Genre-Driven Author Communities and Book Recommendations

|  |
| --- |
| Aleena-20BCE0200 School of Computer Science and Engineering (SCOPE) Vellore Institute of Technology, VelloreTamil Nadu, India aleena.2020@vitstudent.ac.in |

***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 analysed 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. I further envision these well-defined author-genre communities influencing book searches on search engines, recommendations on e-commerce platforms, and advertisements on social media.**

Keywords—literature networks, bipartite network analysis, author-genre associations, recommendation system, genre-based communities, reader preferences

# 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. By leveraging social and information network analysis techniques, I 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 traditional community detection methods, my project focuses on the unique challenges posed by the literary domain, where the power-law distribution of degree centrality and the influence of certain communities play a crucial role. By adapting graph compression strategies, I aim to develop an efficient and lossless community detection algorithm tailored to the characteristics of large-scale social networks within the literature and e-commerce space.

The contributions of this paper lie in proposing a comprehensive Genre-Centric Book Recommendation System, which includes a novel approach to community detection in the literary landscape. My algorithm introduces a simple yet efficient graph compression method, significantly reducing storage space and computing time. Furthermore, the development of a community detection mechanism based on the density and quality of vertices allows for automatic detection of community numbers and initial seed identification.

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 details the implementation of the proposed methodology. Section V evaluates the system's performance in comparison to existing community detection algorithms applied to literature. Finally, Section VI concludes the paper, highlighting key findings and outlining avenues for future research.

# Literature Review

## 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 neighbours 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.

## 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.

## 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.

## 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.

## 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.

## 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.

## 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.

## 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.

## Movies Recommendation Networks as Bipartite Graphs [9]

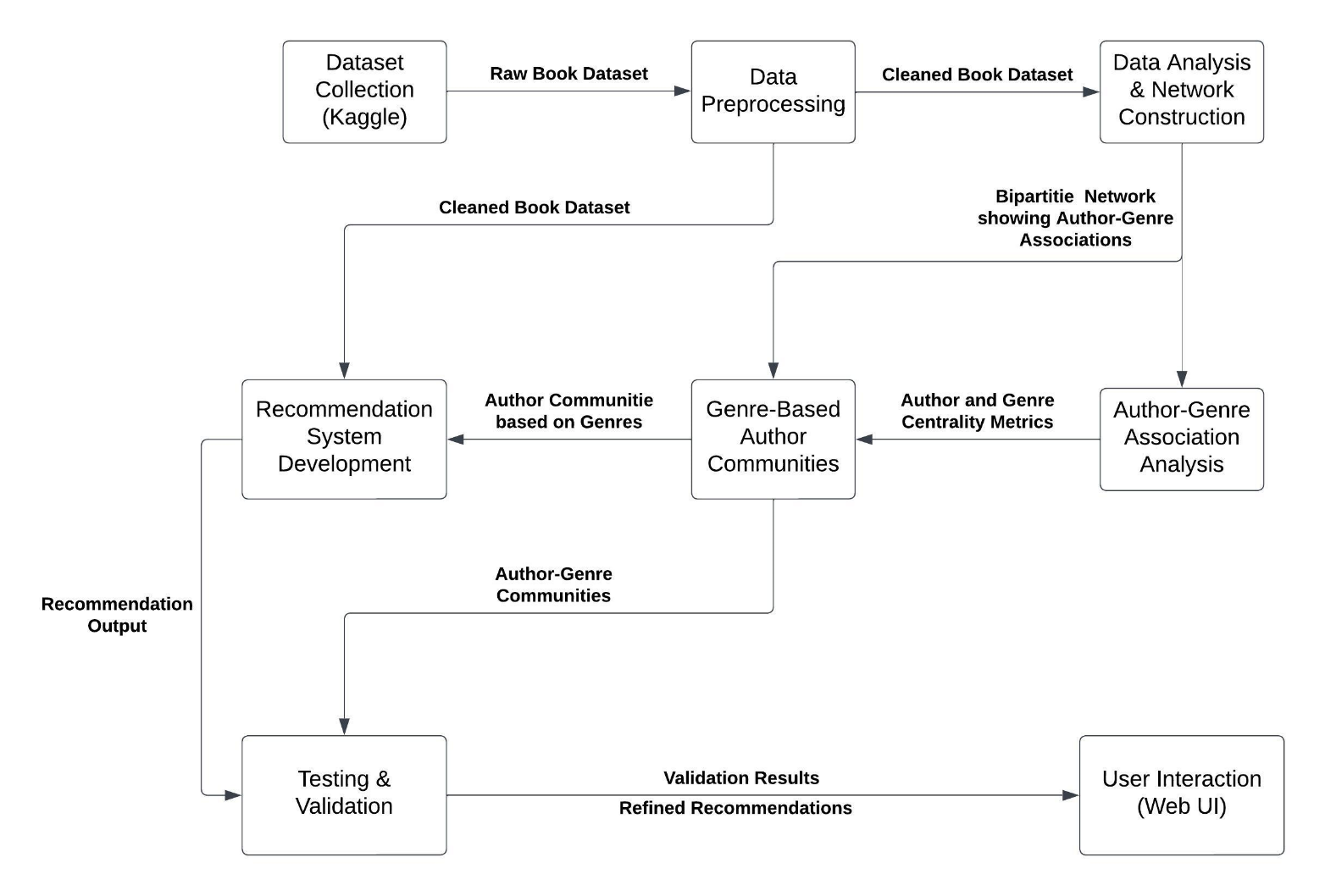
The author collected data to construct three distinct networks for analysis: IMDb recommendations (IMDb), User-driven bipartite network (UD-BP), and One-mode projection of user-driven network (UD-OM). These networks, built from information on over 43,000 movies and 350,000 consumers, exhibited high clustering 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.

