Sidharth Dhawan
IW Checkpoing Form
November 9, 2015

- Intro
  - The goal of my thesis is to learn something about the personality (for example, happiness levels, or different axes of personality as dictated by the Meyers-Briggs personality test) from the way a person writes letters. So far, I have focused my research mainly on the "optimistic / pessimistic" dimension of other personality; however, I may try to analyze others.
  - For my project, I am using a dataset consisting of written letters, available in the Library of Congress. I am using comprehensive lists of letters written by John Adams, George Washington, and Thomas Jefferson.
- Dataset Collection
  - So far, I have built a crawler and collected my data for John Adams and George Washington from the website [INSERT URL]. For each president, I have collected about 6,000 written letters. These letters consist of both businesslike, impersonal letters, and more personal correspondences with loved ones.

- Introductory Analysis of Data
  - Archaic words and features of the dataset
    - I have also done some introductory analysis of the dataset. One of the interesting (and possibly complicating) results I found was the fact that many of the presidents' letters have archaic spellings of words, as well as archaic abbreviations of words, which will be difficult to find within standard NLP datasets such as Wordnet. In order to understand the intent behind these words, it may be useful to approximate words that cannot be found within wordnet, to the closest word within wordnet (using some kind of autocorrect technology).
  - Words most frequently used
    - I have also done some analysis of the words most frequently used by each president (not including stopwords like "the", "and", "or", etc.). I have performed a comparison of these words with the most frequent words used in standard english text collections (such as the Brown corpus). I found that certain words were more commonly used by all of the presidents than by the Brown Corpus – these words generally had to do with governments and wars. Additionally, there were many words that were very unique to each president – I believe this can potentially provide some useful information when trying to understand the personalities of each president. The contrast between Adams and Washington was stark – Adams was more likely to use big words (perhaps reflective of his extensive education) whereas Washington used many abbreviations, and many words dealing with war ("general", "mutilated", etc.).
  - Understanding what words each president uses in a more sentimental context
    - The final piece of introductory analysis I have conducted concerns the most sentimental words that each president uses. In order to find words that were both sentimental and frequently used, I created a new metric consisting of the sentiment score of the word (within "Sentiwordnet") multiplied by the frequency of the word within the dataset. Sorting by this metric, I was able to find very unique lists of words for each president. For example, John Adams' most sentimental words are often those concerned with doling out praise, for example "congratulations", "esteem", etc. In contrast,

Washington's most sentimental words paint the picture of a man who evaluates men more bluntly. His most sentimental words are “honorable”, “courageous”, etc.

- Next steps:
  - More data analysis
    - More thorough data analysis is needed in all of the directions outlined above. For example, the sentimentality metric discussed above contains a great deal of useful information; however, it also contains a great deal of noise. This is mostly due to the fact that many words can have different sentimentality, depending on the context in which they are used. For example, the word “cartoon”, when used as a noun, does not have a positive or negative connotation – it is merely a type of writing. However, when used as an adjective to describe someone, it can contain either positive connotations of being funny, or negative connotations of being overly inane. Thus, parts of speech tagging and greater understanding of context may reduce the amount of noise in these data.
    - Eventually, a metric is needed that will be able to correlate most strongly with optimism. An initial metric (which should be improved upon) is a sum of the metric described above for all of the most frequent words and phrases used by each author – if each of these metric scores is summed together, with a negative weight for words that are used in a negative context, it could give an interesting picture of how optimistic or positive each president is overall. Other potential techniques within may need to be investigated as well – for example, it may be better to investigate the sentimentality of each word or phrase as estimated for the current dataset. To rephrase, there are methods (such as Pointwise Mutual Information) that identify the positive or negative connotation of a given word or phrase within a given dataset. For example, if applied to John Adams' letters and the word “cartoon”, the PMI metric will reflect the overall positivity associated with John Adams' specific ways of using the word “cartoon”.
    - I plan to spend most of the remaining term (until the end of February) testing and fine-tuning different metrics. There should ultimately be roughly five metrics that I will end up testing.

  - evaluation of data with Amazon Mechanical Turk
    - After these metrics have been discovered and fine-tuned, the next step is to evaluate their performance. This will be done with the Amazon Mechanical Turk service, which provides “Human Intelligence Data” for certain tasks – in other words, it pays humans to manually evaluate datasets. Specifically, I will use the Amazon service to get my letters evaluated by human eyes. Each set of ~100 words will be evaluated on a 5-point scale, with 1 being strongly pessimistic, 2 being guardedly pessimistic, 3 being uncertain or very little emotion shown, 4 being guardedly optimistic, and 5 being optimistic. These scored will then be averaged to perform an objective optimistic and pessimistic score for each dataset.