**PS239T Final Project: Coverage of Trump over Time**

**Bergliot Borg Christensen**

**Project overview:**

This project is broadly speaking focused on investigating the characteristics of media coverage of Donald Trump’s presidential campaign. The media have been criticized heavily for not treating his behavior, actions and words critically enough and thereby helping normalize him. In this project, I take a stab at a specific news organization – the New York Times – and attempt to quantify their coverage in terms of both most frequent words used as well as the overall tone.

**Data and analytical approach:**

I wish to tell the story of how NYT’s coverage changed, and if/when the tone and content became critical. I therefore cast a broad net and get all the coverage the words ‘Donald’, ‘Trump’ and ‘President in the period 06.16.2015-11.07.2016. NYT’s API only gives access to snippets, or short segments of articles, and these search parameters amount to app. 8200 snippets all in all. Because these snippets are so short (1-2 sentences), there is reason to doubt that they include the actual words that might indicate critical coverage such as ‘liar’, ‘demagogue’, ‘populist’, ‘assault’, etc. I therefore access full-length articles from LexisNexis to get more in-depth data on articles from the first and last month of the campaign.

I analyze these texts using different tools from textual analysis:

* *Frequency* *analysis*: I get the most frequent words used in different months of the campaign. Unfortunately, as the tables in the markdown show, these words don’t show much change in terms of new more critically laden words cropping up. Instead, all months are dominated by neutral words such as campaign, election etc. I therefore did not pursue this approach further, since there was little point in cataloging changes from some neutral words to others.
* *Sentiment analysis:* Instead, I found a package online that assesses the sentimental value of words or sentences, the Syuzhet package. This allows me to get the sentimental value across eight different sentiments – anger, fear, sadness etc. and overall positive and negative emotions. I could then get the net positivity by subtracting the negative from the positive sentiments to see which valence dominated.

**Results:**

This analysis of the snippets yielded two interesting results:

1. The overall sentiment in the coverage of Trump confirmed that coverage got more negative over time, especially the last month of his campaign.
2. The coverage of Trump got more intense over time across all sentiments – fear, sadness, joy, surprise ect. Sentiments are counted in a non-exclusionary fashion, so an article’s can score high on positive emotion as well as negative emotions. A high score across sentiments therefore means that snippets over time came to evoke more feelings and sentiments across the board, in other words a higher sentiment intensity.

Analyzing full length articles gave somewhat contradictory results in relation to snippets, mainly that coverage over the last month revealed overall positive coverage. Granted, the coverage became less positive, but it never actually became overall negative. This is surprising given my own perception of the coverage, and questions either the Syuzhet package or my use of it.

**Relevance to PS239T**

I have used the tools from PS239 to access, clean and analyze my data. Here is a short list of the most important skills I have used:

* Use Python to access NYT’s API.
* Write parser in Python to extract dates and article body from the output from LexisNexis. For full length articles, LexisNexis only exports via a single pdf containing all articles. I therefore had to write a parser so that I could get only the actual articles and their dates.
* Data-cleaning in R:
  + I used dplyr and other basic functions to subset and merge data.
* Data analysis in R:
  + I used the tm-package to construct frequencies
  + I used ggplot2 to make graphs
  + I found and used the package ‘syuzhet’ to x
  + Create functions and use them via lapply
  + Used ggplot to make graphs.

**Challenges and solutions:**

*Technical challenges:* Overall, I met the greatest amount of challenges in the data cleaning and getting my data ready to feed to the different packages I was using. Any time I ran into a problem – either extracting certain data, subsetting in specific ways or merging dataframes – I googled my issue and generally had success tweaking and applying the different solutions found online. Specific examples of time-consuming challenges I ran into involve the ‘syuzhet’-package, which only accepted objects that were in the form of character vectors, and I had to fiddle with the ‘as.XXX’ option in R to get it right. Another challenge was merging the different objects, fx when I combined extracted full-length articles with the separately extracted date for the article – R had trouble reading the date as an actual date, and this required significant fiddling to fix and use the ‘as-date’ function.

*Conceptual/analytical challenges:* My initial research design only involved analyzing word frequencies and how these changed over time. Upon getting the most frequent words for every month, it was clear that an analysis like this was not possible with my over-time data based on snippets. There was simply not enough variation in the most frequent words to warrant in-depth analysis, and the most frequent words were furthermore rather neutral in character – words like ‘elect’, president’ etc rather than ‘assualt’ etc. I considered removing these neutral and very frequent words, but since I was interested in seeing if the coverage changed over time from neutral to more negative words, manually removing all neutral words might bias my conclusions.

**Next steps:**

This analysis only represents taking the first steps into understanding the coverage of Trump. Ideally, full length articles covering the entire period would be interesting to analyze, and I would also like to move forward with a more in-depth approach that tries to get at the existence and nature of criticism in the coverage of Trump. With another more detailed analysis that removes some neutral words that confound the frequency counts, e.g. ‘said, ‘time’, ‘voter’ etc, it might be possible to see when critical words appeared. But, this requires finding which words to keep, which to remove – and for this project, I chose to focus on getting techninal training and experience in working with R and Python, as well using Stackexchange constructively.