Sarthak Aggarwal ASU ID#1222329095

CSE 472 Project 1

Data Crawler

Project Report

US Elections- Red vs Blue

Overview:
I have completed a project on the US elections. I have analyzed posts related to this topic as well as individuals connected with this debate.

Data Collections:
I have used Mastodon as the primary social media site. I created a developer account to fetch API keys that I used to crawl the data.

Posts: I have gathered many hashtags relating to the US elections. Almost half of them relate to the democrats and other half to the republicans. Then my crawler uses the .search_v2() function to fetch all the posts on Mastodon that used any of these hashtags. Then these posts are saved in all_statuses list. This is then written into the posts.json file. This contains dictionaries with post details along with any comments or boosts on them.

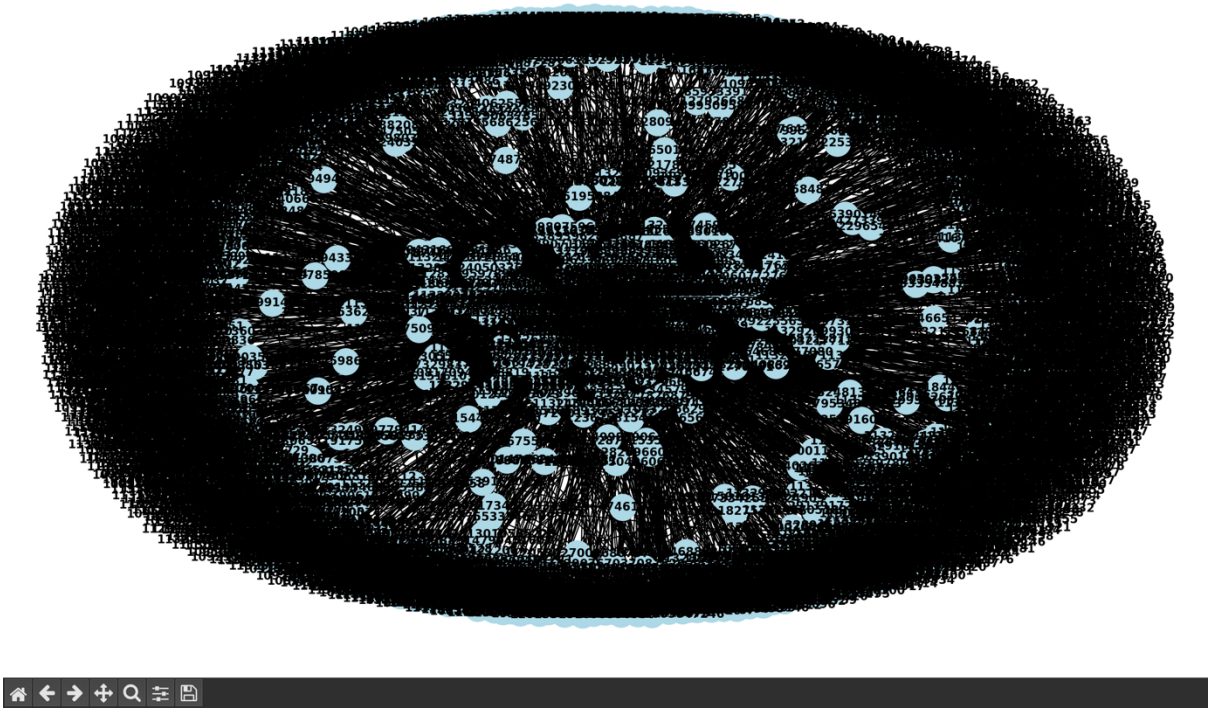
Users: I have gathered 3 users from Mastodon, Trump and Kamala Harris. Kamala Harris's account did not have much followers so I took two of her accounts that equaled the following of Trump's handle. Now I ran a loop to get top 80 followers (max) from all of them. Then I recursed this to get followers of one of those followers to get a data set of 350 users.

