CS225 Final Project Team Contract: Project Goals:

Signatures: Adish Patil (adish2), Yash Gupta (yashg3), Mike Lee (dcl3), Jaehan Kim (jaehank2)

Dataset:

Reddit - http://snap.stanford.edu/data/soc-RedditHyperlinks.html

http://snap.stanford.edu/conflict/

https://pushshift.io/

- Summary:
  - The hyperlink network represents the directed connections between two subreddits (a subreddit is a community on Reddit). We also provide subreddit embeddings.
  - Format: The network is directed, signed, temporal, and attributed.
  - Each hyperlink is annotated with three properties: the timestamp, the sentiment of the source community post towards the target community post, and the text property vector of the source post.
  - Network is directed, temporal, signed and attributed
  - Each post has a title and body.
  - Hyperlink is present either in the body/title of the post
    - Dataset provides one network file for each.
- Project Implementation:
  - Tsv datafile will be converted to csv for the purposes of our project
  - The multiple edges between a subreddit and target_subreddit can be combined into one directed edge
    - Sentiment will be added up and averaged. This will become our edge weights between vertices
    - We will still store the count of the original amount of edges between two subreddits

    - Create a connected graph from the csv file

Traversals:

- We aim to be able to try and implement both BFS and DFS traversals for this project
- Aim to use DFS for our strongly connected components function
- Aim to use BFS to find the shortest path between two vertices

Covered Algorithms:

- Shortest Path
    - Since the edges are weighted using -1 or 1 (representing the sentiment of a post linking two subreddits), we can't use any algorithms that require the edge weight to represent distance between two nodes. Therefore, our shortest path algorithm will be using a modified BFS(geeksforgeeks).
    - If two subreddits are not connected directly, then we can find the shortest amount of subreddits we need to jump through to find how they can be connected through hyperlinks.
    - We can have an input of 2 subreddits that will find the shortest path between them.

Complex or Uncovered Algorithms:

- Iterative Deepening Depth-First Search
    - Perform DFS continuously but with a limited depth: IDSUtil
    - Alternate for BFS as they both return the shortest path
- Strongly connected components
    - Grouping subreddits by common interests by using strongly connected component analysis.
    - Implementing Kosaraju's Algorithm and Tarjan's Algorithm.