Capstone Three: Project Proposal

For my third capstone project, I will be analyzing a dataset of 19,320 bloggers and their blogs collected from blogger.com to determine whether there is a difference in the language used between male and female bloggers. Throughout the course of my project, I will most likely discover other insights about my data, as it includes more features than just gender. The results of this project could be important, not just to me due to personal curiosity, but to researchers in fields involving interpersonal communication. Many inconclusive studies have been conducted on the amount of words women and men use in conversation, yet the amount of studies on differences in vocabulary is far less. Current research has concluded that men speak in more abstract terms, while women are more focused on details. The results of my projects could be supplemental to the efforts of these researchers.

For context, the current stereotype is that women use more words than men. A common assertion is that women use 20,000 words a day while men use 7,000, yet there is no scientific basis for this claim. While subsequent studies have found negligible differences in the amount of words men and women speak per day, far less studies have been conducted on the gender differences in vernacular. In general, these studies have found that men speak in terms of the big picture, while women tend to focus on specifics. This was done by assigning scores of abstractness to specific words and measuring the frequency of these words in the speech and text of men and women. Besides just abstractness, I will seek some other form of NLP analyses that may highlight the differences between the words that males and females use. The criteria of success for this project would be discovering a predictor, abstractness or otherwise, that would be able to predict whether a blog was written by a man or woman accurately. The scope of solution space would be any sort of speech pattern I have available to analyze. The constraint within solution space is that this dataset only contains blogs, so data has not been procured in an experimental laboratory setting. Stakeholders to provide key insight will be my mentor, Eleanor Thomas, as well as my peers who I may share this project with. The key data source will come from the Blog Authorship Corpus, collected by Schler, Koppel, Argamon, and Pennebaker (2006).

The Blog Authorship Corpus has been scraped from blogger.com. After making sure the data is clean and workable, I will analyze it for correlations age, gender, and text used. This will potentially include variables such as word abstractness, sentiment analysis, or maybe even word length. I will then compare the strength of correlations between these variables with gender and age to see what combinations of these features results in the most accurate prediction of the blogger’s identity. My deliverable will be a GitHub repo containing the work for each step of my project, including a slide deck and project report.