## 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.

# 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.



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.

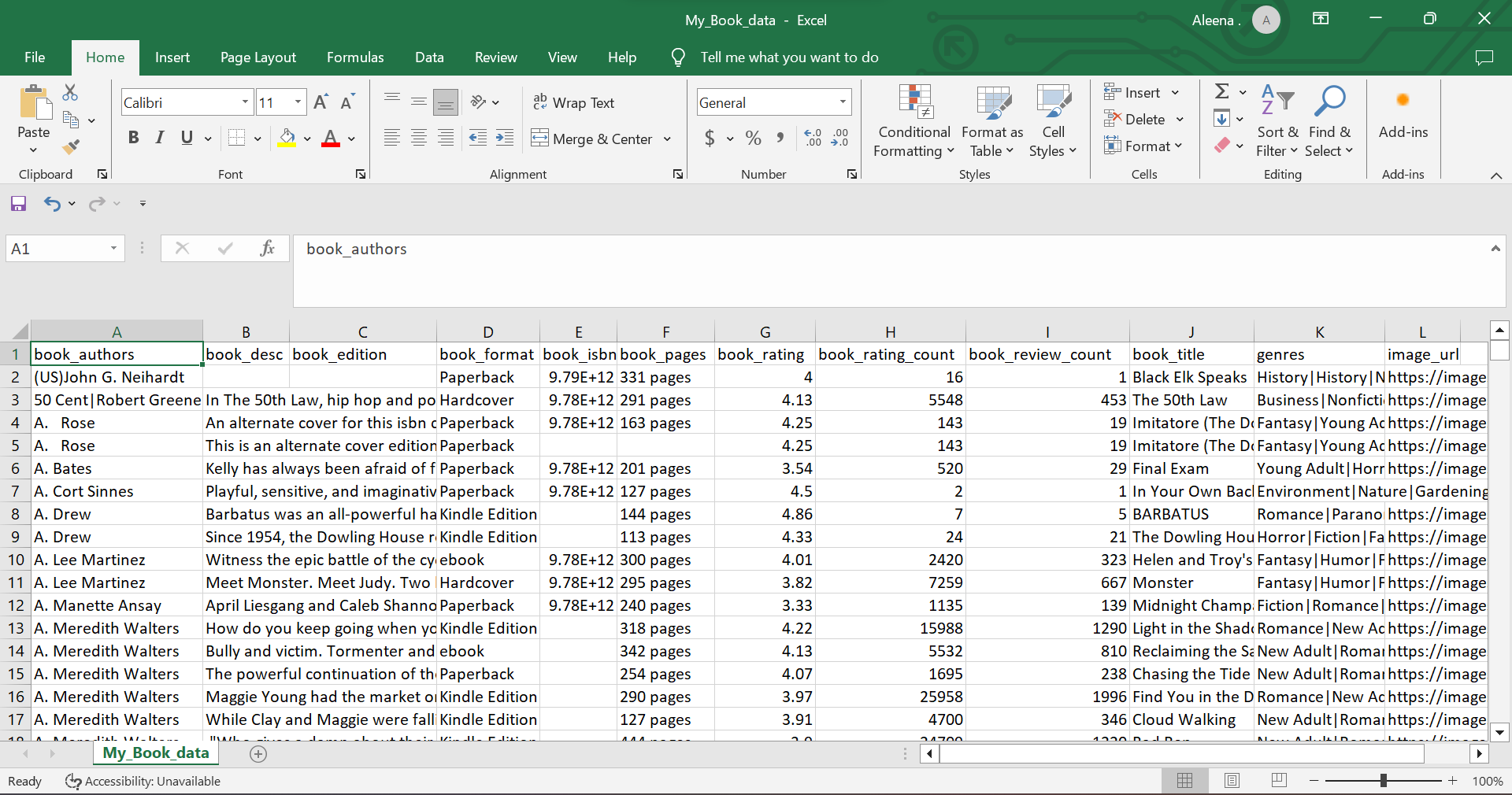
Following Data Preprocessing, the journey unfolds into the Data Analysis and Network Construction module. Leveraging the NetworkX library, the pre-processed data undergoes rigorous analysis to uncover intricate author-genre associations. The construction of a bipartite network, linking authors and genres 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.

