Ania Shaheed
DS 210
7 May 2023

Final Project Report

Project Overview:

Goal: The goal of this project is to analyze a Facebook social network dataset in Rust and use a measure of similarity to identify commonalities and dissimilarities between pairs of users as well as their respective circles.

Motivation: The motivation for this project is to explore the concept of friend recommendations on social media platforms, using the Facebook dataset as an example. My idea is to develop an algorithm that can analyze the social connections between users.

The Data:

Name: ego-Facebook

Link: https://snap.stanford.edu/data/ego-Facebook.html

Description: This Facebook dataset from the Stanford Network Analysis Project (SNAP) contains anonymized data from Facebook's user profiles and friendship networks, as well as information about their networks of online friendships. Specifically, the dataset includes information about a set of Facebook users who have agreed to share their profile information and friendship networks with researchers. The data includes information about users' gender, age, education, and relationship status, as well as a list of their friends on Facebook. There are 10 individuals in the dataset, each with a unique ID.

Files: The dataset contains several files for each ego node. However, I only decided to use 2 out of the 5 for this project: nodeId.edges and nodeId.circles. The edges data is undirected, and while the 'ego' node does not appear, it is assumed that they follow every node id that appears in this file. The circles data contains a list of circles for the node as well as a series of nodes for each circle.

Methodology:

Approach: First, I “read” the Facebook friends list dataset from files into memory. Then, I parsed the dataset into a graph data structure. Each node in the graph represents a Facebook user, and each edge represents a friendship between two users. After, I created a function to test friends of friends using an adjacency list. I was then able to find which lists were most similar or dissimilar. This process was repeated for the circles. Throughout, I wrote tests to ensure correctness.

Metrics: To measure similarity, I decided to use Jaccard similarity. It is defined as the ratio of the size of the intersection of the sets to the size of the union of the sets. In other words, it measures the extent to which two sets overlap. I used this metric because it was the most straightforward and intuitive.

Execution and Output:

Execution: This code was built to analyze the Facebook friends list dataset and output the results. Therefore, this code is compatible with datasets that are in the same format. Additionally, the input is written into the code, so running the program will directly produce the results of the data used without needing any other user input.

Output:

```
$ cargo run --release
   Compiling project v0.1.0 (C:\Users\anias\Documents\BU\Semester 2\DS 210\Project\project)
    Finished release [optimized] target(s) in 1.29s
     Running `C:\Users\anias\Documents\BU\Semester 2\DS
210\Project\project\target\release\project.exe`
-----Data Overview-----

Individual 0
Number of edges: 5038
Number of circles: 24

Individual 107
Number of edges: 53498
Number of circles: 9

Individual 348
Number of edges: 6384
Number of circles: 14

Individual 414
Number of edges: 3386
Number of circles: 7

Individual 686
Number of edges: 3312
Number of circles: 14

Individual 698
Number of edges: 540
Number of circles: 13

Individual 1684
Number of edges: 28048
Number of circles: 17

Individual 1912
Number of edges: 60050
Number of circles: 46

Individual 3437
Number of edges: 9626
Number of circles: 32

Individual 3980
Number of edges: 292
Number of circles: 17

-----Comparing Circles within Indiviudals' Networks-----

Individual 0:
No Overlapping Circles:

Individual 107:
No Overlapping Circles:

Individual 348:
Most similar circles:
("circle1", "circle8", 1.0)
Lease similar circles:
("circle1", "circle11", 0.0)

Individual 414:
No Overlapping Circles:

Individual 686:
Most similar circles:
("circle12", "circle13", 1.0)
Lease similar circles:
("circle7", "circle9", 0.0)

Individual 698:
Most similar circles:
("circle2", "circle0", 1.0)
Lease similar circles:
("circle8", "circle2", 0.0)
```

```
Individual 1684:
No Overlapping Circles:

Individual 1912:
Most similar circles:
("circle22", "circle19", 1.0)
Lease similar circles:
("circle22", "circle25", 0.0)

Individual 3437:
Most similar circles:
("circle8", "circle12", 1.0)
Lease similar circles:
("circle8", "circle19", 0.0)

Individual 3980:
No Overlapping Circles:

-----Friends of Friends Similarities-----

How often are friends of my friends my friends?
[(0, 0.2212624584717608), (107, 0.3863976083707025), (348, 0.16496350364963502), (414,
0.10964912280701754), (686, 0.8038277511961722), (698, 0.08388814913448735), (1684,
0.43068493150684933), (1912, 0.7507537688442211), (3437, 0.7739130434782608), (3980,
0.9122807017543859)]

Most similar users (user 1, user2, similarity):
(205, 52, 1.0)

Most dissimilar users (user 1, user2, similarity):
(2202, 61, 0.0)
```

Conclusions:

The Jaccard similarity for individual 686 is very high (0.80), which means that a lot of their friends are also their friends' friends. The Jaccard similarities for IDs 1912 and 3437 are also quite high (0.75 and 0.77 respectively), the similarities for IDs 107 and 1684 are moderate (0.39 and 0.43 respectively), and the similarities with IDs 0, 348, 414, 698 and 3980 are low to very low. This means that these individuals do not share many friends in common with their friends' friends. The "most similar" and "most dissimilar" analysis did not give many insights, since many friends completely overlapping or had nothing in common at all.

Sources:

https://en.wikipedia.org/wiki/Jaccard_index