Computational Methods for American Politics (40500)
Autumn 2019

Problem Set #3: Text Mining

Remember to submit a ***single*** *rendered PDF (either from .Rmd or a Jupyter Notebook) via GitHub* ***by Wednesday, November 27 at 5 pm****:* https://github.com/macss-cmap19

For this final problem set, suppose you are an undecided American voter. You are interested in exploring both major political parties in America and how they brand themselves and things they choose to focus on. As such, you get the most recent formalization of these things you can find, which are the 2016 party platforms for each of the two major parties (also called “manifestos” in other countries).

You decide to employ your computational skills to get explore that which these parties have to offer. At the end of the analysis, you must make a decision. I.e., answer the following question: Based only on your analysis here, NOT your current political biases, which party would you support in 2020?

Note: In this assignment, I am asking you to lay aside any feelings toward either party or political actor in America, whether good or bad. Rather, approach the question as unbiased as possible and try hard to let your results inform your response. Regardless of your current political proclivities, I promise I won’t (care or) broadcast what your hypothetical response is in this problem set; its just a fun exercise in “data-drive decision making”. Good luck!

GENERAL NLP/PREPROCESSING

1. Load the platforms.csv file containing the 2016 Democratic and Republican party platforms. Note the 2X2 format, where each row is a document, with the party recorded as a separate feature. Also, load the individual party .txt files as a corpus.

2. Create a document-term matrix and preprocess the platforms by the following criteria (at a minimum):
   a. Convert to lowercase
   b. Remove the stopwords
   c. Remove the numbers
   d. Remove all punctuation

3. Visually inspect your cleaned documents by creating a wordcloud for each major party’s platform. Based on this naive visualization, offer a few-sentence-description of general patterns you see (e.g., What are commonly used words? What are less commonly used words? Can you get a sense of differences between the parties at this early stage?

SENTIMENT ANALYSIS

4. Use the “Bing” and “AFINN” dictionaries to calculate the sentiment of each cleaned party platform. Present the results however you’d like (e.g., visually and/or numerically).

5. Compare and discuss the sentiments of these platforms: which party tends to be more optimistic about the future? Does this comport with your perceptions of the parties?

TOPIC MODELS

6. With a general sense of sentiments of the party platforms (i.e., the tones related to how parties talk about their roles in the political future), now explore the topics they are highlighting in their platforms. This will give a sense of the key policy areas they’re most interested in. Fit a topic model for each of the major parties (i.e. two topic models) using the latent Dirichlet allocation algorithm, initialized at k = 5 topics as a start. Present the results however you’d like (e.g., visually and/or numerically).

7. Describe the general trends in topics that emerge from this stage. Are the parties focusing on similar or different topics, generally?

8. Fit 6 more topic models at the follow levels of k for each party: 5, 10, 25. Present the results however you’d like (e.g., visually and/or numerically).

9. Calculate the perplexity of each model iteration and describe which technically fits best.

10. Building on the previous question, display a barplot of the k = 10 model for each party, and offer some general inferences as to the main trends that emerge. Are there similar themes between the parties? Do you think k = 10 likely picks up differences more efficiently? Why or why not?

CONCLUSION

11. Per the opening question, based on your analyses (including exploring party brands, general tones/sentiments, political outlook, and policy priorities), which party would you support in the 2020 election (again, this is hypothetical)?

Bing (binary): https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

AFINN (-5 to 5): http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010