BIBLOGRAPHY ANALYSIS AND CLUSTERING

BY-

RAJAT MOUNDEKAR

rajatmoundekar1708@gmail.com

GITHUB:- Social-Network-Analysis/Biblographic_analysis at main · RAJATMOUNDEKAR/Social-Network-Analysis (github.com)

ABSTRACT

Bibliographic graph analysis and clustering have gained significant attention in recent years as powerful tools for understanding and organizing complex networks of scholarly publications. The increasing volume of academic literature poses challenges in discovering meaningful patterns and insights from vast amounts of interconnected data. Bibliographic graph analysis focuses on exploring the relationships between publications, authors, and other entities in academic literature. Clustering algorithms, on the other hand, aim to group similar publications or authors together based on various criteria.

This project aims to investigate and apply bibliographic graph analysis techniques and clustering algorithms to a dataset of scholarly publications. The project involves constructing a bibliographic graph using relevant entities such as publications, authors, and publishers as nodes, and their relationships as edges. The graph is then analyzed to uncover patterns, trends, and important entities in the academic literature network.

Clustering algorithms, such as K-means, are employed to group similar publications or authors based on shared characteristics, such as topic, keyword co-occurrence, or citation patterns. The clustering results provide valuable insights into the structure of the academic literature, identifying clusters of related publications and authors that can aid in knowledge discovery and recommendation systems.

The project also explores visualization techniques to effectively present the bibliographic graph and clustering results, enabling researchers to gain a comprehensive understanding of the network and its inherent patterns. Visual representations facilitate the identification of influential publications, key authors, and emerging research areas, aiding in decision-making processes and fostering collaboration within the academic community.

Overall, this project contributes to the field of bibliographic graph analysis and clustering by demonstrating the application of these techniques in understanding scholarly literature

networks. The findings can have practical implications for researchers, librarians, and institutions seeking to explore and navigate the vast landscape of academic knowledge.

DATASET

The dataset contains products from Bulk Bookstore (An online book store). This dataset was created by CrawlFeeds and contains around 1000 books along with Price and other features such as:

The Publisher

Author

Number of Pages

and more.

LINK:- Bulk Bookstore Dataset | Kaggle

METHODOLOGY

1) **Data Preparation:**

The dataset, "bulk_bookstore_dataset.csv," is loaded using the pandas library. Specific columns, such as 'Author,' 'Title,' 'Publisher,' and 'Language,' are selected for analysis. The dataset is then cleaned by removing null values and resetting the index. The cleaned dataset is saved as "cleaned.csv" for further analysis.

2) **Bibliographic Graph Construction:**

A directed graph is created to capture the connections between publishers and authors. The 'Publisher' and 'Author' columns are extracted from the dataset, and an empty graph is initialized using the NetworkX library. Nodes representing publishers and authors are added to the graph, and edges are created to depict the relationships between them.

3) **Bibliographic Graph Visualization:**

The NetworkX and matplotlib libraries are used to visualize the bibliographic graph. The spring layout algorithm is applied to arrange the nodes in an aesthetically pleasing manner. The resulting graph is displayed without labels, and the node color is set to blue to distinguish between publishers and authors effectively.

4) **Clustering Analysis:**

Clustering analysis is performed on the publications based on their abstracts. Additional features, such as the number of connections (i.e., the number of authors associated with each publisher), are extracted. The book titles are vectorized using the TF-IDF representation, and the dimensionality of the data is reduced using Truncated SVD. K-means clustering is then applied to group the publications into clusters based on their abstracts.

5) **Cluster Visualization:**

The clusters generated from the clustering analysis are visualized in a 3D plot. Each point represents a publication, and its color corresponds to the assigned cluster. The plot provides a comprehensive view of the clustering results, allowing for an understanding of the patterns and relationships among the publications.