Graph Creation

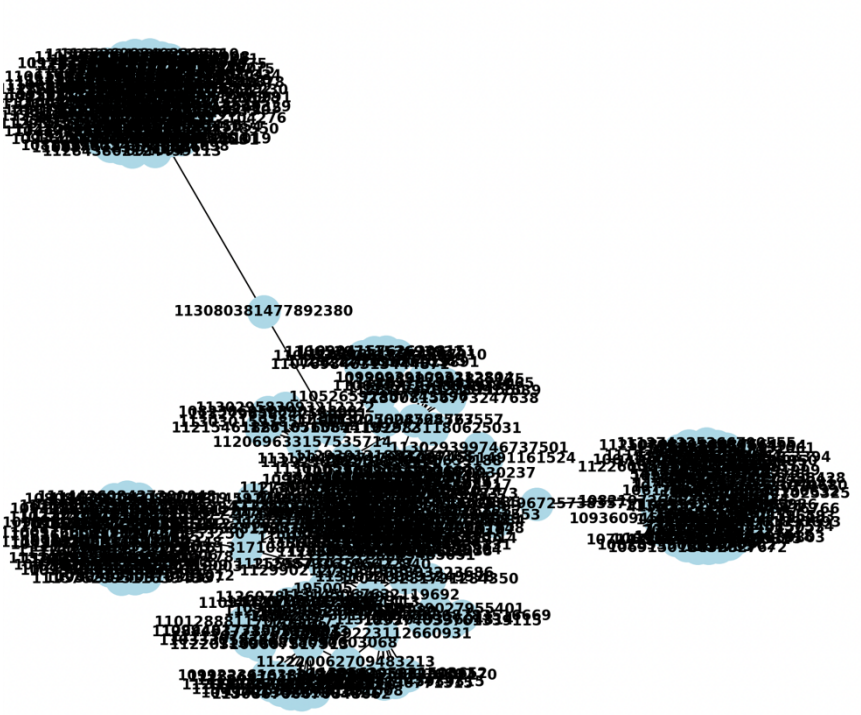
I have used network libraries for creating two graphs.

Post Graph: this Directional graph contains posts, comments and boosts as nodes. These are then connected. One post is connected to its replies and boosts. The nodes are labeled with their post id as an identifier.

Sarthak Aggarwal ASU ID#1222329095



User Graph: This graph contains three seed users. These users are then used to fetch more users and they are connected to each other. Then top users are again used to fetch more users and are connected. The nodes are labeled as their id as identifyer.



Sarthak Aggarwal ASU ID#1222329095

I made a map dictionary to store key as the id and values as its followers. Then I added all the elements as a node to the graph and recurse through the dict to log all the connections.

