Understanding US-China Relations: Text Analysis on Congressional Speeches

Sylvia Ding, Daniel Liang, Angela Zheng

Advised by: Zach Branson, Dani Nedal

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.

Data

- Dataset drawn from the Congressional Record: 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

Methods

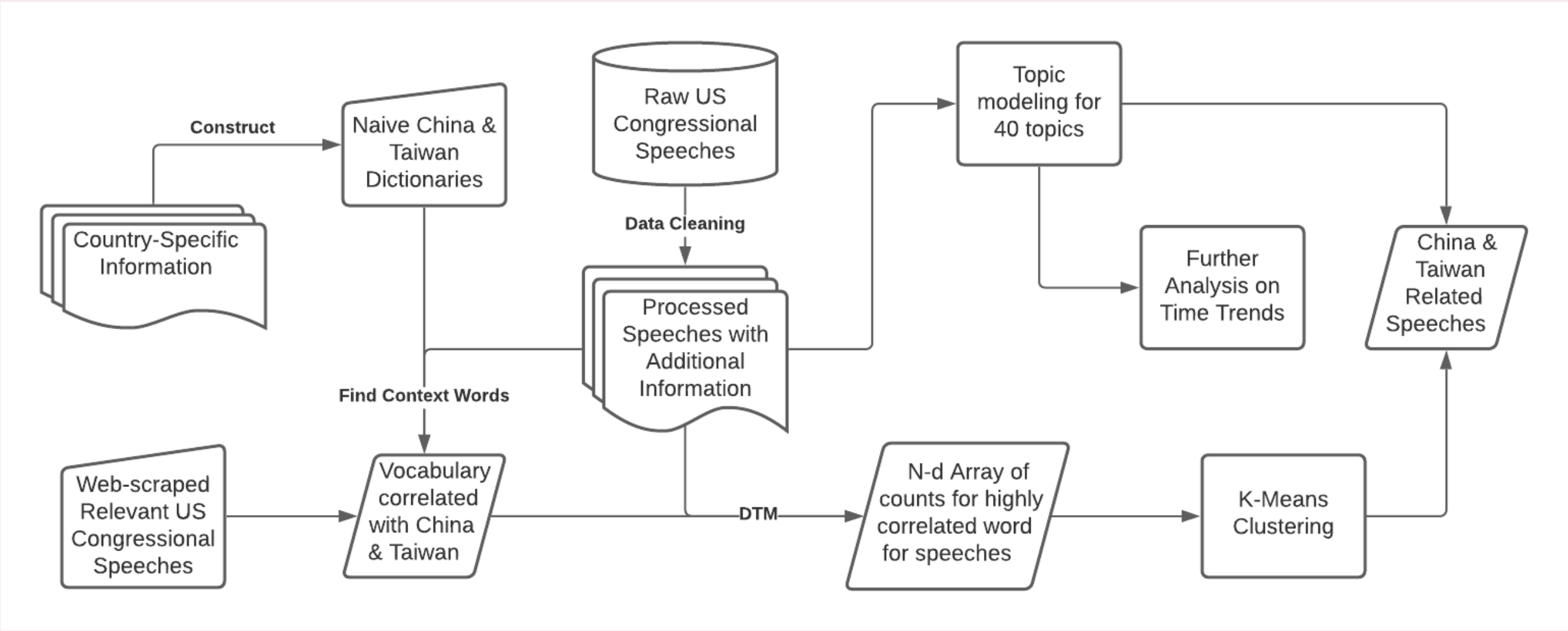
Construct
Naive China & Taiwan Dictionaries
Raw US Congressional Speeches
Topic modeling for 40 topics
Country-Specific Information
Data Cleaning
Further Analysis on Time Trends
China & Taiwan Related Speeches
Processed Speeches with Additional Information
Find Context Words
Web-scraped Relevant US Congressional Speeches
Vocabulary correlated with China & Taiwan
DTM
N-d Array of counts for highly correlated word for speeches
K-Means Clustering

Figure 4. Flow Chart Analysis of Research Approach

- Country-specific information is used to construct our naive dictionaries for China and Taiwan.
- Dictionary words are used as anchor words to find context words in the sample of processed speeches.
- Web-scraped US congressional speeches already labeled as China related after 1995 also provides unique terms associated with China.
- N-d array of counts for the vocabulary words re-represents each speech for classification with unsupervised k-means clustering.

