DATA620 Final Project Proposal

Jonathan Hernandez

My final project will be based on examining questions and answers from stackoverflow that uses the ‘R’ tag. The data set is from kaggle.com: <https://www.kaggle.com/stackoverflow/rquestions>

This project will attempt to combine both topics that was taught in the class: Network Analysis and Text Processing.

- Motivation: By looking at the Q&A of R (statistical programming language), we can do analysis on what words are mostly used in answering these questions and also see users who use these words. We can also see the different types of words for questions as well, see their rank and determine if it is a question that is good to ask. Also looking at the graph of users and questions answered can help visually in who is getting the accepted answers and scores.

- Questions to answer:

1. Is there any difference in frequency distribution between accepted answers and not accepted such as types of words used?

2. Who are the top users with the most accepted answers?

Question 1 can be answered by doing text analysis and NLP to see the words used and splitting the analysis by answers accepted and not accepted and see if there is any correlation is there. Using a t-test will probably be the best choice.

Question 2 can be answered by making a network graph of users where the nodes are users and the edges are if a question was answered. A edge weight is the score and the edge label is if it was accepted or not. The last thing to do is use a method like the island method or other option to see the top users

- Retrieving data: Data will be downloaded from the website mentioned above as 3 seperate csv files. The files are questions, answers and tags. The tags.csv will be a dataset that includes other tags that are used besides the ‘R’ tag.

- Approach: The approach will be to first combine the 3 csv files into one similar to a join in SQL. Next step is to use python’s nltk library and functions to examine each answer (or question) into words and begin doing frequency analysis. The next step will be to take the entire corpus of words and split them based on if they were found to be accepted answers and not accepted, and use a t-test or some other metric to examine the difference. Finally, a network graph can be drawn for question 2 and different metrics can be computed on the graph.