Week 2 Part 2

We found that the subreddit dataset has a categorical value called POST_LABEL, which is a label that indicates whether or not a post is explicitly negative towards the target post. The possible labels are -1 and 1. If the source is negative towards the target, the hyperlink gets a -1, and if positive or neutral it gets a 1. These labels were obtained via a classifier trained on crowd sourced data.

This data set is very large, and we have already attempted to load the data (a tsv file) as a dataframe and creating a graph in python. It might be more optimal to read the data via functions native to networkx. Drawing out the graph in python is unfeasible due to its size (we tried, it took 5 hours to draw the whole thing). Gehpi turned out to be a better choice, and we can look into analysis options provided by Gephi. Neo4j might also provide some solutions.

After loading the data, we can focus on analyzing the categorical value POST_LABEL and seeing how “negative” the links between reddit subreddits are. With this value, we not only know the frequency of links between subreddits, but we can see what the general sentiment is between them- perhaps some subreddits are antagonistic towards one another, and others more collaborative.