Kiki Park

Professor Kontothanassis

CDS DS 210

15 December 2013

Final Project Reflection

What the Project is Supposed to Do

This project aims to analyze a dataset on citation networks from Kaggle, observing academic research trends to identify the most influential papers and authors. The dataset was chosen for its extensive coverage that gave me plenty of nodes to work with and could potentially serve as a small exploration of the dynamics of academic influence and topic evolution.

The original goals/intentions of the project were as follows:

1. To identify influential papers and authors. By calculating centrality measures such as degree, betweenness, and closeness centrality, the project highlights the most influential papers and authors in the network. These metrics provide a quantifiable means of understanding influence and prominence within the academic community.
2. To identify research clusters. Through clustering algorithms, the project identifies various clusters within the citation network. Each cluster represents a coherent group of research papers, often aligned with specific research themes or areas. This analysis is crucial for understanding how academic research is grouped and how different fields or topics are interrelated.

3. To visualize the citation network. The visualization component of the project was supposed to offer a graphical representation of the network, or an overview of the network structure that highlighted key papers, authors, clusters, etc.

How to Run It

To run the project, you should only need to run the function `fn main`defined in the `main.rs` file, or use `cargo run` in a Visual Studio Code terminal after opening the DS210_FinalProject folder.

The project's code is split into six modules which ostensibly accomplish the following:

1. `main.rs`: The entry point of the application. It orchestrates the flow of data through various stages of processing and analysis.
2. `data_preprocessing.rs`: Responsible for reading and preprocessing the dataset. It constructs a directed graph representing the citation network.
3. `centrality_analysis.rs`: This module computes various centrality measures to identify the most influential papers and authors in the network.
4. `clustering.rs`: Used for detecting clusters or communities within the citation network, indicative of different research areas.
5. `visualization.rs`: Handles the graphical representation of the network, aiding in the visual interpretation of the data.
6. `utilities.rs`: A support module providing common functionalities used across the project.

Surprise! The Project Doesn't Do What It's Supposed to Do

My project was unsuccessful. I continuously encountered challenges throughout the development, some of which I was clearly unable to resolve before the deadline. The dataset I downloaded was only about 5GB large, but after unzipping it I discovered that it was in fact about 12GB large! I didn't feel confident that I could write yet another proposal and have it approved and still have time to complete the project, so I decided to proceed and deal with the problems as they came, but I quickly became overwhelmed by trying to handle the large dataset and to maximize efficiency while still preserving accurate parsing and processing of the original data.

Parsing and processing the dataset was incredibly difficult. The initial attempts to construct the graph encountered issue after issue, with programs taking 20 minutes or longer to complete running or Visual Studio Code crashing. The graph constructed from the original dataset had a whopping 4,894,063 nodes and 45,564,149 edges. In trying to make the program run faster, I kept getting outputs with 0 nodes and 0 edges, then managing to get it up to a very, very impressive graph with 3 nodes and 2 edges, and finally giving up once the program constructed a graph that had 4,146,772 nodes and 45,564,149 edges but didn't progress any further (as in I left `fn main` running for over 12 hours and the output was still not finished). I think this can be attributed to incorrect JSON parsing, but I struggled to revise the parsing logic to match the dataset's structure.

The problems with graph construction and data preprocessing unsurprisingly affected the graph visualization. I constantly got blank results. Literally blank. As in there was a new PNG file entitled graph_visualization but it was a naked white square with very helpful lettering that titled the image "Graph Visualization."

In the end, the hours of work feel wasted and this project amounted to what is essentially a failed experiment. I’m not sure why I expected anything different, considering how helpless I am when it comes to coding, especially with Rust. Ha, ha. I suppose I’m just really glad to be guaranteed a minimum score of 40% and not a flat zero.