CS 170A Project Proposal

I'm an avid fan of Reddit, and from the perspective of a data-analysis project, it is perfect because it is basically a crowd-sourced categorization of the web into cultures and topics. I feel like a good way to use to use this data is to start building a model of how different ideas and topics relate to each other. For example, rock music and pop music both fall under the larger category of "music". We could start to see this relationship in Reddit posts if we notice that all posts in the "rock music" and "pop music" subreddits are also being posted in "music", but not all "music" posts are being posted in the "rock music" and "pop music" subreddits. In addition, we can determine which subreddits are founded on contradicting opinions, for example, the subreddits for Hillary Clinton and the subreddit for Donald Trump, by analyzing how different posts are received (positively or negatives) on those subreddits. There are many datasets I could use, but the one I will mainly be using is this one:

https://github.com/umbrae/reddit-top-2.5-million

Which contains the top 1000 posts from each of the top 2500 subreddits. While I am only concerned with cross-posts (posts that were submitted to multiple subreddits), there should still be a significant amount of those types of posts in this dataset, because top content tends to be posted to multiple subreddits. I will mainly be using covariance and correlation analysis, but I may start to look at specific features of the posts (eg the website itself, and the content of the website that was posted) to further increase the accuracy of the hierarchal model of subreddits I will be making.