Primah Muwanga

Data Science 210

Professor Kontothanassis

15 December 2023

Final Project Report

The dataset I have selected is the Reddit Hyperlink Network, which is located in the social network section of the Stanford dataset page. According to the website, this network illustrates the direct connections between two subreddits, representing communities on Reddit. The data spans two years and was extracted from publicly available Reddit data.

In this dataset, each subreddit is a node, and the edges signify hyperlinks between subreddits. Specifically, the file I am working with is a network of subreddit-to-subreddit hyperlinks extracted from the body of posts. This means the hyperlinks extracted were created within the posts that link from one subreddit to another. Each hyperlink originates from a post in the source community and links to a post in the target community. I opted for this dataset because of my interest in Reddit and its ability to connect people. The BU Reddit page, in particular, captures various content, ranging from the funniest and most pointless to the relatable and valuable. Despite the presence of computer science students airing grievances, the content is humorous due to the absurdities at our school, which can all be found on Reddit. Interestingly, I stumbled upon my current roommate on Reddit, highlighting the platform's effectiveness in forging connections. So far, things are working out quite well.

To achieve my goal at the end of this project and have everything working, I started by making a checklist of the first things I needed to do. Over the past 13 weeks, I constantly got ahead of myself during the homework. Initially, I made my read file work before I moved on to creating

my adjacency list. To make the adjacency list using a HashMap, I first needed to understand what these terms meant, which helped a lot. When my adjacency list was achieved, I made the breadth-first-search algorithm logic. This sequential approach has proved instrumental in achieving a deeper understanding of each component and ensuring a well-structured and functional outcome.

Throughout this project, I tried solving the average degree of separation using the breadth-first search algorithm. I found the average to be between 3.5 and 4.9. This number represents the average shortest path length between a specific starting node and all other nodes in the graph. In this subreddit data set, the average degree of separation provides insights into how closely connected or separated different network parts are from a given starting point. With this data set, there is no specific shortest path; I actually learned that lesson when I tried to find the shortest path from a source node to a target node, and every time, the output was just the starting node twice, then the target node. At first, I was confused, and I tried fixing the loops and trying everything, but after multiple attempts, I figured there was no shortest path unless I did DFS, which, given more time with the project, I could have implemented as well. I chose to make tests that tested the adjacency and the read file because these are the first mechanisms needed to move on and create the breadth-first search algorithm that then calculates the max and average degree. My project works by reading a file and then creating an adjacency list that will be used to find the maximum degree of separation and average degrees of separation.

PROBLEMS 2 OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
Node: drawhelix
Average Degree of Separation: 3.96
Maximum degree of separation for drawhelix: 8
---------------------
Node: falstadt
Average Degree of Separation: 4.00
Maximum degree of separation for falstadt: 8
---------------------
Node: slate
Average Degree of Separation: 0.67
Maximum degree of separation for slate: 1
---------------------
Node: animemusic
Average Degree of Separation: 3.75
Maximum degree of separation for animemusic: 7
---------------------
Node: aidb
Average Degree of Separation: 3.87
Maximum degree of separation for aidb: 7
---------------------
Node: avalanche2_support
Average Degree of Separation: 0.50
Maximum degree of separation for avalanche2_support: 1
---------------------
Node: quebeclevis
Average Degree of Separation: 4.08
Maximum degree of separation for quebeclevis: 8
---------------------
Node: imaginarywitches
Average Degree of Separation: 4.11
Maximum degree of separation for imaginarywitches: 8
---------------------
Node: beachcity
Average Degree of Separation: 3.24
Maximum degree of separation for beachcity: 7
---------------------
Node: evilgeniuses
Average Degree of Separation: 3.70
Maximum degree of separation for evilgeniuses: 7
---------------------
```