Daniel Ayabe, Remi Kalbe

Prof. Hood

CS 579

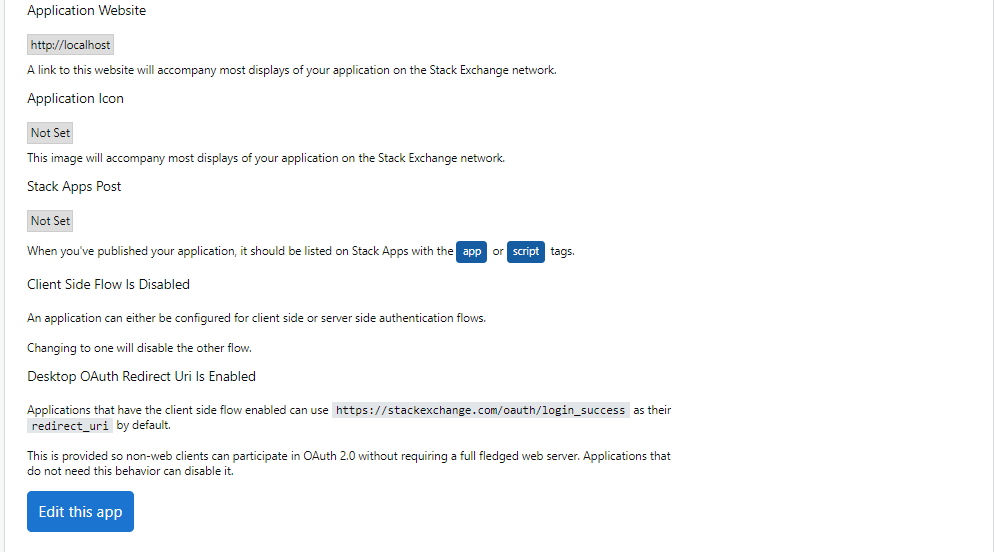
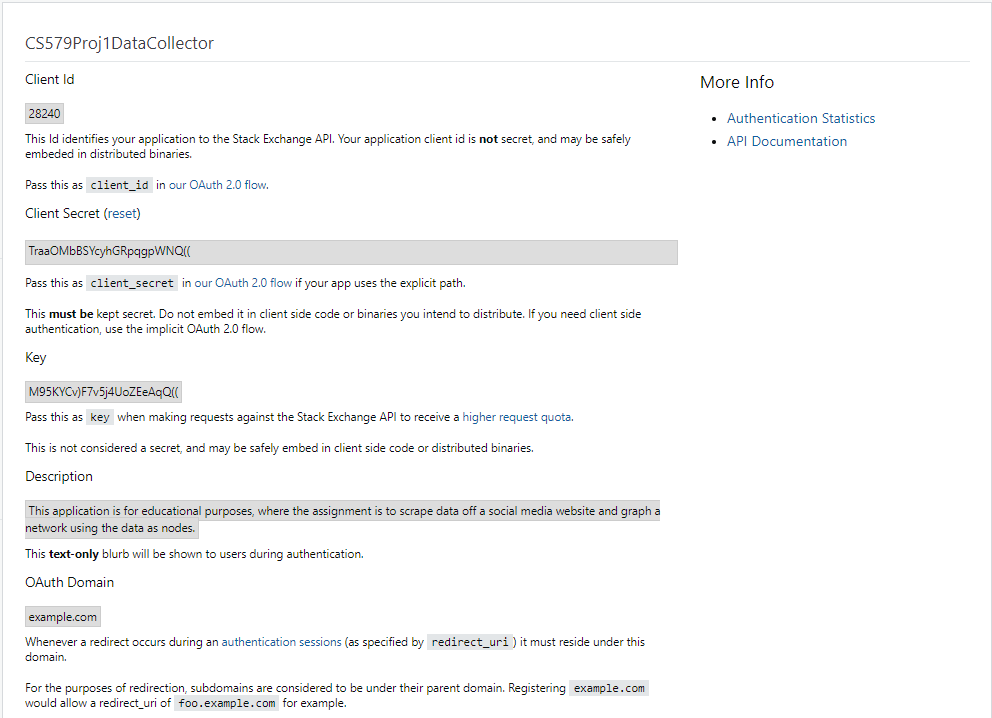
02/20/2024

**Project 1: Social Media Data Analysis**

Data Collection:

For data collection, originally, we had decided on selecting Reddit as our source of social media to analyze, with using API methods in order to gather data. However, when we looked further into checking the data licenses and the privacy licenses, it required too much effort just to use the Reddit data for educational purposes. We then decided to switch to Stack Overflow, where the policies were very simple to follow in order to legally use the data on the website for educational purposes.

