Sam Valenzuela
DS210 Final Report

The Reddit Hyperlink Network

The dataset I used contained interactions between different subreddits, which are forums dedicated to specific topics, on the social network website Reddit. I focused on the columns containing the source subreddit and the target subreddit that was mentioned. The original dataset had about 285,000 entries, which my program was not able to process in a timely manner, so I had to cut down the dataset to the first 10,000 entries. In the end, it still took about 10 minutes for the program to compile.

To construct the graph, I assigned a number uid to each subreddit, making sure that I gave duplicates the same uid. The graph is a directed one because the source and target subreddit matter. I noticed that this graph had a lot of detached nodes, often with a node connected to another one and that being it. I changed it so the console only outputted subreddits that mentioned other subreddits at least 5 times, and that resulted in 395 subreddits from the 4,075 unique subreddits in the condensed dataset. I was unable to get traversals to work as the compiler kept giving me panic errors, so I had to review the graph visually. I pasted the console output to "10k test results.txt".

There are some subreddits that mentioned others a lot more frequently than others. r/france, r/dailydot, r/subredditdrama, r/badhistory, r/hailcorporate, r/circlebroke, and r/dogecoin all had high counts. It's cool to see how these are all different genres, showing the vast range of discussions that can occur on the internet. For example, r/hailcorporate shows examples of unintentional advertising, and because of this, the subreddits they target are a wide variety of genres such as r/pics and r/android. r/dogecoin focuses on the cryptocurrency of the same name and as a result stays mostly within the sphere of cryptocurrencies, such as r/bitcoin and r/litecoin.

There are also certain subreddits that get mentioned more than others. r/askreddit was mentioned the most in this dataset, with 246 mentions, followed by r/iama, r/pics, r/todayilearned, and r/funny. These are all general subreddits that are catered to the common folk and aren't niche-based, which makes sense why they are mentioned the most.

It seems most subreddits mention other subreddits of a similar genre, but there are some funny outliers. Some examples are r/math → r/conspiracy, r/sweden → r/syriancivilwar, r/redditdads → many GTA communities. There are many clusters of subreddits that interact with each other based on genre, but every now and then there are subreddits that are very different and connected to each other. It really shows how vast the internet is and all the connections that can be made.