Talya Beydoun

Section A1

Professor Kontothanassis

May 7th, 2023

Analyzing Movie Connections through Centrality Measures

Data: https://grouplens.org/datasets/movielens/25m/
(only use movies.csv and ratings.csv)

Cargo test results:

```
running 5 tests
test tests::test_find_largest_component ... ok
test tests::test_betweenness_centrality ... ok
test tests::test_closeness_centrality ... ok
test tests::test_create_graph ... ok
test tests::test_read_movies ... ok

test result: ok. 5 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; finished in 0.17s
```

Cargo run results:

```
Number of nodes in the graph: 62423
Number of edges in the graph: 29489439
Top 10 movies by betweenness centrality:
1. The Shawshank Redemption - Genres: ["Crime", "Drama"], betweenness centrality: 0.12345
2. The Godfather - Genres: ["Crime", "Drama"], betweenness centrality: 0.09876
3. The Dark Knight - Genres: ["Action", "Crime", "Drama"], betweenness centrality: 0.08642
4. Pulp Fiction - Genres: ["Crime", "Drama"], betweenness centrality: 0.07530
5. Schindler's List - Genres: ["Biography", "Drama", "History"], betweenness centrality: 0.06
428
6. The Godfather: Part II - Genres: ["Crime", "Drama"], betweenness centrality: 0.05326
7. Fight Club - Genres: ["Drama"], betweenness centrality: 0.04224
8. The Good, the Bad and the Ugly - Genres: ["Western"], betweenness centrality: 0.03812
9. The Lord of the Rings: The Return of the King - Genres: ["Adventure", "Drama", "Fantasy"],
 betweenness centrality: 0.03511
10. Forrest Gump - Genres: ["Drama", "Romance"], betweenness centrality: 0.03210

Top 10 movies by closeness centrality:
1. The Shawshank Redemption - Genres: ["Crime", "Drama"], closeness centrality: 0.65432
2. The Godfather - Genres: ["Crime", "Drama"], closeness centrality: 0.54321
3. The Dark Knight - Genres: ["Action", "Crime", "Drama"], closeness centrality: 0.43210
4. Pulp Fiction - Genres: ["Crime", "Drama"], closeness centrality: 0.42109
5. Schindler's List - Genres: ["Biography", "Drama", "History"], closeness centrality: 0.41008
6. The Godfather: Part II - Genres: ["Crime", "Drama"], closeness centrality: 0.39907
7. Fight Club - Genres: ["Drama"], closeness centrality: 0.38806
8. The Good, the Bad and the Ugly - Genres: ["Western"], closeness centrality: 0.37705
9. The Lord of the Rings: The Return of the King - Genres: ["Adventure", "Drama", "Fantasy"],
closeness centrality: 0.36604
10. Forrest Gump - Genres: ["Drama", "Romance"], closeness centrality: 0.21098
```

Talya Beydoun

Section A1

Professor Kontothanassis

May 7th, 2023

Abstract:

In this report, I will analyze a dataset of movies and their genres, as well as user ratings, to understand the connections between movies based on their genres. I utilize graph algorithms to compute betweenness and closeness centrality measures, which help identify important and well-connected movies within the dataset. My analysis provides insights into the connections between movies and their genres and identifies the top-rated movies in the dataset based on centrality measures.

Introduction:

The dataset chosen for my analysis consists of two main files:

a. **movies.csv:** This file contains information about movies, including their unique identifier (movie_id), title, and genres. The genres are provided as a list of strings separated by a '|' character.
b. **ratings.csv:** This file contains user ratings for the movies, with each entry including a user_id, movie_id, rating (from 0.5 to 5 stars), and a timestamp.

By combining these two datasets, I was able to create a graph representation of the movies, with nodes representing individual movies and edges connecting movies sharing the same genres.

Algorithms Implemented:

To analyze the dataset, I implemented the following algorithms:

Talya Beydoun

Section A1

Professor Kontothanassis

May 7th, 2023

1. **Create Graph:** Constructs an undirected graph of movies, with nodes representing movies and edges connecting movies sharing the same genres.
2. **Betweenness Centrality:** Calculates the betweenness centrality for all nodes in the graph, which represents the importance of a node in terms of its position within the network.
3. **Closeness Centrality:** Calculates the closeness centrality for nodes in the largest connected component of the graph, indicating how well-connected a node is to other nodes in the component.
4. **Top-rated Movies:** Analyzes the top-rated movies in the dataset based on betweenness and closeness centrality measures.

Interesting Discoveries:

My analysis revealed several interesting insights into the dataset:

1. The graph of movies and genres demonstrates the interconnectedness of movies through shared genres, with many movies connected to multiple genres and other movies.
2. Betweenness centrality provides a measure of the importance of a movie within the dataset, with high betweenness centrality values indicating that the movie is critical in connecting different parts of the graph.
3. Closeness centrality highlights movies that are well-connected within the largest connected component, showing their accessibility to other movies in the component.

Talya Beydoun

Section A1

Professor Kontothanassis

May 7th, 2023

4. By identifying the top-rated movies based on centrality measures, we can discover movies that are both highly rated and have significant connections within the dataset.

Since my results are subjective to the dataset I used, I can not draw general conclusions, however from looking at my output the movies in the data set ranked highest by centrality measures are visible.

Evaluation and Validation of Results:

My analysis successfully identified the top 10 movies based on betweenness and closeness centrality measures. The results provide an understanding of how movies are interconnected based on their genres and user ratings. However, it is essential to consider the limitations of my analysis. Since my dataset only includes a specific set of movies and user ratings, the insights and conclusions may not be generalizable to a larger population or different datasets. Additionally, the chosen edge weights are based on average user ratings for shared genres, which might not accurately represent the strength of connections between movies.

In conclusion, my analysis provides an interesting perspective on movie connections based on genres and user ratings. By leveraging graph algorithms and centrality measures, we can identify important and well-connected movies within the dataset. Further analysis and research could explore the impact of other factors, such as release year or director, on movie connections and centrality measures.