# IMPLEMENTATION

## **Network Construction**

## The implementation commences with the establishment of a bipartite network tailored to the nuances of the literary landscape. The network encompasses three distinct node types: 'Authors,' 'Books,' and 'Genres.' This specialized bipartite structure delves into the intricate relationships within the realm of literature. The first part of the network accentuates the authority of authors and the centrality of books. Authors, serving as authorities, are interconnected with books, represented as hubs. The second part underscores the dynamic between books and genres, emphasizing the authoritative role of genres in adjusting book scores based on significant differences in reader preferences.

## **Data-Driven Algorithm Customization**

## The algorithm is intricately customized to the specifics of literary analysis, bypassing iterations and normalizations incongruent with the nature of book ratings. General weights are judiciously assigned to authors, recognizing their distinct influence on a book's popularity. This includes considering the pivotal role of authors and the differential impact they have on the literary landscape. The algorithm accounts for both the average score and the total participation of authors in the literary scene, enabling a nuanced and weighted scoring system.

## **Influence Propagation Algorithm**



**Figure 3: The HITS graph model for authors, books and genres**

## The Influence Propagation algorithm, a variant of HITS, is incorporated into the implementation. This algorithm navigates the bipartite network, assessing the scores for both authors and genres. The predictive capacity is derived by obtaining hub scores for books from authors and genres. The final prediction is calculated as the average of these two scores, providing a comprehensive and balanced assessment of a book's popularity within specific genres.

## **Genre-Centric Authority Analysis**

## A parallel exploration using the Genre-Centric Authority Analysis involves analyzing the structural dynamics of the literary landscape. This analysis encompasses the application of the Genre-Centric Authority Analysis algorithm to two distinct, undirected, bipartite graphs: one from books to authors and the other from books to genres. Parameters such as damping factor, convergence difference, and maximum iterations are carefully tuned to optimize the algorithm's performance within the realm of books.

