B113 | R113 | RC113 | ReC113 | BW113 = (RC113+ReC113)/2 |

$$Rating(1,1) = \frac{R111 * BW111 \ + \ R112 * BW112 \ + \ R113 * BW113}{BW111 \ + \ BW112 \ + \ BW113} \quad (1)$$

Therefore, in general –

$$\text{Rating}(a, g) = \frac{\sum_{k=0}^{nag} R_{agk} \ * \ BW_{agk}}{\sum_{k=0}^{nag} BW_{agk}} \quad (2)$$

where, Rating(a,g) is the overall rating given to the author a and genre g pair, nag is the number of books written by author a of genre g, Ragk is the Book rating of the book k of genre g written by author a, and BWagk is the book weight of the book k of genre g written by author a. The author A1 would have also written several other books B121, B131, and more, of other genres G2, G3, and so on. Similarly, there will be several other books B211, B311, and more, of genre G1 which have been written by other authors A2, A3, and so on. Hence, in order to compute the probability of seeing an A1G1 pair in the dataset, a Joint Probability will be calculated, which will be the product of two probabilities (assuming the two probabilities as being independent of each other) PAG(a,g) and PGA(a,g). Let the number of books written by author A1 of genre G1 be C11, and total number of books written by author A1 be CA1. Then, PAG(a,g) is the probability of finding a book written by an author 'a' which has the genre 'g' -

$$PAG(1,1) = \frac{C11}{C_A 1} \quad (3)$$

Therefore, in general –

$$PAG(a, g) = \frac{Cag}{C_A a} \quad (4)$$

where, Cag is the number of books written by author a of genre g CAa is the number of books written by author a (all genres). Let the number of books of genre G1 written by author A1 be C11, and total number of books of genre G1 be CG1. Then, PGA(g,a) is the probability of finding a book of genre G1 which is written by author A1 -

$$PGA(1,1) = \frac{C11}{C_G 1} \quad (5)$$

Therefore, in general –

$$PGA(a, g) = \frac{Cag}{C_G g} \quad (6)$$

where, Cag is the number of books written by author *a* of genre *g*, and $C_G$g is the number of books of genre g (all authors). Thus, the Joint Probability of seeing a particular author-genre pair in the dataset is –

$$P(a, g) = \ PAG(a, g) * \ PGA(a, g) \quad (7)$$

$$P(a, g) = \frac{Cag^2}{C_A a * C_G g} \quad (8)$$

where, P(a,g) is the joint probability of an author a and genre g pair occurring in the dataset, Cag is the number of books written by author a of genre g, $C_A$a is the number of books written by author a (all genres), and $C_G$g is the number of books of genre g (all authors). The weight of a particular author-genre pair will be directly proportional to the rating given to it and will also be directly proportional to the probability of its occurrence, so the overall weight will be a product of these two –

$$W(a, g) = \ Rating(a, g) * \ P(a, g) \quad (9)$$

$$W(a,g) = \frac{\sum_{k=0}^{nag} R_{agk} * BW_{agk}}{\sum_{k=0}^{nag} BW_{agk}} * \frac{Cag^2}{C_A a * C_G g} \quad (10)$$

where, Cag is the number of books written by author a of genre g, $C_A$a is the number of books written by author a (all genres), and $C_G$g is the number of books of genre g (all authors).

### 3.3 Author Genre Association

This module performs community detection and analysis on the bipartite graph of 1% sample of the dataset, representing relationships between authors and genres in the dataset. The resulting communities, link weights, and common genres are visualized and saved for further analysis. The code reads a previously saved 1% sample bipartite graph from a pickle file ("1%_Sample_author_genre_bipartite.gpickle") using NetworkX. The bipartite graph is visualized using matplotlib. Nodes are categorized into genres and authors based on their bipartite attribute. The unique genres and authors are printed along with their counts. The bipartite graph is projected onto the set of authors to create a new graph (`sample_author_projection`). Degree centrality is calculated for authors. Louvain community detection algorithm is applied to identify communities within the projected graph. Degree centrality and associated genres are printed for each author. The projected graph is visualized using matplotlib, showing connections between authors. Weights are calculated for each author-author edge in the projected graph based on common genres and their weights in the original bipartite graph. A new graph (`C`) is created to represent the communities, including edges with the computed weights. The community network is visualized using matplotlib. Data is collected for each author-author edge in the community network, including community ID, link weight, and common genres. The data is stored in a DataFrame (`communities_df`) and printed. A new community network (`C1`) is created by excluding inter-community edges. The filtered community network is visualized, and the data is collected and stored in a new DataFrame (`communities_df2`).

