IPPP Final Project

Julia Karon and Adriana Rakshana

Proposal Question

We want to explore the relationship between a candidate's online presence (here represented by Google Trend data) and the actual percentage of the vote received by each candidate. Initially we were hoping to use social media data from either Facebook or Reddit but it was a bit difficult to obtain a coherent usable data set from either so we settled on using the Google trends data. We are focusing our analysis on Congressional races since the data from presidential races could potentially be misleading (Trump had a massive online presence but received less than half of the popular vote). We would like to analyze multiple election cycles to see if an increased online presence over time has led to an increase in the vote for any given candidate. While this data may be misleading for incumbents since they have a massive political advantage regardless of online presence, this analysis could prove informative when it comes to candidates running multiple times in the same district who have to consider how much increasing their online presence will realistically increase their odds of being elected.

Datasets

Google Trends - Google Trends shows how often a particular search-term is entered relative to the total search-volume across various variables, including region, time and related queries. We hope to input specific key terms/topics that relate to each candidate in order to find who was searched more, i.e. more popular by region, in this case 'by metro'. For example, we would input the terms Franken, Al Franken, Minnesota Senators, or use the Topic feature that automatically includes all related queries to the topic "Senator Al Franken" and obtain the trends csv that includes the data with the variables of time and search frequency as an index of 100. Given the multiple data files that would be needed, we hope to use the python feature of pytrends that will help automate the download of data when we list the required key terms, regions and time period. Using code written by GitHub members, we can directly input key terms and immediately get a Pandas dataframe with the relative frequency of words over time and over a certain region. We then hope to compare the frequency of a given candidate with the percentage of the vote obtained from the FEC datasets.

https://trends.google.com/trends/
https://github.com/GeneralMills/pytrends

Congressional Results - The Federal Election Commission (FEC) has published election results from every federal election up to 2014 which detail the percentage of the vote received by each candidate in any given congressional race. We would have to use a different dataset for each election cycle but the datasets themselves are a bit more manageable.

https://transition.fec.gov/pubrec/electionresults.shtml