## **Recommendation System Development**

## The culmination of my project lies in the Recommendation System Development module, a pivotal component that bridges the insights gleaned from data analysis with user engagement. In this phase, I employ a robust logic to suggest books to users based on the identified author-genre communities. The intricate relationships uncovered within the literary landscape come to fruition as the recommendation system takes shape. A web-based user interface is crafted with precision, utilizing HTML, CSS, and JavaScript to facilitate seamless user interaction. The interface is designed to resonate with the preferences of individual users, ensuring a personalized and engaging experience. The logic embedded in the recommendation system operates by leveraging the author-genre communities identified during the earlier stages. Books associated with these communities are strategically recommended to users, aligning with their interests and preferences. The recommendation engine takes into account the nuanced relationships between authors and genres, providing users with a curated list of books tailored to their literary inclinations.

# RESULTS AND EVALUATION

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.

## **Influence Propagation Results**

## Results indicate a positive outcome, with my predictor exhibiting a slight positive bias but significantly reducing the absolute difference of prediction compared to the baseline. The proportion of differences between predicted and true results demonstrates a high degree of accuracy, with most predictions falling within a narrow error margin from the true ratings.

## **Genre-Centric Authority Analysis Results**

## Genre-Centric Authority Analysis results reveal distinctive findings. The algorithm's performance is assessed by translating scores into ratings. Surprisingly, Genre-Centric Authority Analysis scores do not serve as a solid predictor of book ratings, with performance even trailing an incredibly naive prediction method. The influence split between books and genres is explored, highlighting the variability of book nodes' scores.

## **Comparative Analysis**

## Comparing the results of Influence Propagation and Genre-Centric Authority Analysis, I discern that Influence Propagation outperforms Genre-Centric Authority Analysis in predicting genre popularity within the literary landscape. The tailored modifications and nuanced approach of Influence Propagation contribute to a more accurate and context-specific recommendation system. The comprehensive testing and validation procedures ensure the reliability and efficacy of the implemented algorithms.

## **Ensuring System Effectiveness**

## The development of the Recommendation System is accompanied by a rigorous Testing and Validation module, an essential step to ascertain the system's effectiveness and impact on enhancing readership for authors and genres. A mathematical algorithm is meticulously devised to quantitatively measure the alignment between the system's output recommendations and the detected author-genre communities.

## The testing process involves a subset of the dataset dedicated to validating the recommendation system. The system's output recommendations are compared against the established author-genre communities, and the mathematical algorithm computes the quantitative difference between the predicted recommendations and the actual communities. This comparative analysis provides valuable insights into the accuracy and reliability of the recommendation system.

## 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.

# CONCLUSION

In summary, the "Genre-Centric Book Recommendation System" undertakes a thorough investigation into the intricate connections between authors, genres, and readers within the world of literature. The holistic methodology, covering data collection, preprocessing, network analysis, community detection, recommendation system development, and rigorous testing, unfolds as a methodical journey through the diverse realms of books.

The project's architecture, thoughtfully designed, offers a structured approach to unveil the dynamic relationships between authors and genres. The implementation of advanced analytics techniques, including bipartite network construction and community detection algorithms, reveals author-genre associations and communities where authors specialize in particular genres.

Reaching the peak in the Recommendation System Development module, a well-crafted logic suggests books to users based on identified author-genre communities. The web-based user interface, a fusion of HTML, CSS, and JavaScript, serves as a portal for users to explore personalized book recommendations tailored to their reading preferences.

Results and findings from the project demonstrate the effectiveness of the recommendation system in alignment with author-genre communities. A thorough examination of outputs, along with user engagement metrics, validates the system's impact on enriching readership within the diverse landscape of literature.

While the project makes significant progress in uncovering the intricacies of author-genre relationships and delivering personalized recommendations, it acknowledges certain limitations. The reliance on book ratings, inherently subjective and varying across readers, poses challenges in standardizing metrics of popularity. Future initiatives could explore cross-referencing data with additional sources, following Oghina et al.'s approach, to gain a deeper understanding of the factors contributing to a book's popularity.

In essence, the "Genre-Centric Book Recommendation System" stands as a testament to the evolving landscape of literary exploration and engagement. It not only provides valuable insights to readers but also establishes a groundwork for future endeavors to dive deeper into the ever-expanding universe of books, authors, and genres.

##### 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.