CS 579: Online Social Network Analysis

Project I - Social Media Data Analysis

Team Members:-

Bhumiben Hiteshbhai Patel(A20502661)

Vismaya M (A20519405)

Social Media Used: Reddit

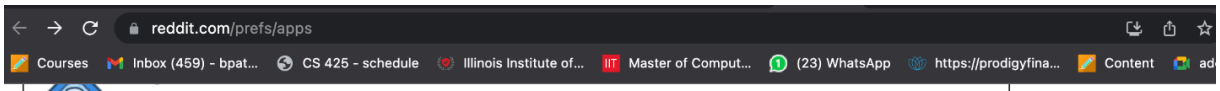
1. Introduction:

In this project, we visualized and extracted the data from Reddit using Python Language to find the relation between the author of the post and author of comments on the post. The data has been collected through the API provided by Reddit. Then we built a network model using python from the data extracted. In the last step, we calculated network measures for the obtained network.

2. Methodology

2.1 Data Collection and Processing:

We have created a Reddit developer account and got the Key and named as Project1
We use Praw to connect and crawl data from Reddit.



Spend reddit gold creddits
Mod Note
Approve and ban users
New Modmail
Moderate Subreddit Configuration
Edit My Subscriptions
Edit structured styles
Read Wiki Pages
Wiki Editing
Vote
My Subreddits
Submit Content
Moderation Log
Moderate Posts
Moderate Flair
Save Content
Invite or remove other moderators
Read Content
Private Messages
Report content
My Identity
Manage live threads
Update account information
Subreddit Traffic
Edit Posts
Moderate Wiki
Make changes to your subreddit
moderator and contributor status
Manage My Flair
History

revoke access

Developers: jankii27

reddit on mobile web (installed)

reddit on mobile web

My Personally identifying information
Spend reddit gold creddits
Mod Note
Approve and ban users
New Modmail
Moderate Subreddit Configuration
Edit My Subscriptions
Edit structured styles
Read Wiki Pages
Wiki Editing
Make changes to my Identity
Write subreddit content
History
Vote
My Subreddits
read moderator
Submit Content
read moderator
Moderation Log
Moderate Posts
Moderate Flair
Spend reddit gold creddits
Read reddit gold creddits
Save Content
Invite or remove other moderators
My Email Address
read vote
Read Content
Read my Identity
Update account information
Private Messages
Chat read
Report content
read subreddit content
My Identity
Manage live threads
Read My Subscriptions
Subreddit Traffic
History
Edit Posts
Edit My Subscriptions
Chat write
Moderate Wiki
Make changes to your subreddit
moderator and contributor status
History
write vote
Manage My Flair

revoke access

Developers: reddit

developed applications

Project1
personal use script
AOtDTmgh99EXMLnBvjXGZg

Creating this project for Assignment

edit

Developers: jankii27

create another app...

DataSet Information:

```
Data:
                 Source              Target
0             doktorinjh          Garandhero
1             doktorinjh            Sirtoshi
2                zonye10          doktorinjh
3                zonye10   jasontheguitarist
4                zonye10      TheGoldenPanda
5                zonye10           ozbourne8
6                zonye10    encephalophiliac
7                zonye10       iBleeedorange
8                zonye10          Garandhero
9                zonye10            Sirtoshi
10          beetnemesis         Metallkiller
11            bothunter          beetnemesis
12            bothunter         Metallkiller
13  Kindletookawaymyjob           bothunter
14  Kindletookawaymyjob         CaptainMuon
15               fwl200  Kindletookawaymyjob
16  Kindletookawaymyjob              fwl200
17              Yojenkz             itisike
18  Kindletookawaymyjob             Yojenkz
19  Kindletookawaymyjob       AkumaBengoshi



Count of Nodes & Edges:
Nodes: 197 | Edges: 261
```

The following process was followed for data collection and cleaning.

1. Praw was used to connect the python client to Reddit API
2. A request for subreddit python/programming was made with the top 25 posts.
3. For each of the posts, a recursive call was made to get the comments with an upper limit of 10 comments.
4. A data frame was created with the post author as the node and the comment author as the destination.
5. The author of the post and the author of the comment were represented as nodes. The edge
   Will represent the connection between the author of the comment and the author of the post.
6. In a recursive call, a csv is made and data is stored under columns Source and Target. Columns of csv represent the nodes while each row represents an edge between the two nodes.
7. To ensure that the author is not repeated as a node we have made a list of previously added nodes and at each successive recursion, this list is cross-referenced to prevent multiple nodes.

