Homework 1 Reflection

Sian contributed by combing through the news articles and the provided Scientific articles to pull out relevant events regarding BSE/mad cow disease and the labs. She also worked with the rest of the group to analyze the voter registration data and picked out the pattern of missing addresses. In collaboration with Sarah, she helped determine some of the key stakeholders. Sian also coded the timeline visualization that shows the events by type (politics-, BSE-, and lab-related).

Sarah helped comb through news articles and data. Within the articles, she looked for relevant political events and any stories mentioning John Torch. She also helped to analyze the voter registration data for any suspicious patterns. Although we brainstormed visualizations as a group, Sarah drew up the sketches and created the network graph included in the final report. She also assisted with compiling the final report.

Cindy helped with data cleaning and data visualization in R. She cleaned the data by lower-cased data (ex. There are 2 democrat party groups in the voter registry dataset, one in uppercase and the other in lowercase), and wrangled data to filter out variables that are useful in creating the visualizations. She created the cleaned dataset and pushed it to the repository for reference. She also tried to help with creating a word could using the exported word list created by Yuhan using AntConc which helps the team to diagnose themes/stories of the city.

Yuhan mainly contributed to the large-scale text analysis of the news articles. She used Python to rename and clean the news articles, renamed the news files based on time to facilitate review, and combined the news into a comipled_news file for large-scale analysis. She used AntConc to gather high-level linguistic insights into the news articles, which supported some hypotheses from the events timeline and the generation of the word cloud.

One of the main obstacles we encountered was putting the data in the context of the question. When we first approached the assignment, the messy and atypical data was unlike anything we had encountered in our previous classes. This was especially true when we realized that none of the data connected neatly to one another or came in the typical CSV format. Without previous experience with large quantities of text data we initially struggled to come up with a clear plan for understanding, compiling, and analyzing the data. To address this, we decided to split into 2 small groups– one focused on understanding the spreadsheet data, and the other on the text data. The spreadsheet data group performed exploratory data analysis, applying skills from previous classes, and the text data group read through the scientific reports and news articles to create a timeline of major events. After this process, the two groups came together to compare findings. The text data group came up with some hypotheses based on the stories, which could be supported by patterns in the voter registry discovered by the data analysis group. Combining our insights, we used the questions to guide our further analysis. Eventually, we were able to develop a hypothesis that we felt was supported by both the timeline and visualizations.

Since this was our first time solving this type of assignment, we struggled to plan and divide the work effectively. A big factor in these types of assignments is time, for homework two we should at least familiarize ourselves with the data and question earlier in the week. After doing so, we can regroup and discuss our impressions, findings, and thoughts, and apply them to the context of the problem. Based on peoples' interests we can once again split into groups to analyze the data. Given our common goal and understanding of the question, each group would be better able to extract insights from the data and seamlessly join together their findings. Finally, we can compile all the discoveries into hypotheses backed by data.