Text Analysis of U.S. Congressional Records

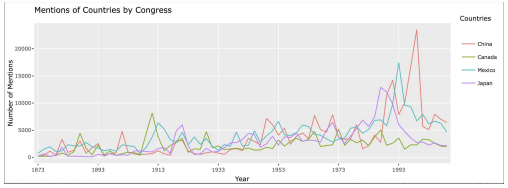
Adam Behnke, James Mahler, Parvathi Meyyappan, Youna Song
Advisors: Dani Nedal, Zach Branson

Introduction

In order to understand U.S. political action over the years, it may be important to analyze the countries that drive U.S. foreign policy during certain time periods. We perform text analysis to determine what countries are being talked about most often on the floor of the U.S. Congresses during crucial time periods. We also explore the context surrounding the mentions of different countries to examine the types of words or language used when describing specific countries. Finally, we include variables such as political party affiliation to analyze, for example, how Democrats and Republicans talk about different countries in different ways.

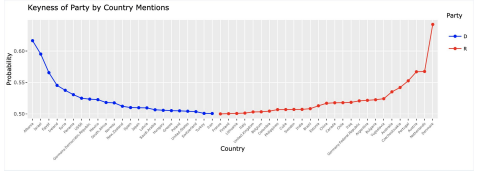
EDA

- We included a visualization tool to analyze and compare how often nations are discussed in congress, for 43rd to 111th congresses
- The tool incorporates features that allows users to choose the count types, show the proportions of counts, show trend lines, or filter the counts by party



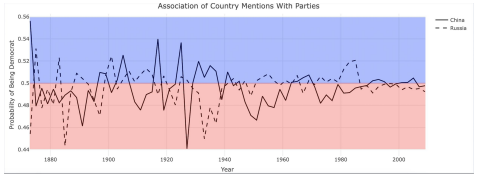
Keyness

To explore if Democrats and Republicans talk about different foreign nations, we developed a logistic regression model to assess the likelihood that a speaker was a Democrat or Republican given that they mentioned a particular country. We are able to use this model to assess the relationship between the mentions of each country and representing a particular political party for each Congress. The following plot illustrates this relationship for the 80th congress (1947-1949). Mentioning Albania or Israel in Congress during these years was a fairly strong indicator of being Democratic, while mentioning Denmark was a strong indicator of being Republican.



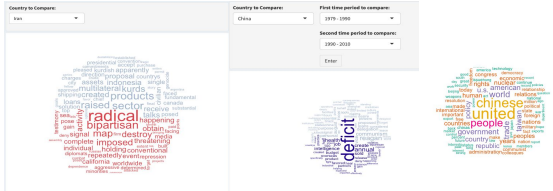
Data

The data is provided by the United States Congressional Record (available by HeinOnline). It contains speeches from the 43rd to 111th Congresses spanning the years 1873 to 2010. Each congress is in session for two years. We analyze all text spoken on the floor of each chamber of Congress and we also consider information related to the speakers' party affiliation, age, and gender, among others. We worked with over 30GB of this text data and cleaned/parsed it to extract relevant information.



Word Clouds

Some of the main visualizations on the R Shiny app are word clouds that help the user better understand the language used when speaking about certain countries. We used contextual words surrounding every country mention and created frequency tables to use as the basis of the word clouds. It's broken down into 3 different explorations. One is where the user can see aggregated contextual words spoken in congress over 1945 - 2010 in general. The next visualization lets the user compare contextual words and phrases that are unique to each party and similar between both. The third visualization is also a difference and similarity word clouds but letting the user choose 2 different time periods for various countries to compare.



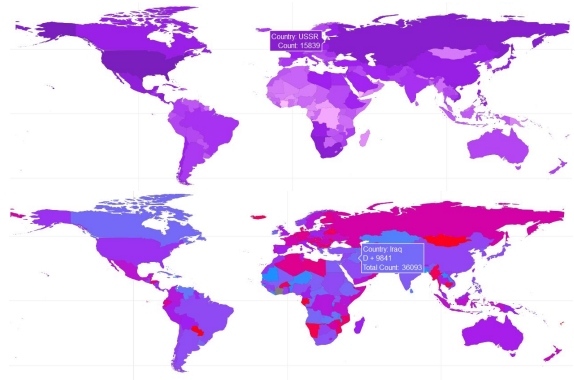
Left: Word cloud shows the political party comparison of contextual words involving the country Iran. It's important to note that these counts are over 1945 to 2010 so amass quite a bit of time.
Middle: Word cloud of the most frequent distinct contextual words for China between the time periods of 1979 to 1990 and 1991 to 2010.
Right: Word cloud of the most frequent similar words for China between the time periods of 1979 to 1990 and 1991 to 2010.

References:

Data: https://data.stanford.edu/congress_text
Senate Years: https://www.senate.gov/legislative/YearstoCongress.htm

Interactive World Heatmaps

One of the main tools included in our R Shiny app is an interactive heat map that shows the number of times each country was mentioned in a particular time period (separated by each Congress). This allows users to get a quick look at which countries are being talked about most during a particular time period and could lead researchers to study why a particular country is or isn't being talked about or why one political party is talking about certain countries more often than the other. Below are two example heatmaps.



Top: Heatmap for all country mentions in 1985-1987
Bottom: Differences in mentions between the Republican and Democratic parties in 2003-2005.

We see that the USSR is being mentioned frequently in 1985-1987, which makes sense as the Cold War was taking place. In 2003-2005, we see Democrats mentioning Iraq much more frequently than Republicans.

Conclusion and Future Work

Our findings can provide for further research in understanding U.S. foreign policy. Using our application, users can trace the trends of country mentions and explore the context surrounding different countries during specific time periods. Limitations of our research included the difficulty in extracting information from extremely large volumes of noisy text which could have led to inaccurate country counts, despite many efforts to account for the noise.

One particularly interesting application could be to trace the trends in country mentions to help explain and predict patterns in US political action. In addition, more advanced sentiment analysis could be done using NLP techniques to better understand the text or topic modeling could be done for further analysis.