Classification of Posts using Llama8

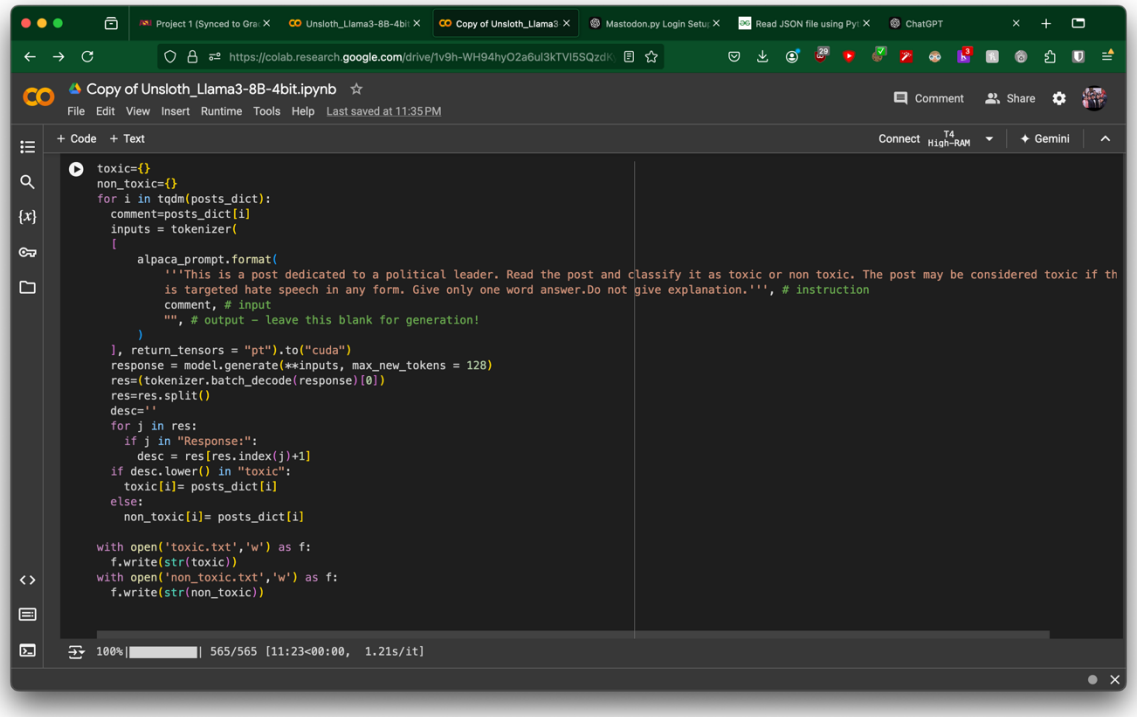
I used Llamaba 3.1 8b. I made a dictionary from my posts.json file and fed it to the LLM. The instruction feeded is :"

his is a post dedicated to a political leader. Read the post and classify it as toxic or non toxic. The post may be considered toxic if there
is targeted hate speech in any form. Give only one word answer.Do not give explanation.

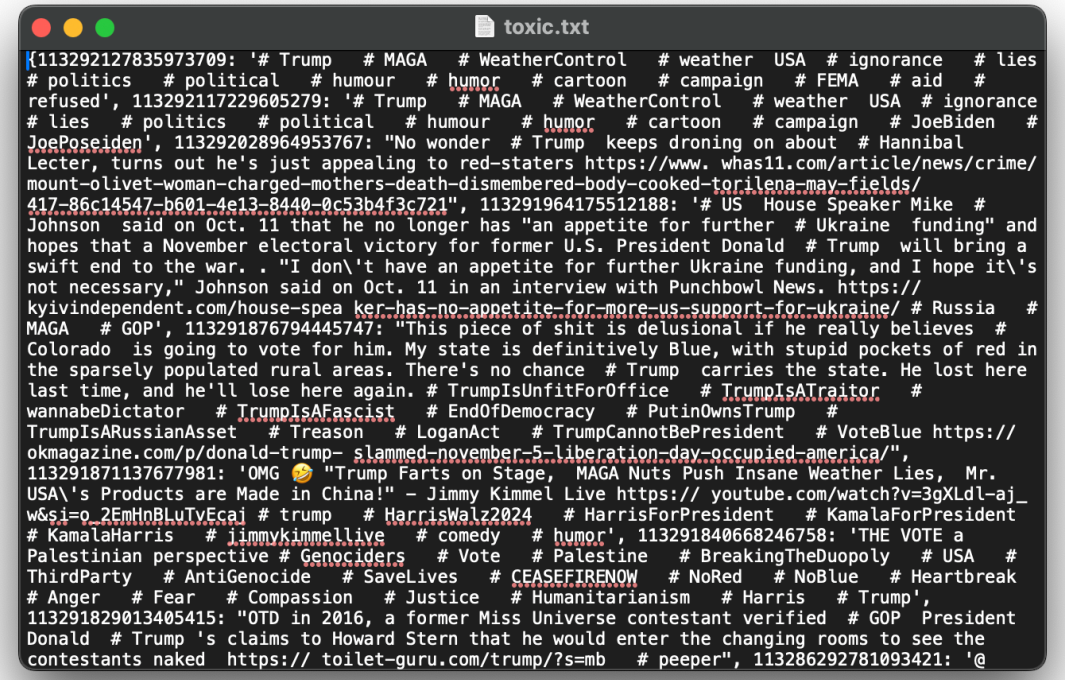
"