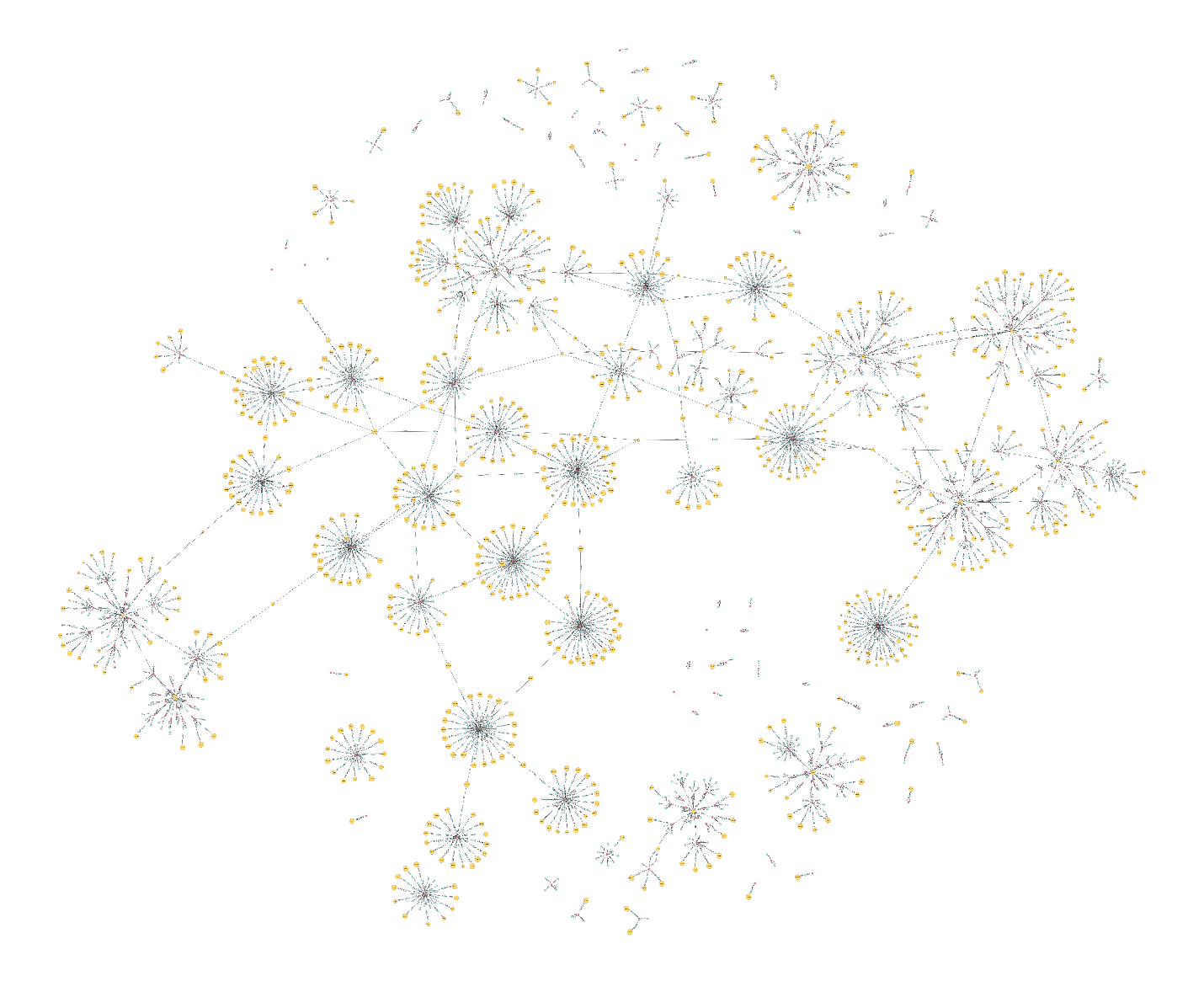
Next, we decided to create a Python script that would gather data of questions asked, filtering the questions, comments, authors, and highest upvotes specifically. The code can be found in the collection folder of our project repository. To make sure that we did not violate any of the policies mentioned beforehand, we had to create a stackapps project do receive a client ID in order to use the Stack Exchange API, which is shown below in the 2 screenshots provided.



Data Visualization:

For the data visualization, we specified certain filters to gain specific aspects of information from each stack overflow post. The user’s account and user IDs, number of answers and question the user has asked and given, the number of upvotes and downvotes, the acceptance rate of the user’s answers, their location, and lastly, their reputation.

Next, we decided to use a graph analysis software to visualize our network as a graph. We chose graph-tool because it is a Python module, and we wrote our code in Python. Its memory usage and computational time are also efficient when graphing too. Along with that, we also needed to install pixi for managing the dependencies when using graph-tool. Our format for our data input files were in the .parquet file format because this file format enables less data scanned at ingestion, meaning that faster scans are performed with lower query and memory costs because the file format enables you to only read the columns of data that are needed. Below is the screenshot of the graph.



The users are represented in yellow; the questions are represented in red; and the answers are represented in blue. There are a significant number of instances where a user only replied to one question (when the users are in the outer edge of the circle) as supposed to some instances where the users reply to more than one question (when the users are in the center of the circle). In addition, there are also questions and answers with no user, most likely due to deletion of accounts.

Network Measures:

Degree Distribution + 2 more network measures (plot as histogram)

Discussion of Results:

What insights do these results provide?

What further questions do these results raise?

What would be the next step to investigate further be?

References

* <https://stackoverflow.com/legal/terms-of-service/public>
* <https://stackapps.com>
* <https://api.stackexchange.com/docs>
* <https://stackapps.com/apps/oauth/register>
* <https://graph-tool.skewed.de/>