***Author-Author Edge Weight Computation***

Let there be authors A1 and A2 in the bipartite graph. A1 is connected to genres G1, G2 and G3, while A2 is connected to genres G1, G3, and G4. Then in the author-projected graph there will be an <A1,A2> edge because of genres G1 and G3. Hence, the weight of this edge in the projected network is defined as such,

$$W_P(A1,A2) = \frac{W(A1,G1) * W(A2,G1) + W(A1,G3) * W(A2,G3)}{2} \quad (11)$$

Therefore, in general –

$$W_P(x,y) = \frac{\sum_{k=0}^{n} W(x,G_k) * W(y,G_k)}{n} \quad (12)$$

where, n is the number of common genres between authors x and y, and W(x,Gk) is the weight of the edge between author x and genre Gk in the bipartite graph, similarly W(y,Gk) is the weight of the edge between author y and genre Gk in the bipartite graph.

### 3.4 Recommendation System

This module provides an interactive book recommendation system based on the user's input, leveraging community data from previously analyzed graph. It allows the user to interactively search for a book and receive recommendations for similar books. The recommendations are influenced by the community structure, suggesting books by authors within the same community as the selected book's author. The code loads a processed book dataset from "Processed_BookData.csv" using pandas. A random 1% sample of the dataset (`df_sample`) is created for testing. The code loads a CSV file containing community data ("Communities_1%_BookData.csv") that was previously generated. The user is presented with a list of books to choose from (output of `showBooks()` function). The user is prompted to enter a book title and the number of book recommendations to get. The user's input (book title and number of recommendations) is taken. The details of the selected book are displayed, including author, number of pages, rating, rating count, and genres. The function `get_all_authors` retrieves other authors from the same community as the selected book's author. The function `get_n_books` provides book recommendations based on the authors obtained from the community. The recommendations are displayed, including the book title and author. The user can input the book title and the number of recommendations they want to receive. Output: The code outputs details about the selected book and a list of recommended books along with their authors.

## IV. Results and Discussion

To gauge the effectiveness of the recommendation system, a rigorous testing and validation process is executed. A subset of the dataset is held as a test set, with the remaining portion utilized for training. The true average rating in the test set is compared to my predicted average rating, showcasing the system's accuracy. Additionally, a baseline prediction is established using the median rating for all books in the dataset, serving as a benchmark for comparison.

### 4.1 Preprocessing

The Data Preprocessing Module starts with the input of raw book data in CSV format. Its primary goal is to transform and clean the dataset, resulting in a processed dataset saved as "Processed_BookData.csv." Several key steps are performed, including the removal of unnecessary columns, handling missing values, and converting the "book_pages" column from object type to integer. Multiple data cleaning tasks are carried out, such as removing rows with multiple authors, filtering out non-English author names and book titles, and converting the genre information from a "|" separated string to a list. The final step involves saving the processed dataframe into a new CSV file for subsequent analysis.

```
df.head()
```

| | book_authors | book_desc | book_edition | book_format | book_isbn | book_pages | book_rating | book_rating_count | book_review_count | book_title | genres | image_url |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | (US)John G. Neihardt | NaN | NaN | Paperback | 9.79E+12 | 331 pages | 4.00 | 16 | 1 | Black Elk Speaks | History\|History\|Native American History | https://images.gr-assets.com/books/1530306364l... |
| 1 | 50 Cent\|Robert Greene | In The 50th Law, hip hop and pop culture icon ... | NaN | Hardcover | 9.78E+12 | 291 pages | 4.13 | 5548 | 453 | The 50th Law | Business\|Nonfiction\|Self Help\|Psychology\|Philo... | https://images.gr-assets.com/books/1348220810l... |
| 2 | A. Rose | An alternate cover for this isbn can be found ... | NaN | Paperback | 9.78E+12 | 163 pages | 4.25 | 143 | 19 | Imitatore (The Donna Chronicles #1) | Fantasy\|Young Adult\|Adventure\|Adult | https://images.gr-assets.com/books/1515218745l... |
| 3 | A. Rose | This is an alternate cover edition for 9780692... | NaN | Paperback | NaN | NaN | 4.25 | 143 | 19 | Imitatore (The Donna Chronicles #1) | Fantasy\|Young Adult\|Adventure\|Adult | https://images.gr-assets.com/books/1518725503l... |
| 4 | A. Bates | Kelly has always been afraid of final exams. N... | NaN | Paperback | 9.78E+12 | 201 pages | 3.54 | 520 | 29 | Final Exam | Young Adult\|Horror\|Thriller\|Fiction\|Young Adul... | https://images.gr-assets.com/books/1262303980l... |