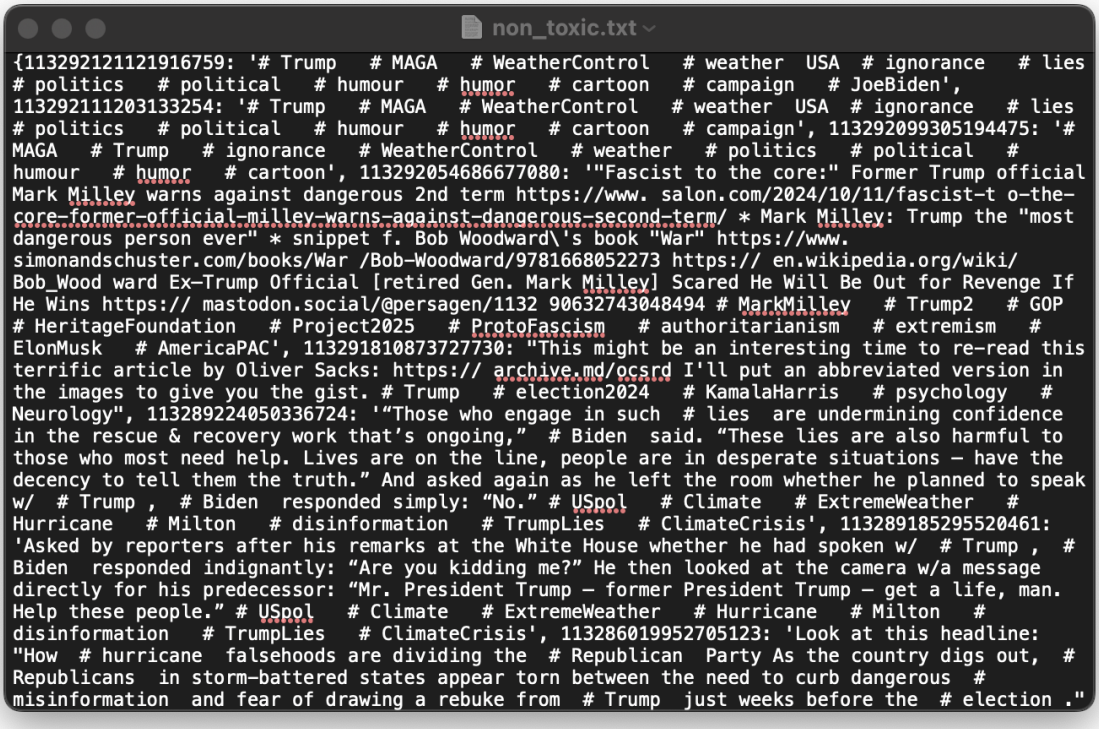
Then I parse the output to check if it returns toxic or non-toxic. Depending on that, I put it in toxic.txt or non_toxic.txt

Link https://colab.research.google.com/drive/1v9h-WH94hyO2a6ul3kTVI5SQzdKyXQiU?usp=sharing

