Text as Data: Homework 1

Problem 1

We can use <p> tags to find chunks of text in the data (I do this after cutting down excess text on the front and back end in my script). However, this doesn't necessarily identify who the speaker is.

I define each speaker based on when their name appears in capital letters with a colon at the beginning of the string.

I deal with this issue by concatanating all of the text and then splitting based on a regular expression.

I then filter out the audience behavior with a regular expression that targets capital text in parenthesis.

Problem 2

See CSV file. Note that I used NLTK's stopwords instead of the non-working one.

Problem 3

The previous speaker thing wasn't helpful because Lehrer basically was in between each statement the candidates made as the moderator. (There were only about 10-14 statements where the candidates directly followed each other.) When Romney was the previous speaker, there was slightly more negativity among the other two speakers at the beginning.

I wouldn't say any of the patterns were all that interesting.

The stemmers were all about the same in pattern, though some were a little higher overall.

In plotting the rates (that is proportion of negative or positive words rather than just number per statement), it seems that Obama had two spike areas where he hit 1 for negative statements. (All words were negative). I'm guessing this is because there were few non-stop words and he said something like "no." Through most of the debate, Romney was slightly more negative in proportion than Obama, but this would make sense because Obama as incumbent would want to cast his presidency in a positive light.

There were a few cases in the number of negative words where the candidates seem to trade off. This would make sense if Obama (for example) criticised Romney and then Romney responded by criticizing Obama and vice versa. Not sure why there would be the drop in between, though. The way I plotted it there would be a lag, so not sure.

Obama said about 300 fewer substantive words than Romney. However, they had almost the exact same proportion of negative words (.071 percent and .072 percent).

Honestly, I don’t really see anything that interesting in these data from just the positive and negative words.