*Figure 3 Dataset before preprocessing*

```
df_filtered[1:len(df_filtered):2000]
```

| | book_authors | book_pages | book_rating | book_rating_count | book_review_count | book_title | genres |
|---|---|---|---|---|---|---|---|
| 1 | A. Bates | 201 | 3.54 | 520 | 29 | Final Exam | Fiction,Horror,Teen,Thriller,Young Adult |
| 2001 | Ann Rule | 418 | 4.01 | 1691 | 52 | Worth More Dead and Other True Cases | Crime,Mystery,Nonfiction,True Crime,True Story |
| 4001 | Bret Lott | 535 | 3.65 | 34402 | 650 | Jewel | Fiction |
| 6001 | Christina Worrell | 534 | 3.96 | 78 | 21 | Shade's Loss | Demons,Fantasy,Paranormal,Shapeshifters,Vampir... |
| 8001 | Dean Koontz | 386 | 3.81 | 29987 | 1497 | The Good Guy | Fiction,Horror,Mystery,Suspense,Thriller |
| 10001 | Enid Blyton | 183 | 3.99 | 4354 | 96 | The O'Sullivan Twins | Childrens,Fiction |
| 12001 | Graeme Simsion | 432 | 3.57 | 81802 | 9108 | The Rosie Effect | Chick Lit,Contemporary,Fiction,Humor,Romance,W... |
| 14001 | Jack G. Shaheen | 156 | 4.23 | 13 | 1 | The TV Arab | Nonfiction |
| 16001 | Jessica Park | 338 | 4.26 | 8755 | 1173 | Flat-Out Celeste | Contemporary,New Adult,Romance,Young Adult |
| 18001 | Julie Garwood | 379 | 4.26 | 37996 | 1388 | The Secret | Historical,Historical Fiction,Historical Roman... |
| 20001 | Kurt Held | 383 | 4.13 | 1983 | 28 | Die rote Zora und ihre Bande | Childrens,Classics,European Literature,Fiction... |
| 22001 | Louis L'Amour | 240 | 4.02 | 2616 | 72 | Borden Chantry | Fiction,Historical,Historical Fiction,Westerns |
| 24001 | Max Byrd | 448 | 3.50 | 100 | 9 | Jefferson: A Novel | Cultural,Fiction,France,Historical,Historical ... |
| 26001 | Nicholas Sparks | 204 | 3.81 | 70190 | 3390 | At First Sight | Chick Lit,Fiction,Romance,Womens Fiction |
| 28001 | Pittacus Lore | 399 | 4.32 | 23477 | 2132 | The Fate of Ten | Fantasy,Science Fiction,Young Adult |
| 30001 | Roddy Doyle | 307 | 3.75 | 18083 | 755 | Paddy Clarke Ha Ha Ha | 20th Century,Adult Fiction,Coming Of Age,Conte... |
| 32001 | Simon Winchester | 242 | 3.84 | 84160 | 5126 | The Professor and the Madman: A Tale of Murder... | Biography,Books About Books,History,Humanities... |
| 34001 | Terry Pratchett | 155 | 3.74 | 49492 | 1423 | Eric | Fantasy,Fiction,Humor |
| 36001 | William Woodruff | 407 | 3.97 | 605 | 56 | The Road to Nab End | Autobiography,Biography,British Literature,Eur... |

*Figure 4 Dataset after preprocessing*

### 4.2 Network Construction

Building on the preprocessed data, the Network Construction Module creates a bipartite graph representation of the relationships between authors and genres. This graph is saved as "1%_Sample_author_genre_bipartite.gpickle." The construction involves calculating weights based on book ratings, counts, and probabilities, resulting in a comprehensive representation of author-genre associations. The module also includes a visualization component, displaying the bipartite graph for a 1% sample of the dataset. This visualization helps understand the structure of author-genre connections.