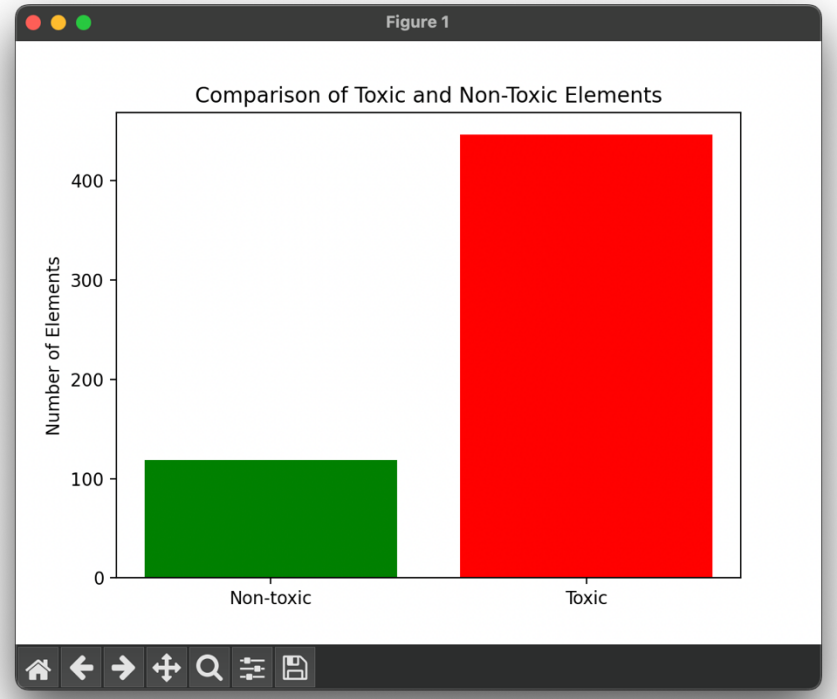


Sarthak Aggarwal ASU ID#1222329095





Sarthak Aggarwal ASU ID#1222329095



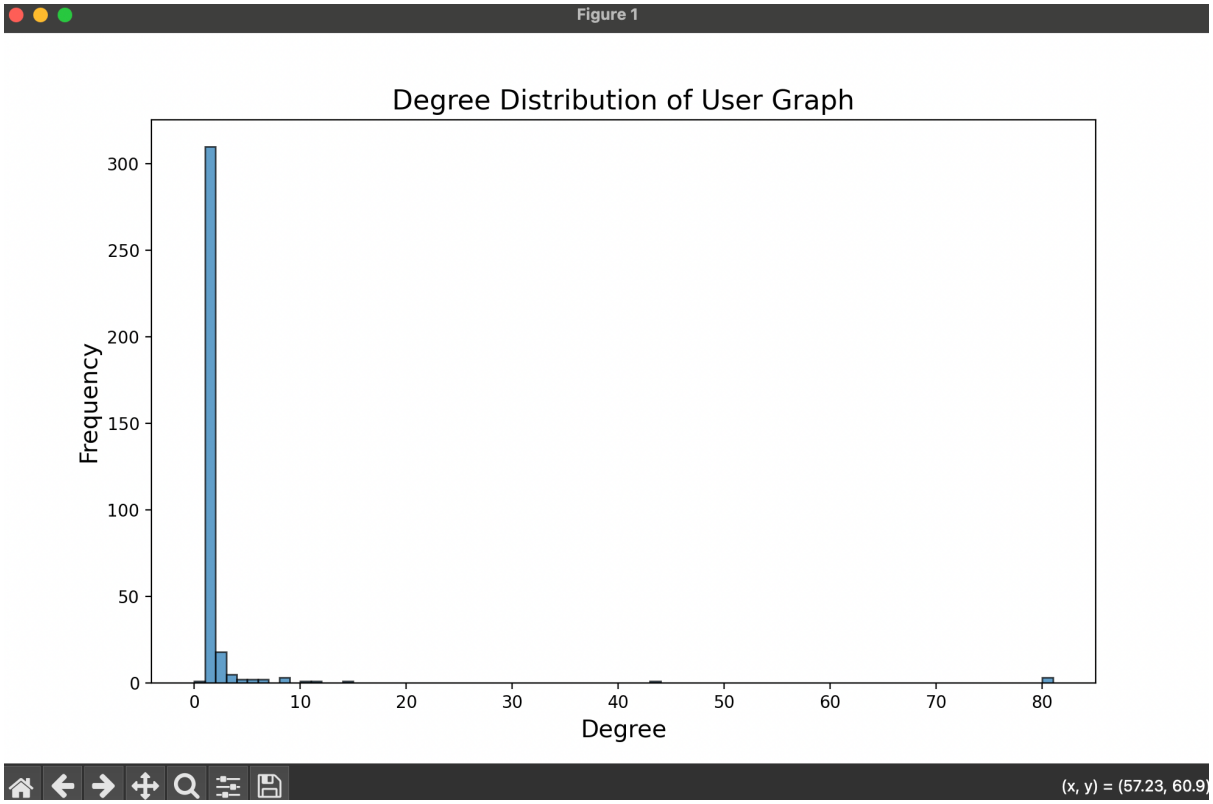
On my analysis, I found that toxic comments and posts are about 78% of the posts and non_toxic are about 22%.
This is displayed in the histogram above.

