Understanding US-China Relations: Text Analysis on Congressional Speeches

Sylvia Ding, Daniel Liang, Angela Zheng
Advised by: Zach Branson, Dani Nedel

Introduction

- Our research takes a look at US-China relations by exploring United States congressional speeches from the 1900s
- Research Goal: Identify and distinguish between speeches about the People's Republic of China (Mainland China) and the Republic of China (Taiwan), and conduct further analyses over time and across political parties.

Methods

- Data: Congressrecord.com: transcripts of 770,123 US Senate speeches from 1945 to 1990
- Worked on random sample of 10,000 speeches
- Two additional variables explored in conjunction:
  1. Political affiliation of the speaker
  2. Year that the speech is made

EDA

- Comparison wordcloud showing most frequently used words from two datasets.
- A selection of Senate speeches excerpted from the internet that were manually annotated to be about the People's Republic of China. Bottom: a random sample of Senate speeches.
- Speeches after 1995 can be found online, manually indexed by topic.
- One goal: look at words in manually indexed "China" speeches to create training data for a machine learning topic classification tool.
- Another method: use dictionary of terms relating to China, for example "Far East" or "Freedom" and search for occurrences of such words and the words around it, as context words.

Topic Modeling & Time Trends

- Topic modeling is an unsupervised statistical model for discovering the abstract "topics" that occur in a collection of documents.
- The topics are essentially clusters of similar words with assigned frequencies unique to each topic.
- Speeches are grouped depending on the relative statistics of words in each. The results are also non-deterministic as topic modeling produces distributions of topics in each document.
- We performed topic modeling on 40 topics, looked into the frequent words in each, and found several topics that are likely discussing international relations.
- We also performed topic modeling on context words of vocabulary words from the naive dictionaries to reveal speech trends.

Figure 1. Comparison wordcloud showing most frequently used words from two datasets.

Figure 2 & 3. Comparison wordcloud showing most frequently used words from each party.

Figure 6. Distribution of log gamma values for different topics faceted by group

- * Our "Red", "Blue", and "Mixed" categories speeches based on the total mentions of words directly associated with China in the naive dictionary.
- * By inspection of the top words shown in Figure 6, topics colored in green are most likely international topics that may discuss China related things.
- * Not surprisingly, these topics correspond to higher log gamma values for speeches mentioning China more. However, not all topics with high log gamma scores for the "Red", "Blue", or "Mixed" groups are discussing international things. They may be false positives even though they mention China.

Figure 7. Rolling Average of China and Taiwan Counts for Parties across Time

- * According to the time trend overall, Democrats are more likely to mention China and Taiwan.
- * The apparent spikes in the graph are 1978, 1984, 1988, 1989, and 1993 for Democrats and 1950-1956 and 1974 for Republicans. These spikes are also significant as they don’t correspond to spikes in the count of total sampled speeches in different years.

Figure 8. Top Words for Selected Topics

- * As we can see above, these topics with clusters of words and corresponding log values seem to be talking about international issues, with indicative words like "foreign", "world", "communist", "league", "terrorism", and so on.

Figure 9. Rolling Average of Gamma Values for International Topics across Time

- * Gamma values indicate how likely speeches’ content is related to specific topics.
- * Note: values are similar for democratic and republican parties, only with some obvious differences for topics related to Soviet and communist nations after 1980, which is more likely discussed by democrats.

Context Words

- Sen. Bill Nelson of Florida: There is at this [moment in orbit the first Chinese astronaut. Their successful launch of] a piloted spacecraft marks the beginning of a new chapter in the history of human exploration of space.

Figure 10. Example of Context Words

- Context words: words that appear directly before and after a keyword of interest. Shown above, the words in red are the 5 context words before and after the keyword of the Chinese "astronaut".

Figure 11. Context Word Topic Modeling over 6 topics

- Context words here are the 50 words before and after the words in China/Taiwan Dictionary used as anchor words in the selected speeches.

Conclusion

Our research hopefully builds the groundwork for further classification with other supervised machine learning algorithms in the following ways:

1. We filtered out a potential set of speeches likely related to China and Taiwan.
2. We identified a useful set of congressional speeches related to China after 1995 that can be used as training data.
3. We obtained a useful set of words highly correlated with China which can be the material for implementing a bag of words approach in classification.