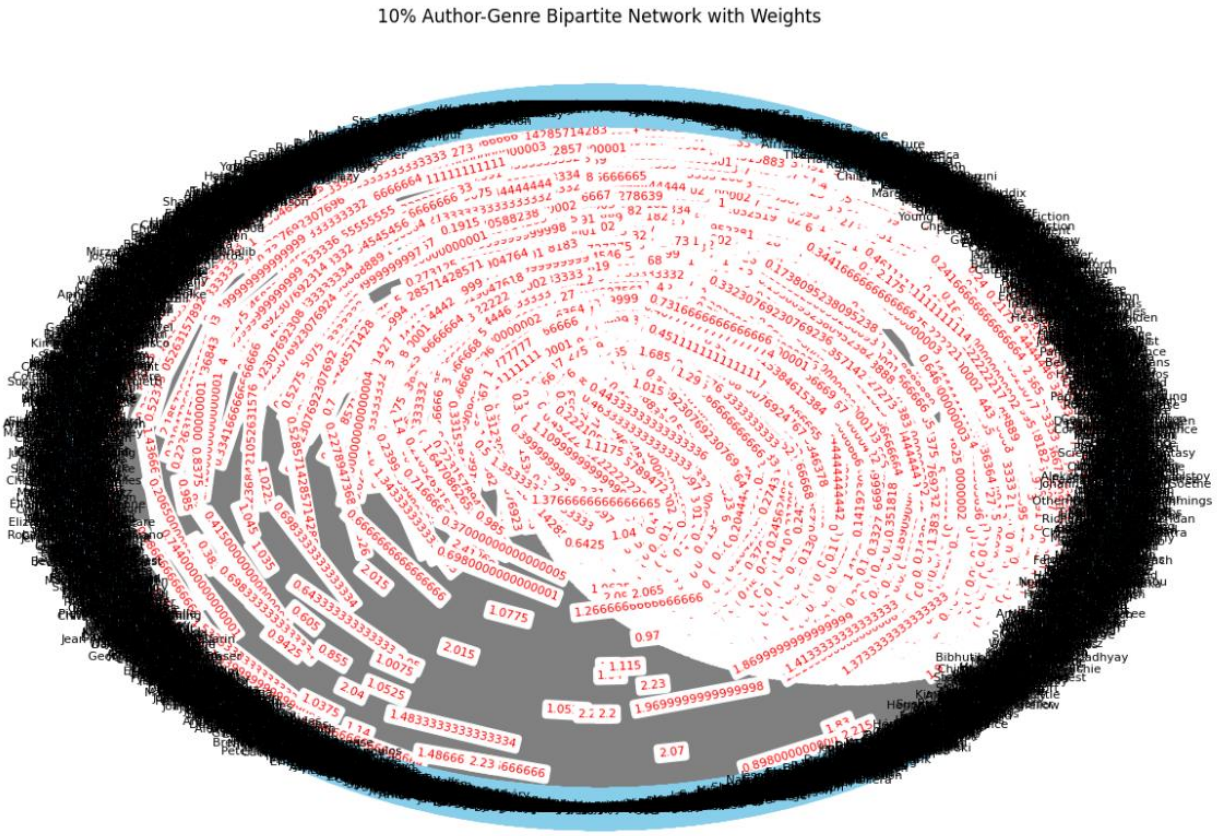



*Figure 5 Author-Genre Bipartite Network created for 10% of the preprocessed dataset*

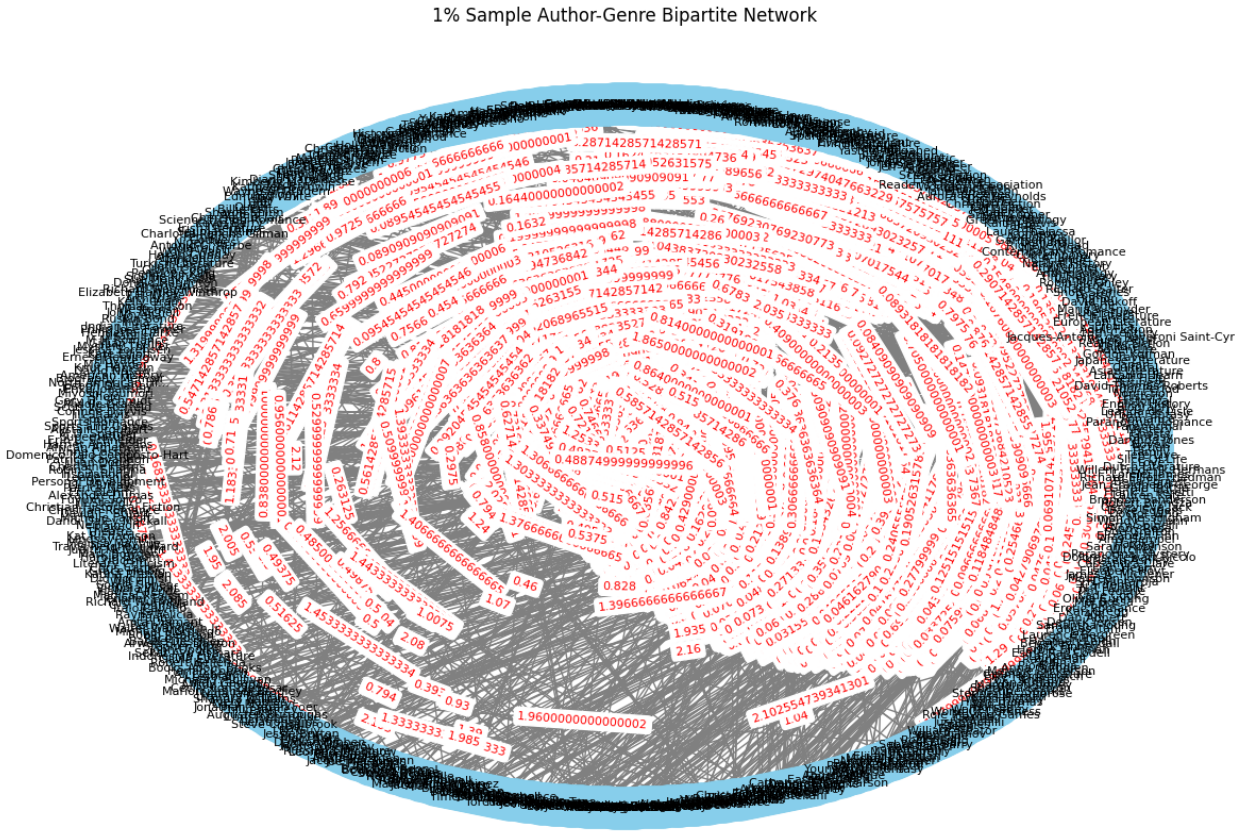



*Figure 6 Author-Genre Bipartite Network created for 1% of the preprocessed dataset*

