https://snap.stanford.edu/data/twitter7.html
The data is 467 Twitter posts from 20 million users over a 7 month period. This is estimated to be about 20-30% of all public tweets published on Twitter during that time. Some categorical variables would be Author, Time, and Content. I would use R to process this data. First, I would load the data via csv files. If there are any text files, I would covert them to csv files. Next, I would use R libraries to clean up and shape the data for use. I would use Model building techniques I learned in Data 621, to create models such as for multilinear distribution. Based on summary statistics of the models such as the R-squared value, I would choose best model that would suggest useful to use. These models would be useful to compare degree centrality across categorical groups and describing hypothetical outcomes such as studying to find if content increased with time.