As the last step in this collection part, we have saved the data(only the IDs of the accounts) in the data frame in 2 columns named 'Source', and 'Target'. And saved this data in a nodedate.CSV file.

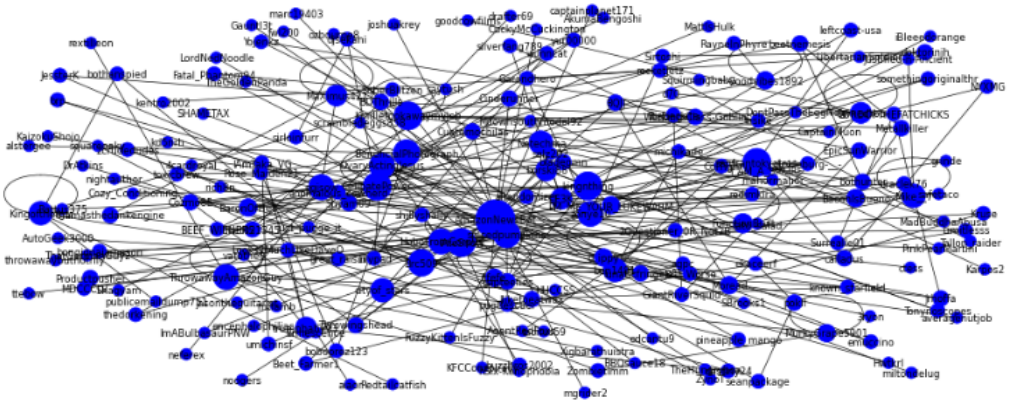
2.2 Data Visualization:

Software Used:- Python

Packages Used:- Networkx, Pandas, Matplotlib

The data has been read from the CSV file and we have stored each column of the data frame in separate variables. Labeled formats of the data have been displayed. The more nodes there are, the more congested the display becomes. The nodes with a higher degree have correspondingly greater sizes due to the visualization settings. Meaning that the node will be more significant the higher its degree.

We have got total of 197 nodes and 261 edges.



2.3 Network Measures Calculation:

Software Used: Python

Packages Used: Networkx, Matplotlib, Praw

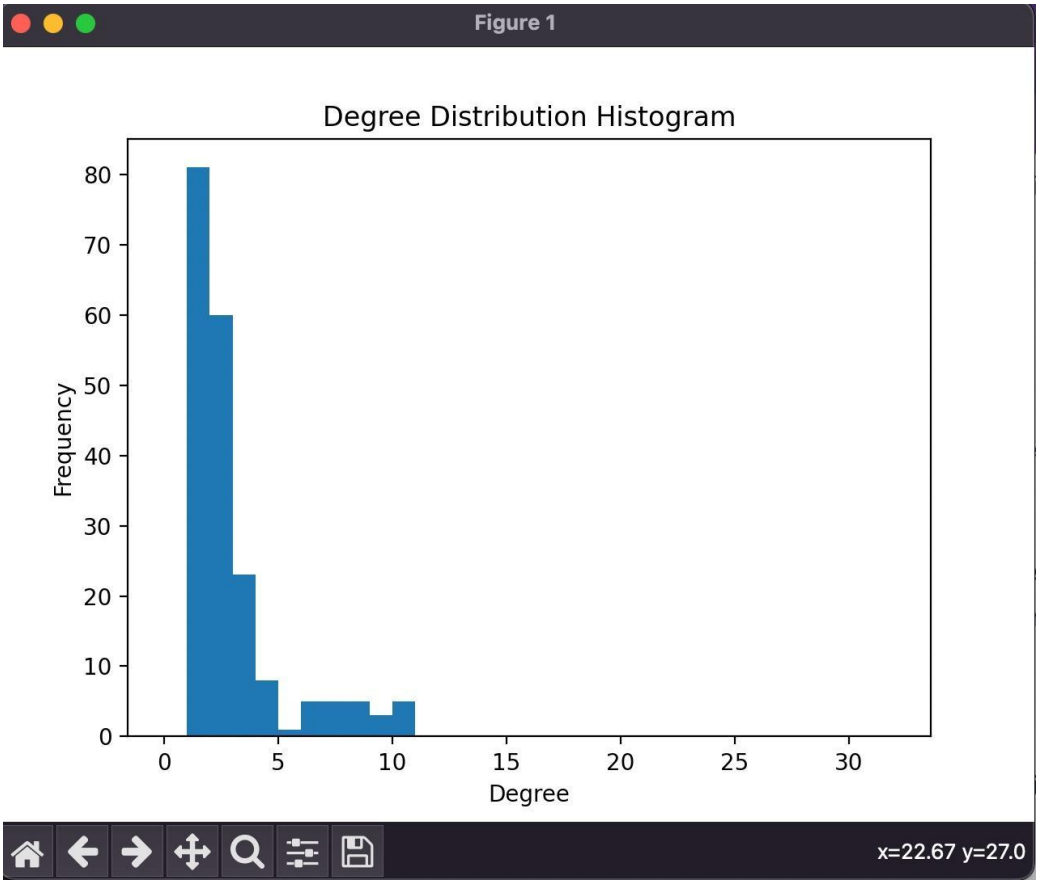
The praw module is used for interacting with the Reddit API