### 4.3 Author-Genre Association & Community Detection

The Author-Genre Association Module operates on the bipartite graph from the previous module. Its key functionalities include applying the Louvain community detection algorithm, creating a weighted projection of the graph onto authors, and calculating degree centrality for authors. The module identifies genres associated with each author and visualizes the projected graph, showcasing communities of authors with shared genre interests. Further analysis explores and interprets these communities, shedding light on patterns of collaboration and genre preferences.

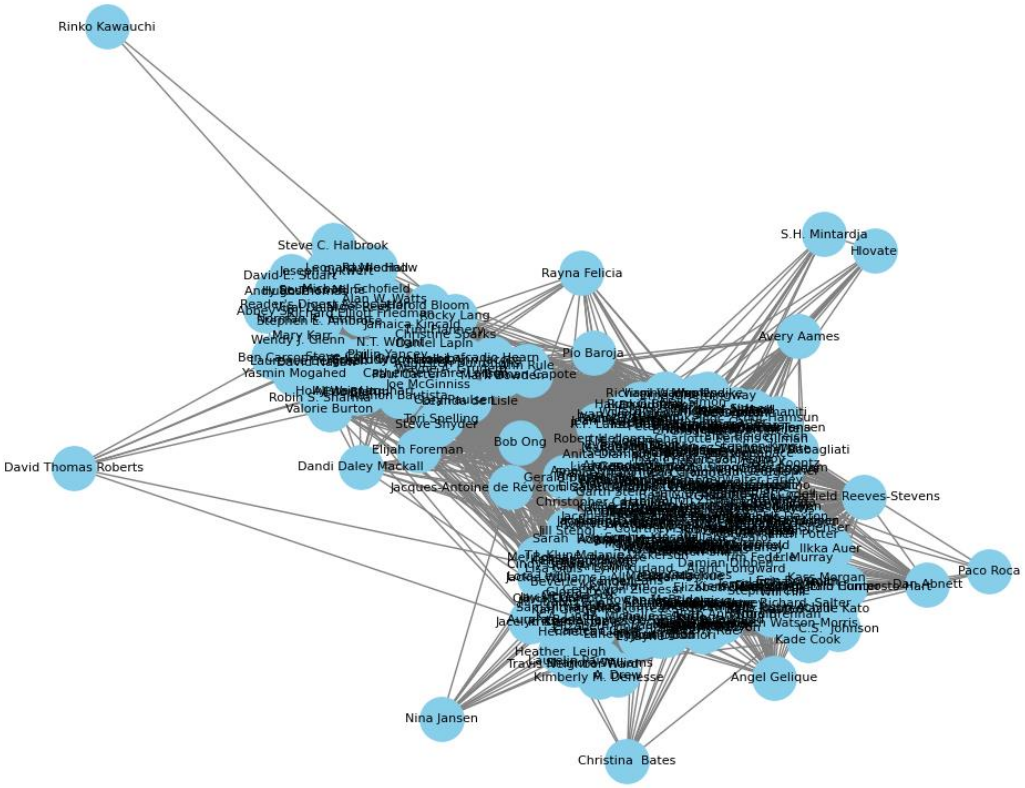



*Figure 7 Author-projected network of author-genre bipartite network for 1% of the dataset*

