Linh Bui
DS210 - Final Project Write Up

Goodreads Book Recommendation

Project Overview + Objective

As a frequent reader, I often gravitated toward books within the same genre or those recommended by friends with similar interests. While these recommendations are valuable, they can sometimes limit exposure to a broader range of books. To address this, I envisioned a way to discover books that might interest me, transcending genre boundaries based on attributes such as average rating, review count (popularity), number of pages (preferred book length), publisher (indicative of style), and more.

This project aims to analyze the relationships between books by constructing a graph where:

- Nodes represent individual books.
- Edges represent similar relationships between books, determined by shared or similar attributes.

The project involves building and analyzing a graph representation of a book dataset, focusing on relationships driven by attributes like shared authors, similar average ratings, number of pages, publishers, ratings count, and text review count. This approach enables the identification of patterns, clusters, and significant connections within the dataset, offering a fresh perspective on book discovery.
The graph provides a tool to uncover patterns, such as highly connected books, clusters of related books, and the overall structure of connections across the dataset. By leveraging these insights, readers can explore books that resonate with their preferences without being constrained by specific genres.

Dataset Overview

Dataset: https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks
The dataset contains data for 11,123 books taken from Goodreads, including the following attributes:

- Book ID: A unique identifier for each book.
- Title: The book's title, including series information if applicable.
- Authors: Names of authors who contributed to the book.
- Average Rating: A floating-point score representing the average reader rating for the book.
- ISBN and ISBN13: Standard book identifiers (10-digit and 13-digit formats).
- Language Code: A string representing the primary language of the book (e.g., eng, spa).
- Number of Pages: The total number of pages in the book (may be missing for some records).
- Ratings Count: The total number of ratings the book has received.
- Text Reviews Count: The total number of textual reviews for the book.
- Publication Date: The date the book was published, formatted as MM/DD/YYYY.
- Publisher: The name of the publishing entity.

Graph Construction

A directed graph was constructed using the petgraph crate, representing each book as a node. Edges were created between nodes based on weighted similarity criteria, including shared authors, similar ratings, similar page counts, shared publishers, and similarities in review statistics. For example, books with average ratings differing by less than 0.5 were connected with a weight of 1.0, while books with the same publisher received an additional weight of 0.2. Only meaningful connections where the total edge weight exceeded 0, were included to prevent graph overload.

Linh Bui
DS210 - Final Project Write Up

The project implemented several analysis functions to extract meaningful insights from the graph. The first function identified the top five books with the highest degree (most connections), revealing highly connected nodes like "The Iliad" and "Anna Karenina." These books acted as hubs in the graph, suggesting their centrality in the dataset. The degree distribution function analyzed the spread of connection counts across all nodes, highlighting the structural diversity of the graph. Most books had relatively low degrees, but a few nodes displayed extremely high degrees, forming major hubs. Lastly, the project identified the two most similar books based on edge weights, often uncovering pairs with shared attributes like authorship or publication.

Code Sections

The project is divided into three main modules: main.rs, dataprep.rs, and graph.rs, each serving a distinct purpose in the overall functionality. The main.rs module acts as the entry point of the program, orchestrating the workflow by integrating the data preparation and graph analysis modules. It begins by invoking the parse_csv function from dataprep.rs to read and preprocess the dataset. If successful, it builds a graph using the build_graph function from graph.rs, where books are represented as nodes, and edges represent relationships based on weighted similarity criteria. The program then performs analyses, such as identifying highly connected books, calculating degree distributions, and finding the most similar pairs of books. The results, including statistics about the graph and insights into book relationships, are displayed in the console.

The dataprep.rs module is responsible for parsing and cleaning the dataset. A Book struct is defined to represent the attributes of each book, including its title, authors, average rating, and other metadata. To handle missing or inconsistent data, the struct includes optional fields and default values, preventing runtime errors. The parse_csv function uses the serde crate to deserialize the CSV file into a vector of Book structs. Invalid rows are skipped, with detailed error messages logged for transparency. By providing a clean and consistent dataset, this module lays the foundation for effective graph construction and analysis.

The graph.rs module handles the construction and analysis of the graph, which is implemented using the petgraph crate. The build_graph function creates nodes for each book and establishes edges based on weighted similarity criteria, such as shared authors, similar ratings, similar page counts, shared publishers, and similarities in review statistics. These weights ensure that connections are meaningful and relevant. Additionally, the module includes functions for analyzing the graph. The find_highly_connected_nodes function identifies the top five books with the most connections, while the analyze_degree_distribution function calculates the distribution of degrees across all nodes, highlighting the most connected books. Finally, the find_most_similar_neighbors function identifies pairs of books with the strongest similarity relationships based on edge weights.

Output

Highly Connected Books: Books with many connections often share common attributes, such as popular authors or genres.

Linh Bui
DS210 - Final Project Write Up

Degree Distribution: Reveals how connected the books are and highlights any hubs or outliers in the dataset.
Most Similar Neighbors: Identifies pairs of books with strong relationships, for recommendation systems or clustering.

```
Parsed 11123 books
Edges added: 90865164
Graph built with 11123 nodes and 90865164 edges.
Top 5 highly connected books:
Title: The Iliad, Connections: 84024
Title: Anna Karenina, Connections: 74585
Title: 'Salem's Lot, Connections: 73847
Title: The Odyssey, Connections: 71203
Title: A Midsummer Night's Dream, Connections: 66330
Top 10 Degree Distribution:
Degree: 84024, Count: 1, Percentage: 0.01%
Degree: 74585, Count: 1, Percentage: 0.01%
Degree: 73847, Count: 1, Percentage: 0.01%
Degree: 71203, Count: 1, Percentage: 0.01%
Degree: 66330, Count: 1, Percentage: 0.01%
Degree: 64405, Count: 1, Percentage: 0.01%
Degree: 63608, Count: 1, Percentage: 0.01%
Degree: 56827, Count: 1, Percentage: 0.01%
Degree: 56818, Count: 1, Percentage: 0.01%
Degree: 56690, Count: 1, Percentage: 0.01%
Books with degree greater than 50000:
Title: The Histories, Degree: 56818
Title: Anna Karenina, Degree: 74585
Title: Macbeth, Degree: 56690
Title: Treasure Island, Degree: 55716
Title: The Picture of Dorian Gray, Degree: 64405
Title: The Great Gatsby, Degree: 56827
Title: Jane Eyre, Degree: 54343
Title: Gulliver's Travels, Degree: 53219
Title: The Iliad, Degree: 84024
Title: A Midsummer Night's Dream, Degree: 66330
Title: Romeo and Juliet, Degree: 52367
Title: 'Salem's Lot, Degree: 73847
Title: The Brothers Karamazov, Degree: 63608
Title: The Odyssey, Degree: 71203
Title: The Secret Garden, Degree: 51830
Title: Sense and Sensibility, Degree: 55147
Most similar books are 'Harry Potter and the Half-Blood Prince (Harry Potter  #6)' and 'Harry Potter and the Order of the Phoenix (Harry Potter  #5)' with similarity score 1.30
```

Linh Bui
DS210 - Final Project Write Up