DS210 Movie Graph Project Report

Date: 6/05/2023

Introduction and Overview

I analyze a dataset from Kaggle called FilmTV movies dataset, which includes two files: filmtv_movies - ENG.csv and filmtv_movies - ITA.csv. We chose the ENG version, containing 40,303 movies with 19 attributes.

Data Cleaning

To prepare the dataset for analysis, I utilized pandas in Python to remove unnecessary attributes. I dropped all NaN rows in the actors and title (movie titles) columns and excluded movies with only one actor to increase the chances of finding connections between actors in the graph. After assessing the impact of these changes on the dataset, I identified movies from various countries. To reduce the number of paths not found by the BFS algorithm, I retained only movies from the United States and kept the title and actors attributes. The cleaned data consisted of 15,518 rows and 2 attributes. In Rust, I applied .trim() to eliminate any spaces.

Six Degrees of Separation:

The Six Degrees of Separation concept suggests that any two individuals worldwide can be connected through no more than six intermediate links. In this project, this principle applies to actor relationships within the movie industry. The shortest path between two actors is calculated based on their collaborations in movies.

Conclusion and Findings

Movie Graph Project:

I investigated actor relationships in the movie industry by calculating the shortest path between two randomly chosen actors in the movie dataset. We 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. We designed the graph implementation to be modular for easy adaptation and expansion.

The program outputs:

1. The number of vertices (actors) and edges (movies) in the graph.
2. The randomly chosen starting and ending actors.
3. The shortest path between the two actors, including intermediate actors.
4. The number of actors needed to reach the end actor from the starting actor (excluding the start and end actors themselves).
5. If no connection exists between the two actors, the program outputs "No path found between [start actor] and [end actor]."

Test Cases:

We have incorporated a set of test cases in the tests module, which 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.