**Stacey Schwarcz**

**Data 620**

**Fall 2016 Final Project Proposal**

**Assignment instructions**

*Your proposal should describe at a high level what you’re seeking to accomplish, and your motivation for performing this analysis. A guiding question or hypothesis to test is one good way to start. If you are going to work in a small group (encouraged!), you should also list your partners’ names.*

*You should briefly describe your data sources, plan for doing the work, and up front concerns. If you are working in a group, please describe the roles and responsibilities of each group member.*

*We’ll treat this proposal as a planning document, not a blue print containing “firm, fixed requirements.”*

**Goal**

This project will explore the network of stack exchange users for a particular topic as well as analyze the connections between the text of user comments and the score those comments received.

**Data Source Description:**

The data source that I plan to use is the Stack Exchange Data Dump by Stack Exchange, Inc.,

Published September 12, 2016

Stack Exchange is a network of question-and-answer websites on topics in varied fields, each site covering a specific topic, where questions, answers, and users are subject to a reputation award process.

The data dump is an anonymized dump of all user-contributed content on the Stack Exchange network. Each site is formatted as a separate archive consisting of XML files zipped via 7-zip using bzip2 compression. Each site archive includes Posts, Users, Votes, Comments, PostHistory and PostLinks. Schema information is in the readme.txt which is part of the Archive index containing all files which can be found here: https://archive.org/download/stackexchange.

**Plan for doing the work**

The data dump is divided by subject matter, with some subjects containing much larger archives than others. I plan to begin by downloading one of the smaller subject archives to make set up and testing easier. Once I have the code working on the smaller data set I will download a larger data set so there is more data to work with for network and text analysis.

*Network Analysis*

For the network analysis I will use the Posts and Comments files, and extract post id and user id. Users will be the nodes, and edges will be connections between users. I will consider a user connected to another user if they commented/posted on the same post. I will process the data into the correct format and use networkx to create a network diagram and analyze network measures including degree centrality and eigenvector centrality to determine which users have the most influence in the network. I will compare the top results for each of these to the users with the highest reputation scores in the users file. I also plan to use the Island method to separate into subparts, and then analyze cliques and triads among the user network.

*Text Analysis*

For the text analysis I will use the comments file, specifically the score and text of the comment. I will look for connections between the length of a comment or any particular words in the comment that are associated with a higher score, to see if it is possible to predict the strength of a comment based on text patterns. In the dataset the score is the net of up and down votes that a post receives from registered stack exchange users even if they did not post or comment on a question.

**Up front concerns.**

I will need to determine how to extract 7zip files and to work with the resulting XML files in the archive. I will then need to get the data into the correct format so that I can create a network model. I don’t anticipate this being very difficult, but I have not worked with 7zip or XML files in python before.

The user id only exists if the user is still active, so I will exclude data where the user id is blank.

Initially I thought I would consider the score on the post or comment and whether the user is a poster or commenter for a particular post as part of the network analysis, but this appears to be more complex than I initially thought, so I will likely not pursue this.