Network Measures

I used functions in networkx to get the degree distribution histogram, closeness centrality histogram and betweenness centrality histogram.

Degree distribution histogram

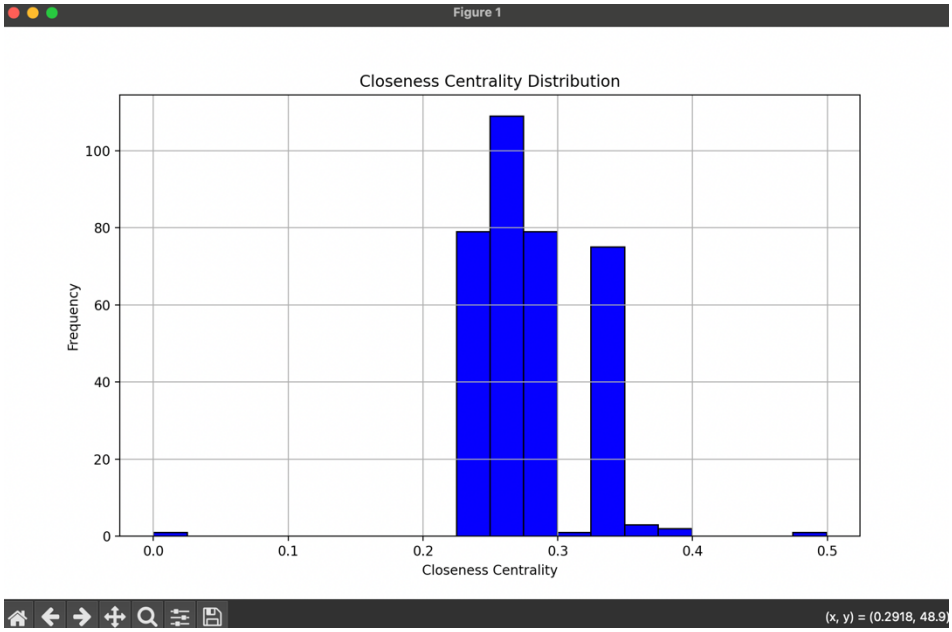
Sarthak Aggarwal ASU ID#1222329095



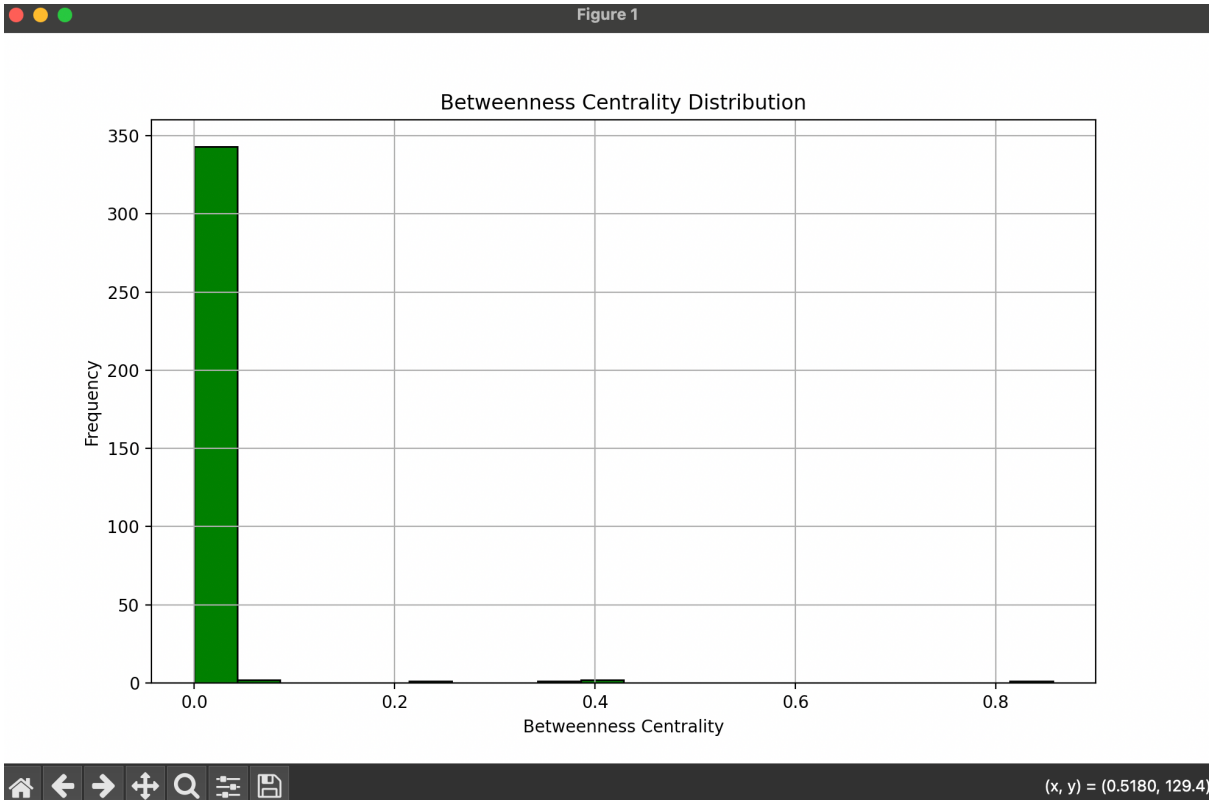
I fetched all the degrees for all the nodes of the graph. Then I plot the histogram.

Closeness centrality histogram

Sarthak Aggarwal ASU ID#1222329095



Betweenness centrality histogram



1 hop friend average : Too long! Look at it in the terminal while running.
Global Average Number of Friends (Across Entire Network): 2.10

Sarthak Aggarwal ASU ID#1222329095

Analysis:

When studying the social media followers of Trump and Harris for the 2024 elections interactive graphs display a pattern; most users have only a few connections while a small number have many connections that point to them being influential figures, with broad networks in the system. These influential individuals likely have a role in spreading content or opinions within their communities. The distribution of closeness centrality highlights that the majority of users are grouped around a value. Indicating they are neither too distant, from the network hubs nor entirely isolated. This suggests that the network allows for communication, between its sections to a certain extent. Though the distribution of centrality is unevenly skewed – most users exhibit betweenness values while only a few play crucial roles as connectors linking distinct parts of the network. These pivotal nodes wield influence in connecting separate communities and present opportunities, for strategic messaging or influence efforts.

References:

Mastodon documentation: https://mastodonpy.readthedocs.io/en/stable/
Networkx documentation: https://networkx.org/
Llama 3-8b: https://colab.research.google.com/drive/1lnnyfcR4rI_qRTmwqy-isiXdutGPFciu?usp=sharing
OpenAI ChatGPT: https://chatgpt.com/