DS210 Movie Graph Project Report

Date: 6/05/2023

Introduction and Overview:

In this project, I analyze the FilmTV movies dataset from Kaggle, specifically the ENG version containing 40,303 movies with 19 attributes. The goal is to investigate actor relationships in the movie industry by calculating the shortest path between two randomly chosen actors based on their collaborations in past movies.

Data Cleaning:

Using pandas in Python, I cleaned the dataset by removing unnecessary attributes and rows with NaN values in the actors and title columns. I also excluded movies with only one actor and retained only movies from the United States. The final cleaned dataset consisted of 15,518 rows and 2 attributes: title and actors. In Rust, I applied .trim() to eliminate any spaces.

Six Degrees of Separation:

The concept of Six Degrees of Separation suggests that any two individuals worldwide can be connected through no more than six intermediate links. In this project, I apply this principle to actor relationships within the movie industry.

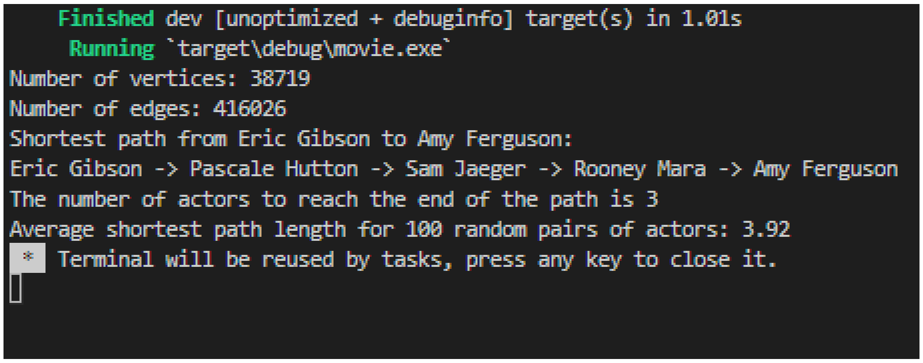
Conclusion and Findings:

Movie Graph Project:

I employed the Breadth-First Search (BFS) algorithm to determine the shortest path between actors, treating actors as vertices and movie collaborations as edges in an undirected graph. The graph implementation is modular for easy adaptation and expansion.

The program outputs the following information:

1. The number of vertices (actors) in the graph is 38,719, and the number of edges (movies) is 416,026.
2. An example of the shortest path between two actors (Eric Gibson and Amy Ferguson) includes three intermediate actors: Pascale Hutton, Sam Jaeger, and Rooney Mara.
3. The average shortest path length for 100 random pairs of actors is 3.92, indicating that two random actors are connected through approximately 4 other actors in the movie graph.
4. The output suggests that the movie industry has a relatively small world network where actors are closely connected through their collaborations.



```
    Finished dev [unoptimized + debuginfo] target(s) in 1.01s
     Running `target\debug\movie.exe`
Number of vertices: 38719
Number of edges: 416026
Shortest path from Eric Gibson to Amy Ferguson:
Eric Gibson -> Pascale Hutton -> Sam Jaeger -> Rooney Mara -> Amy Ferguson
The number of actors to reach the end of the path is 3
Average shortest path length for 100 random pairs of actors: 3.92
 *  Terminal will be reused by tasks, press any key to close it.
```

// Output should change each time it is run.

Test Cases:

A set of test cases in the tests module address MovieGraph creation, BFS algorithm functionality, and edge addition functionality. To run the tests, navigate to the project directory and execute the command cargo test