using Python.

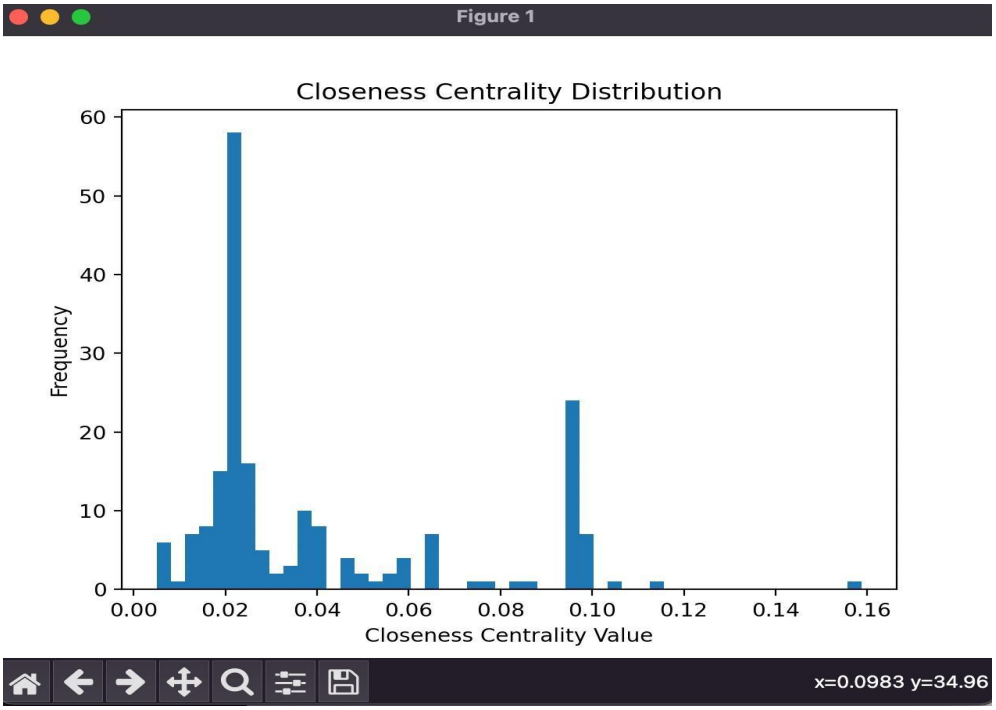
In this step, we have calculated various network measures like Degree Distribution, Degree Centrality, Closeness Centrality, Betweenness Centrality, Katz Centrality, and Page Rank. We have depicted the histogram for each of these measures. The average and the median values of this centrality are also calculated.