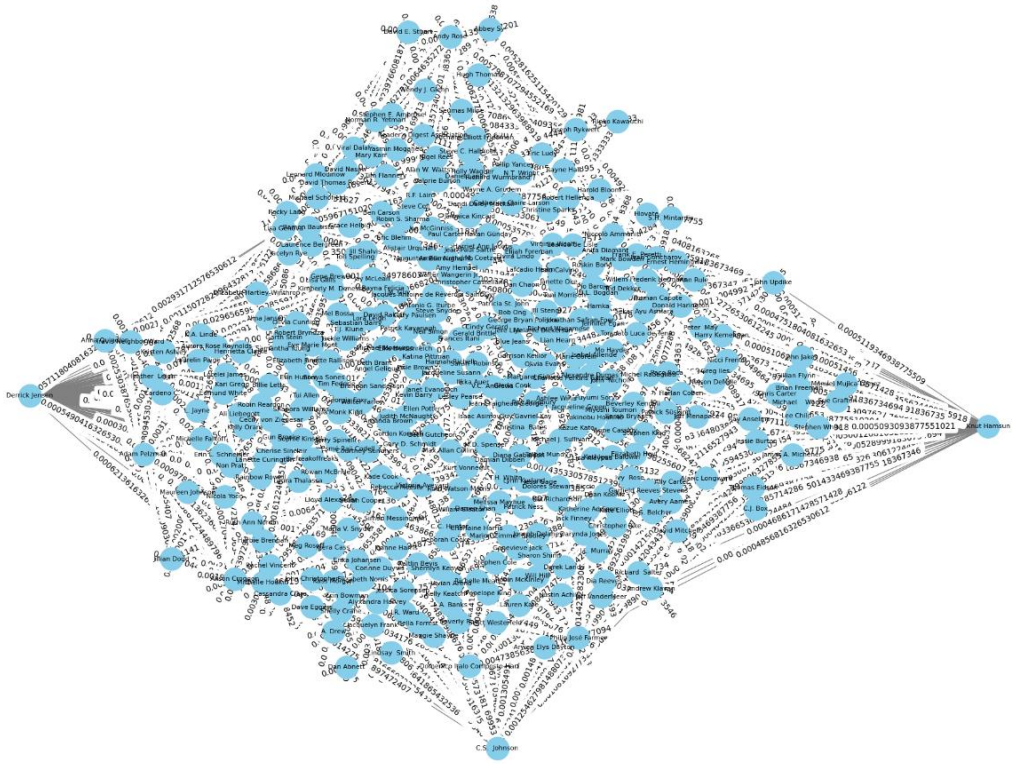


*Figure 8 Community Network from 1% of the dataset*

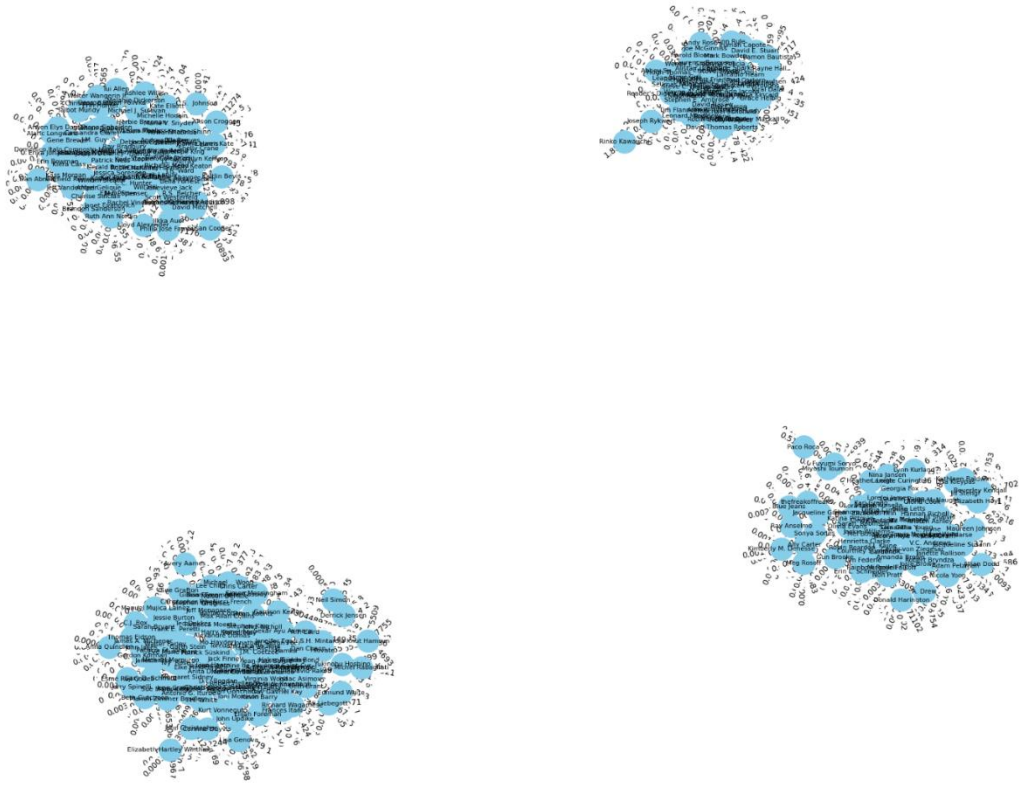


*Figure 9 Community network for 1% of the dataset excluding inter-community links*