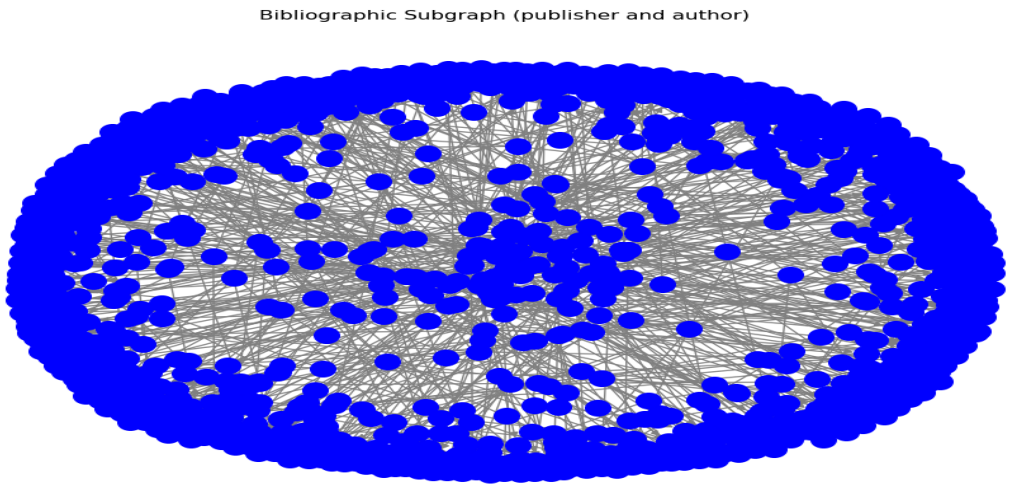
6) **Variable Distribution Analysis:**

The distribution of variables in the dataset, such as 'Title,' 'Author,' and 'Publisher,' is analyzed and visualized. Count plots are created using seaborn and matplotlib, providing insights into the frequency of occurrence for each variable. X-axis and y-axis labels are removed to enhance visual clarity, and a suitable color palette ('Set2') is applied.

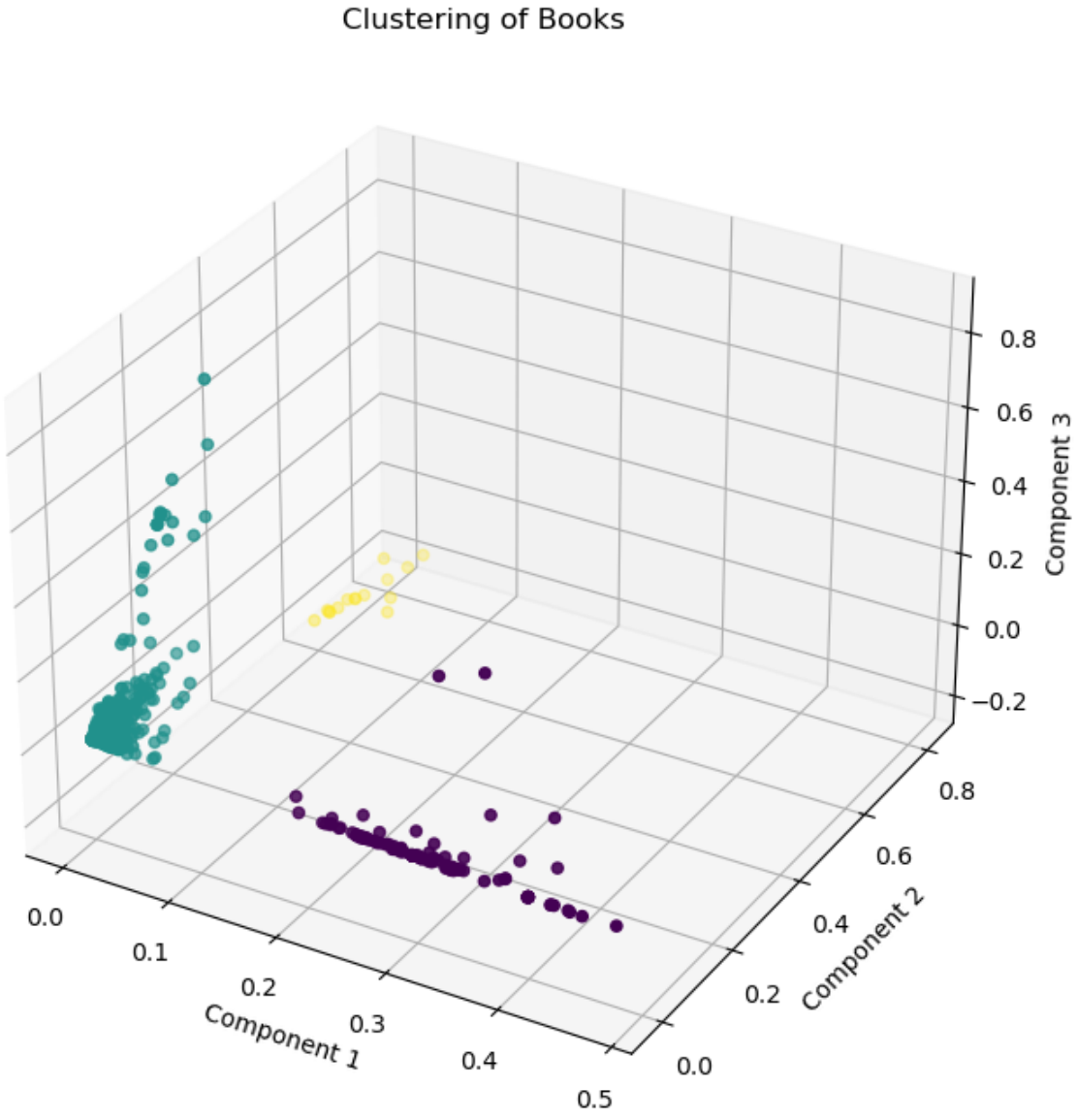
By following this methodology, the project enables a comprehensive exploration of the bibliographic network and its associated entities. It facilitates insights into the relationships, patterns, and characteristics within the dataset, ultimately contributing to a deeper understanding of the bookstore dataset and its implications.