2.3.1. Degree Distribution

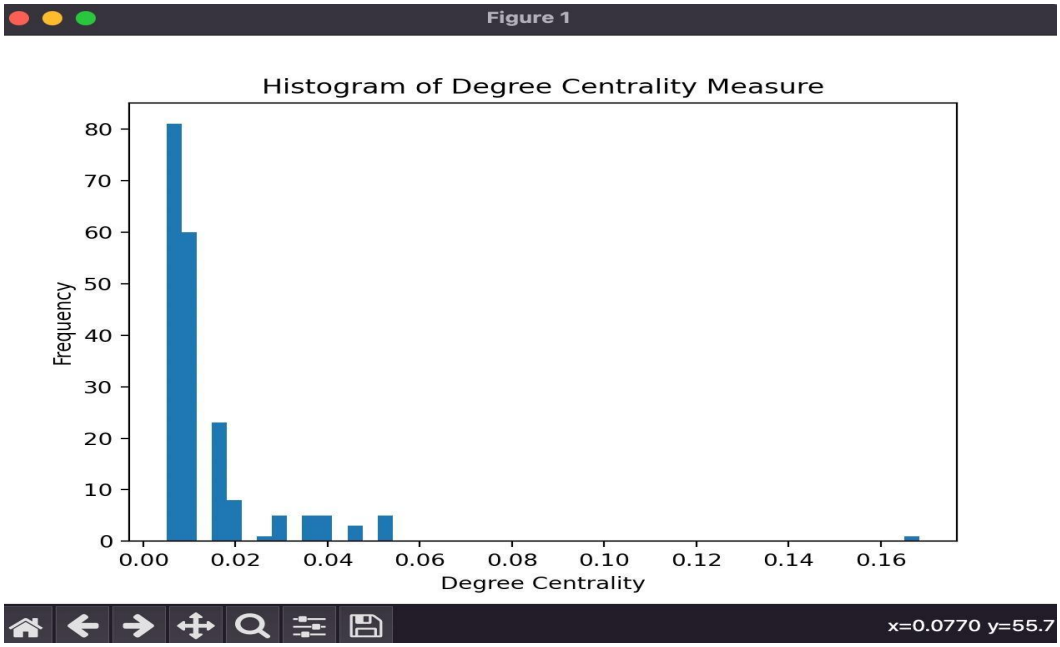
It is calculated for the nodes and the number of connections a node has with other nodes determines its degree within the network. The figure represents the degree distribution using a histogram for “nodedate.csv”



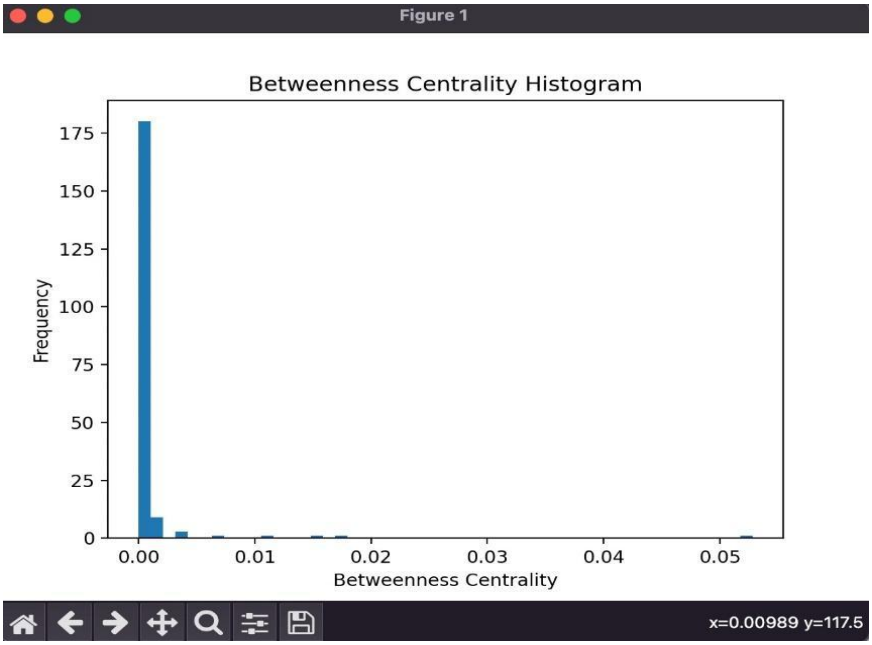
2.3.2. Closeness Centrality: It is a measurement of the shortest path on average between each vertex and its neighbor



