Dealing with text

Text mining with Python

Why text mining?

- Results in real time
- Results from large volumes of text
- (Perhaps!) more objective

Sentiment analysis

- Naive Bayes Classifiers
  - Use word frequencies for classification
  - Construct a set of positive and negative words (or phrases)
  - Calculate word (phrase) frequencies and classify

Sentiment analysis

- Corpus of positive negative words
  - 'http://ptrckprry.com/course/ssd/data/positive-words.txt'
  - 'http://ptrckprry.com/course/ssd/data/negative-words.txt'
- Ideally, construct domain specific corpus
- Analyze sentiment
  * at the word level
  * at the sentence level
  * for named entities
  * for chunked words
  * over time

Concept identification

- Find frequently used words
  - frequency distributions
  - word clouds

- Identify named entities

- Look for adjectives (descriptive words)

- Look for differences in concept identification when comparing texts

Complexity

- Examine complexity of a text
  * Ease of understanding
  * Depth of ideas

Changes over time

- word dispersion over time
- frequency of word use over time
- sentiment of words/phrases over time

What we've done

- Which candidate has a more positive outlook
- Does a candidate get more positive or more negative as the debate proceeds
- How have sentiments changed from the first debate to the third debate
- What each candidate emphasizes (word cloud)
- How that emphasis changes from one debate to the next
- How the candidates differ on erudition
- The differences between our candidates (in erudition) and past Presidents
- Or has speech just gotten less complex since Washington
- We can tag words by their 'part of speech'
- And examine sentiment at that level
- Which candidate uses more positive words as descriptors and which more negative words (gives us a sense of the outlook of the candidate)
- What are the top words that both candidates use (perhaps what is important in this election)
- How do they differ on the implications of the things behind the words (sentiment)
- What words are different (differences in what they think is important)
- Are they positive or negative about these differences
- How the frequency of use of a word changes from the beginning of debate 1 to the end of debate 3 (interesting changes in the way they address each other)
- Identified bigrams and trigrams in the text (helps identify concepts)
- Figured out conditional frequency distributions to see how word use changes from one debate to the next
- Identified named entities in the text. This can be used to figure out which ones are positively viewed and which ones are negatively viewed