Text as Data: Homework 2

In this homework assignment we're going to analyze the first presidential debate from the 2012 election.

Problem 1

To analyze the debate, we first need to load the debate and parse the content. On the coursewebsite, you'll find the file `debate1.html`. Download the file and open it in a browser. We will use `BeautifulSoup` to parse HTML file containing the debate transcript.

- Load the webpage into `Python` and use `BeautifulSoup` to create a searchable version of the debate. What tags can you use to identify statements?
- Note that not all of the statements contain information about the speaker. Devise a rule to assign the unlabeled statements to speakers.
- For substantive reasons, we would like to define a single statement as any *uninterrupted* speech from a candidate. We'll say a candidate is interrupted when the transcript says that a new speaker has begun. In other words, cross talk doesn't count as an interruption. Create a list with just the text (not the tags) of each statement as an element. Some statements are split among several tags; these will need to be concatenated according to the rule you devised above. Remember to filter out notes about audience behavior

Problem 2

Now we're going to do some more preprocessing to create a dataset that includes useful information about our texts. We will use a curated dictionary list from Neal Caren. The positive words are at `http://www.unc.edu/~ncaren/haphazard/positive.txt` and the negative words are at `http://www.unc.edu/~ncaren/haphazard/negative.txt`.

- Load the positive and negative words into `python`
- Use the `porter`, `snowball` and `lancaster` stemmers from the `nltk` package to create stemmed versions of the dictionaries.
- Using the original and stemmed dictionaries, we're going to create a statement by statement data set of the speech. The data set should have the following columns:

1) Statement number (place in debate)
2) Speaker
3) Number of non-stop words spoken
4) Number of positive words
5) Number of negative words
6) Number of lancaster stemmed positive words
7) Number of lancaster stemmed negative words
8) Number of porter stemmed positive words
9) Number of porter stemmed negative words
10) Number of snowball stemmed positive words
11) Number of snowball stemmed negative words

To create the data set, create a set of nested dictionaries that map each statement in the list created in Problem 1 to the each of the attributes described above. To calculate the values for items 3 - 11 above, you'll need to do the following to each statement:

- Discard punctuation
- Remove capitalization
- Remove stop words with the list of words provided here: `'http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop'`
- Tokenize the words
- Apply each of the stemmers, determining which of the words appear in the corresponding stemmed dictionaries

Write your dataset as a csv file and save it to a working directory. Turn it in with your homework.

Problem 3

Using our new data set, let's make some observations about the debate

- Load the data into `R`
- Create a visualization that compares the overall positive and negative word rate for Obama, Romney, and Lehrer. What patterns do you notice? There is no one right answer, be creative!
- Using your data set, examine trends in each candidate's statements and Lehrer's speeches. Do you notice any

i) Trends in the measured tone?

ii) Response to the other candidate's tone (examining who spoke previously)?

iii) Overall interesting patterns? (this is an intentionally vague question)