### 4.4 Recommendation System

The Recommendation Module utilizes the processed book data, communities data, and user input to provide personalized book recommendations. Users can input a book title, and the module displays details of the chosen book. Leveraging information from the community structure, the module recommends books from authors within the same community, enhancing the personalization of suggestions. Users also have the option to specify the number of recommended books, providing flexibility in the recommendation process.

```
showBooks()
[23]
...
1.      Burlian
2.      Third Grave Dead Ahead
3.      The Sisters Who Would Be Queen: Mary, Katherine, and Lady Jane Grey: A Tudor Tragedy
4.      Unemployable!
5.      Glimpses of Unfamiliar Japan
6.      The Contest
7.      Pauliska, ou la perversité moderne
8.      The Study Series Bundle
9.      Love, Dishonor, Marry, Die, Cherish, Perish
10.     The Patchwork House
11.     Sunshine
12.     The Future Eaters: An Ecological History of the Australasian Lands and People
13.     Avoiding Commitment
14.     Succubus Blues
15.     Guy Noir and the Straight Skinny
16.     Rhapsodic
17.     Daughter of the Earth and Sky
18.     The Grey King
19.     The Complete Photo Guide to Hand Lettering and Calligraphy: The Essential Reference for Novice and Expert Letterers and Calligraphers
20.     Water Walker
21.     Until Jax
22.     Bleeding Violet
23.     Fame, Glory, and Other Things on My To Do List
24.     Great Mysteries of the Past: Experts Unravel Fact and Fallacy Behind the Headlines of History
25.     Taken
...
359.    Who Wrote the Bible?
360.    Bound
361.    Nooit meer slapen
362.    Bidadari Bidadari Surga
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

*Figure 10 List of Books in 1% of the dataset*

```
searchBook()
[14]
✓ 1m 29.4s
...
Book Title: The Future Eaters: An Ecological History of the Australasian Lands and People
Author: Tim Flannery
Number of Pages: 432
Rating: 4.21
Number of Ratings: 580
Genres: ['Anthropology', 'Australia', 'Cultural', 'Environment', 'History', 'Natural History', 'Nonfiction', 'Science']


