Social Network Analysis of reddit.com

Dimitrios Economou, Thomas Horn, Josh Weaver

Objective

- Cluster subreddits using metrics and clustering methods
- Compare results to LDA Topic Modeling
- Explore topical differences between different subreddits and types of subreddits (discussion-based or content-based)

Finished Work

- Scrubbed comment objects of data we are uninterested in
- Generated statistical data from our dataset including average number of comments per thread and per subreddit
- Initial clustering using Kmeans with Cosine Similarities of a TF-IDF matrix

Remaining Work

- Cluster subreddits using different metrics
- Cluster subreddits using different clustering techniques
- Find ideal number of clusters for our dataset using NbClust
- Perform Topic Modeling
- Repeat on larger dataset

Difficulties

- Very large dataset
- Takes a lot of time to process and analyze
- Learning some natural language processing
- Formatting data for various algorithms

Questions?