**Text As Data**

**Final Paper Prospectus**

Laura Buchanan - lcb402

Patrick Kraft - pk1622

For our final project, we propose using topic model methods on American National Election Survey (ANES) data to uncover the degree of political knowledge of the 5,000(?) respondents. Political knowledge is widely used when studying political behavior or public opinion. However, measurement of political knowledge based on answers to political trivia may misclassify respondents who cannot recall these facts, but do indeed have a coherent network of political ideas.

The ANES data has several useful features. Our project will focus on the data made up of open-ended survey responses. For this portion of the survey, respondents write freely about what they like and dislike about the republican and democratic parties. We suspect that the number of topics a respondent mentions in their answers, or the detail with which they speak about the topics they mention, will covary with the other political knowledge measures available in the data.

All responses from a single participant will be considered a single document. First, we will find the topics present over the entire corpus, starting with a simple off-the-shelf baseline model, and improve performance as we learn the “knobs” of topic models as they are covered in class. Then, for each document, we will count the number of topics present. This will be our first measure. Second, we will look at the length of the document. Third, we will look at document length relative to the number of topics. We suspect that the first measure will correlate with political knowledge, displaying a wide spectrum of political insight. The second measure will act like something of a control or baseline; can we simply predict political knowledge by the length of response? Finally, the third measure may help us gain insight into potential cases where a respondent focuses on very few topics, but has a wealth of insight into the topics they choose to mention.

The dataset is available from American National Election Studies. For his PhD, Patrick has cleaned this data to some degree, using a dictionary method to correct for misspellings. The dataset may be further cleaned in two different ways. First, we may account for “eloquence” of the respondents, using readability or type-to-token ratio of the responses. We would like to capture political intelligence specifically, rather than general intelligence. Similarly, we may correct for education level, which is also available in the data. Second, we may create a list of “talking points” that could possibly turn up in the responses. We suspect that specific phrasing could suggest that a respondent is parroting a political message, rather than comprehending and reframing the message in their own words. This may be too difficult a task to complete for this project, but a good first pass try would include finding common n-grams, up to n = 6, and use a web search to determine whether these are known slogans. Respondents would then be “punished” if they primarily use these slogans, as that suggests that they do not have a deep political knowledge.

To assess the performance of our model, we will compare our results to the other political knowledge scores, perhaps using a method such as the Adjusted Rand Index. An evaluation such as this will be useful, because we can check, first, if our method is performing better than chance, and, if so, is it performing poorly, well, or excellently. However, we will also take a qualitative look at the performance our model, to determine if we are able to find cases of political knowledge not captured in the standard measure of political knowledge.