----- Other Books you may like ------

1.      In Cold Blood by Truman Capote
2.      Magical Thinking: True Stories by Augusten Burroughs
3.      Incidents in the Life of a Slave Girl by Harriet Ann Jacobs
4.      Lit by Mary Karr
5.      The Jesus I Never Knew by Philip Yancey
6.      The Drunkard's Walk: How Randomness Rules Our Lives by Leonard Mlodinow
7.      sTORI Telling by Tori Spelling
8.      Fearless: The Heroic Story of One Navy SEAL's Sacrifice in the Hunt for Osama Bin Laden and the Unwavering Devotion of the Woman Who Loved Him by Eric Blehm
9.      The Way of Zen by Alan W. Watts
10.     Tortured for Christ by Richard Wurmbrand
11.     Killing Pablo: The Hunt for the World's Greatest Outlaw by Mark Bowden
12.     Who Will Cry When You Die?- Hindi by Robin S. Sharma
13.     Ghost Wars: The Secret History of the CIA, Afghanistan, and bin Laden from the Soviet Invasion to September 10, 2001 by Steve Coll
14.     Surprised by Hope: Rethinking Heaven, the Resurrection, and the Mission of the Church by N.T. Wright
15.     Ang Paboritong Libro ni Hudas by Bob Ong
16.     Grace's Guide: The Art of Pretending to Be a Grown-up by Grace Helbig
17.     January First: A Child's Descent into Madness and Her Father's Struggle to Save Her by Michael Schofield
18.     A Small Place by Jamaica Kincaid
19.     Reclaim Your Heart: Personal Insights on Breaking Free from Life's Shackles by Yasmin Mogahed
20.     Winterdance: The Fine Madness of Running the Iditarod by Gary Paulsen
```

*Figure 11 Output of the recommendation system*

### 4.5 General Analysis

The integrated system progresses cohesively from data preprocessing to network construction, author-genre association, and finally, book recommendations. This comprehensive approach ensures that the dataset is cleaned and prepared for analysis, meaningful associations between authors and genres are captured, and users receive personalized book recommendations based on community structures.

The combined effort of these modules results in a robust system for book data analysis and recommendations. The structured workflow ensures that each step contributes to the overall goal of understanding author-genre associations and providing personalized book suggestions. This system can be a valuable tool for users seeking tailored book recommendations based on their preferences and the collaborative patterns of authors within the dataset.

Validation is a critical step to ensure that the system not only aligns with the identified author-genre communities but also resonates with user preferences. User feedback and interactions with the recommendation interface are carefully monitored and analyzed. The system's ability to enhance readership for authors and genres is gauged by assessing the user engagement metrics and the extent to which the recommended books align with users' expectations.

## V. CONCLUSION

The project demonstrates a systematic and comprehensive approach to handling and analyzing book data, emphasizing data preprocessing, network construction, author-genre association, and book recommendations. The Data Preprocessing Module serves as the foundation, ensuring that the dataset is cleaned and transformed into a structured format suitable for subsequent analysis. This step is crucial for obtaining reliable and meaningful insights from the dataset. The Network Construction Module, through the creation of a bipartite graph, effectively captures the intricate relationships between authors and genres. The incorporation of weighted edges based on book ratings, counts, and probabilities enriches the representation, providing a nuanced view of author-genre associations. The visualization of the bipartite graph offers an intuitive understanding of the collaborative patterns and genre preferences within the dataset. The Author-Genre Association Module further delves into the network structure, applying the Louvain community detection algorithm to identify communities of authors with shared genre interests. The calculated degree centrality for authors enhances the analysis, offering insights into the prominence of authors within their respective communities. The visualization of the projected graph provides a clear depiction of author communities, contributing to a deeper understanding of collaborative patterns and genre preferences. The Recommendation Module leverages the information obtained from the previous modules to provide personalized book recommendations. By considering the community structure, the module enhances the relevance of recommendations, offering users a more tailored and enjoyable reading experience. The integration of user input for book selection and recommendation quantity adds a user-friendly dimension to the system.

While the project achieves a robust framework for book data analysis and recommendations, several avenues for future enhancement and expansion are worth exploring. Firstly, the incorporation of additional data sources, such as user reviews and external genre databases, could enrich the analysis and improve the accuracy of recommendations. This could involve implementing sentiment analysis on user reviews to better understand the qualitative aspects of book preferences. The exploration of dynamic network analysis techniques could also be considered for future work. As reading trends and author collaborations evolve over time, analyzing temporal patterns within the dataset could provide a more nuanced understanding of the dynamics of the literary landscape. Furthermore, enhancing the recommendation algorithm by incorporating machine learning models could lead to more sophisticated and accurate predictions. Collaborative filtering or content-based recommendation systems could be explored to capture subtle nuances in user preferences and improve the precision of book suggestions. To enhance user engagement, the development of a user interface could be considered, providing a seamless and interactive experience for users to explore recommendations, visualize author communities, and gain insights into their reading habits.

In summary, the project lays a solid foundation for book data analysis and recommendations, and future works could focus on incorporating advanced techniques, additional data sources, and interactive features to further elevate the user experience and the depth of insights gained from the analysis.

## VI. BIOGRAPHY OF AUTHOR

[1]Aleena is B.Tech Computer Science and Engineering student at Vellore Institute of Technology, Vellore, India, to be graduated in 2024. Her interests include Backend development, data visualization and analysis, and database management, and her research interests are Information Extraction and Retrieval.

[2]Dr. Tamizharasi T is an Assistant Professor Sr. Grade 1 from the Department of Database Systems in the School of Computer Science and Engineering, at Vellore Institute of Technology, Vellore, India.

## REFERENCES


**[1]** Zhao, X., Liang, J., & Wang, J. (2020). A community detection algorithm based on graph compression for large-scale social networks. Information Sciences. doi:10.1016/j.ins.2020.10.057

**[2]** Sarma, D., Mittra, T., & Shahadat, M. (2021). Personalized Book Recommendation System using Machine Learning Algorithm. International Journal of Advanced Computer Science and Applications, 12, 212-219.

**[3]** Saraswat, M., & Srishti. (2022). Leveraging genre classification with RNN for Book recommendation. International Journal of Information Technology, 14(7), 3751-3756.

**[4]** Reddy, S. R. S., Nalluri, S., Kunisetti, S., Ashok, S., & Venkatesh, B. (2019). Content-based movie recommendation system using genre correlation. In Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018, Volume 2 (pp. 391-397). Springer Singapore.

**[5]** Mounika, C., Poojitha, K. V. V. M., Supraja, P. D. L. S., Vidyullatha, P., Priya, P. K., & Gantasala, G. P. (2023, May). Advanced Graph Analytics Algorithms On Genre Based Recommending System. In 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 738-743). IEEE.

**[6]** Maaref, Rihem. (2022). Link prediction in Social Network Analysis: Steps and Algorithms.

**[7]** Cao, H., Holstein, D., & Lee, C. Building a Predictor for Movie Ratings Final Report.

**[8]** Oghina, A., Breuss, M., Tsagkias, M., & De Rijke, M. (2012). Predicting imdb movie ratings using social media. In Advances in Information Retrieval: 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings 34 (pp. 503-507). Springer Berlin Heidelberg.

**[9]** Grujić, J. (2008). Movies recommendation networks as bipartite graphs. In Computational Science–ICCS 2008: 8th International Conference, Kraków, Poland, June 23-25, 2008, Proceedings, Part II 8 (pp. 576-583). Springer Berlin Heidelberg.

**[10]** Scofield, C., Silva, M. O., de Melo-Gomes, L., & Moro, M. M. (2022, March). Book genre classification based on reviews of portuguese-language literature. In International Conference on Computational Processing of the Portuguese Language (pp. 188-197). Cham: Springer International Publishing.