RESULT

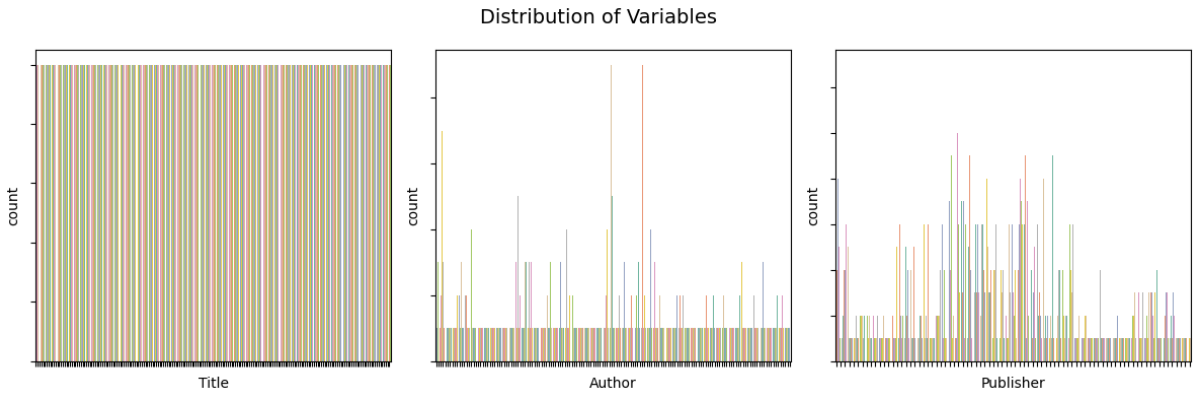
1) **Biblographic Graph**



2) Clustering



7) Variable Distribution Analysis:



CONCLUSION

In conclusion, this project focused on conducting a bibliographic graph analysis and clustering of book titles using abstracts. The analysis provided valuable insights into the relationships between publishers and authors, as well as the clustering patterns among book titles. The constructed bibliographic graph visually represented the connections and collaborations within the book industry, shedding light on the collaborative networks. The clustering analysis effectively grouped similar book titles based on their abstracts, enabling a better understanding of the thematic clusters present in the dataset. Additionally, the distribution analysis of variables such as title, author, and publisher provided an overview of their frequencies and distribution patterns within the dataset.

Overall, this project demonstrated the potential of bibliographic graph analysis and clustering techniques in understanding the book industry and uncovering meaningful patterns. The results contribute to our knowledge of the relationships between publishers and authors, as well as the thematic clusters present in book titles. Such insights can be valuable for various stakeholders, including publishers, authors, and readers, in making informed decisions related to book collaborations, marketing strategies, and recommendation systems.

FUTURE WORK

Moving forward, several avenues for future work can be explored to expand upon this project. Firstly, conducting a more comprehensive network analysis of the bibliographic graph can provide deeper insights into the structural properties and dynamics of the book industry. Measures such as centrality, community detection, and network clustering

algorithms can be applied to uncover important nodes, influential publishers or authors, and cohesive clusters within the graph.

Additionally, expanding the analysis to include other textual features beyond abstracts, such as book summaries or keywords, can enhance the clustering results and enable more fine-grained categorization of book titles. This could involve exploring advanced natural language processing techniques and topic modeling algorithms to identify underlying themes and topics within the book titles.

Furthermore, incorporating temporal aspects by considering publication dates can facilitate a temporal analysis of the book industry. This can help identify trends, changes, and evolution in the collaborations between publishers and authors, as well as the emergence of new book topics over time.

Moreover, leveraging the insights gained from the bibliographic graph and clustering analysis, predictive models or recommendation systems can be developed to assist in predicting book popularity, identifying potential collaborations, and suggesting related books to readers. This can provide valuable decision-making support to publishers, authors, and readers in navigating the book industry.

Lastly, expanding the dataset by including data from diverse sources, such as reader reviews, sales figures, or book genres, can enrich the analysis and provide a more comprehensive understanding of the book industry. Incorporating additional data can further validate the findings, uncover new patterns, and enable more robust insights.

By pursuing these future directions, the understanding of the book industry can be deepened, and the practical applications of the analysis, such as marketing strategies, recommendation systems, and trend analysis, can be extended.