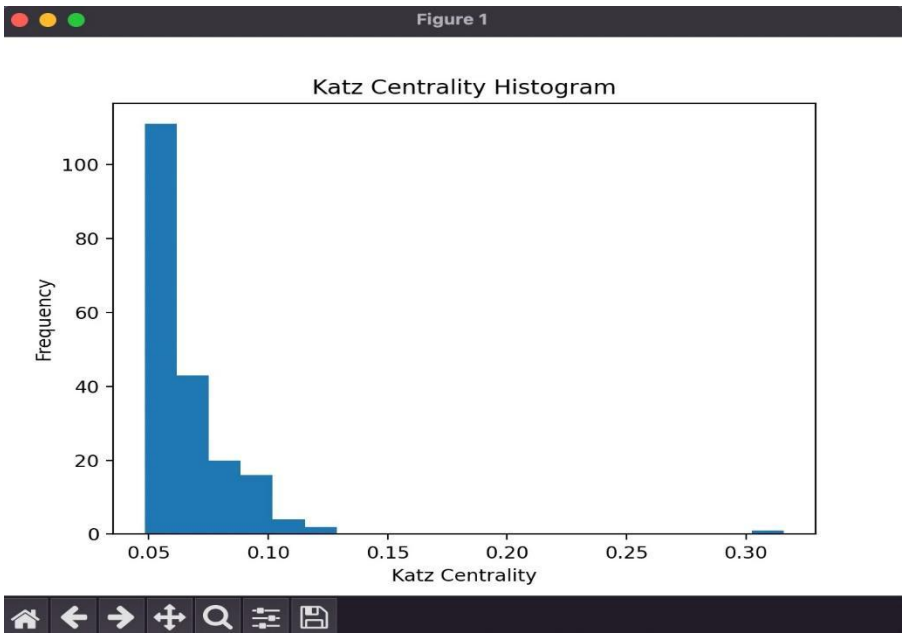
2.3.3. Degree Centrality: The number of times the node has.



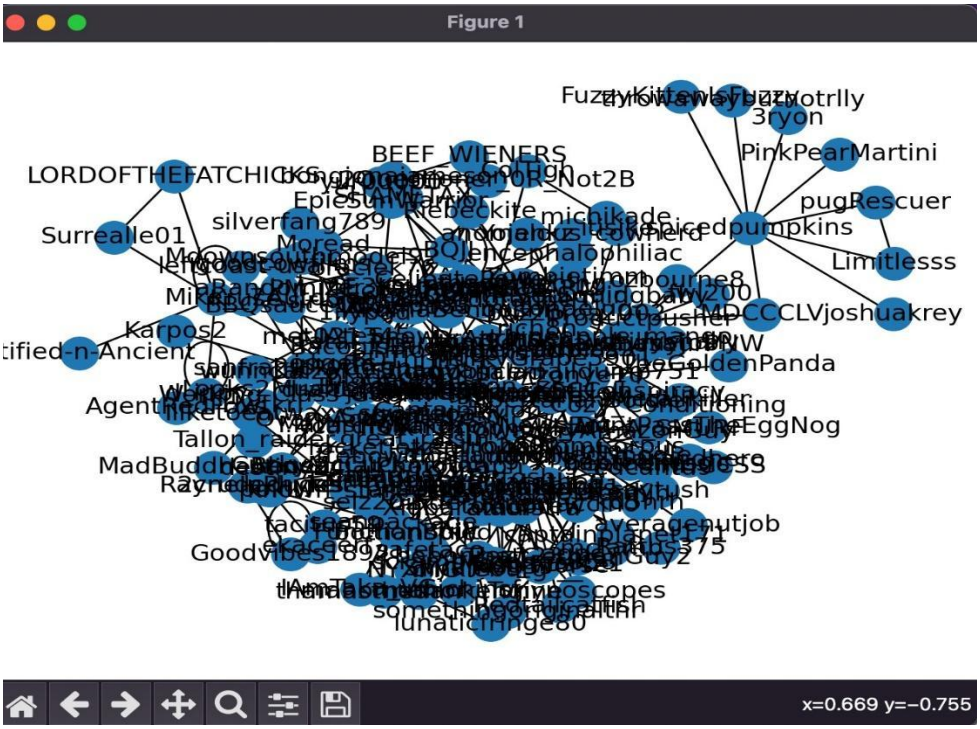
2.3.4. Betweenness Centrality: It is a centrality metric based on the shortest routes in a network.



2.3.5. Katz Centrality: It is the measurement of the centrality in a network.



2.3.6. PageRank Centrality: It is used to know how important a node is by the number and quality of edges to each node.



References:

1. Software Used To run python code : Google Colab
2. Matplot lib, https://matplotlib.org/
3. Pandas, https://pandas.pydata.org/
4. NetworkX Documentation : https://networkx.org/documentation/networkx-1.9/

5.Praw : https://praw.readthedocs.io/en/stable/

Contribution:

Bhumiben Patel

Developed code for collecting and processing, Visualization data for project requirements, and report formation.

Vismaya M

Developed code for network measures and calculations and report formation.