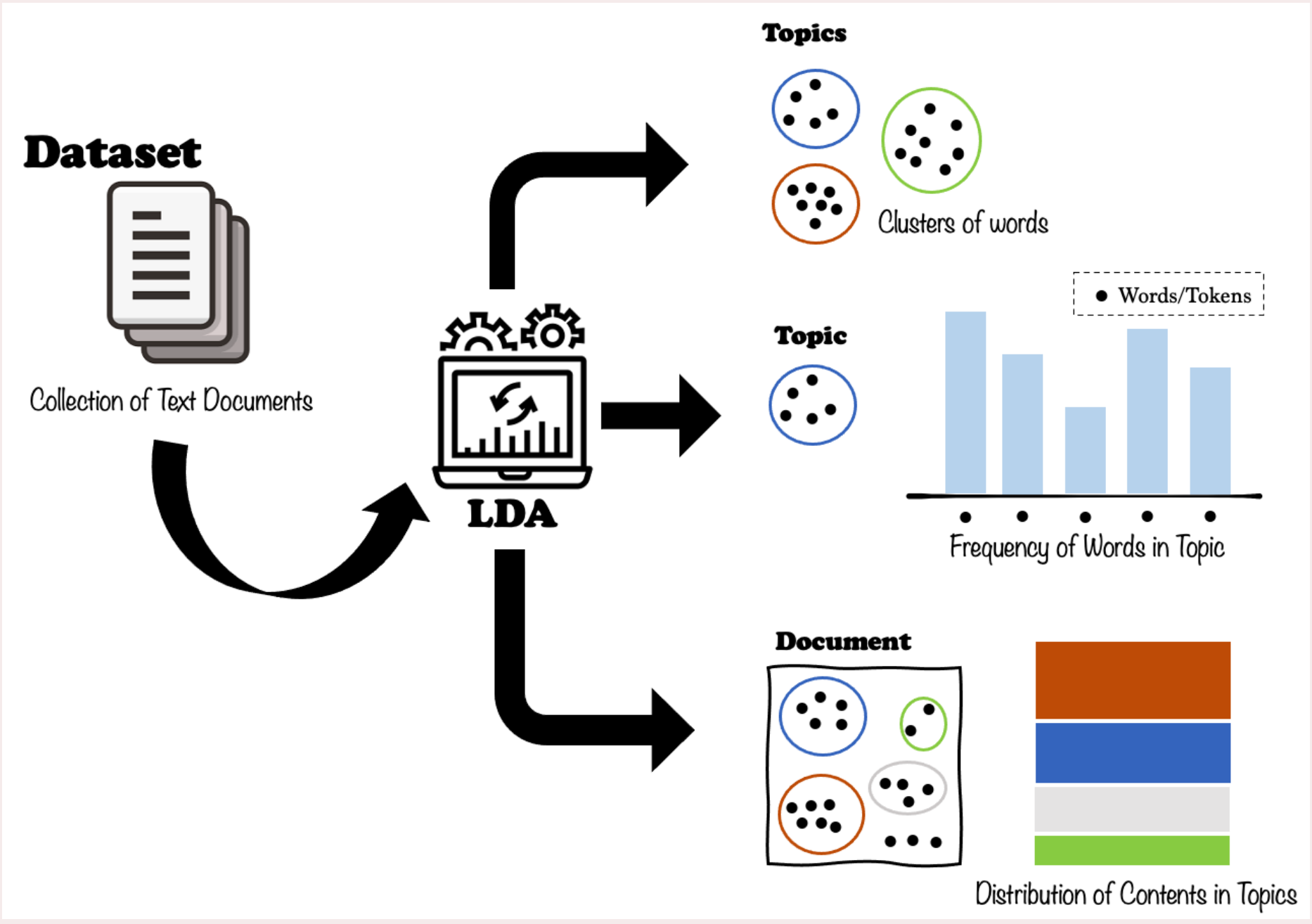
Dataset
Collection of Text Documents
Topics
Clusters of words
Topic
Words/Tokens
Frequency of Words in Topic
LDA
Document
Distribution of Contents in Topics

Figure 5. Topic Modeling Process

- 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 related things.
- We also performed topic modeling on context words of vocabulary words from the naive dictionaries to reveal speech trends.

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 this occurrence of the keyword "Chinese".

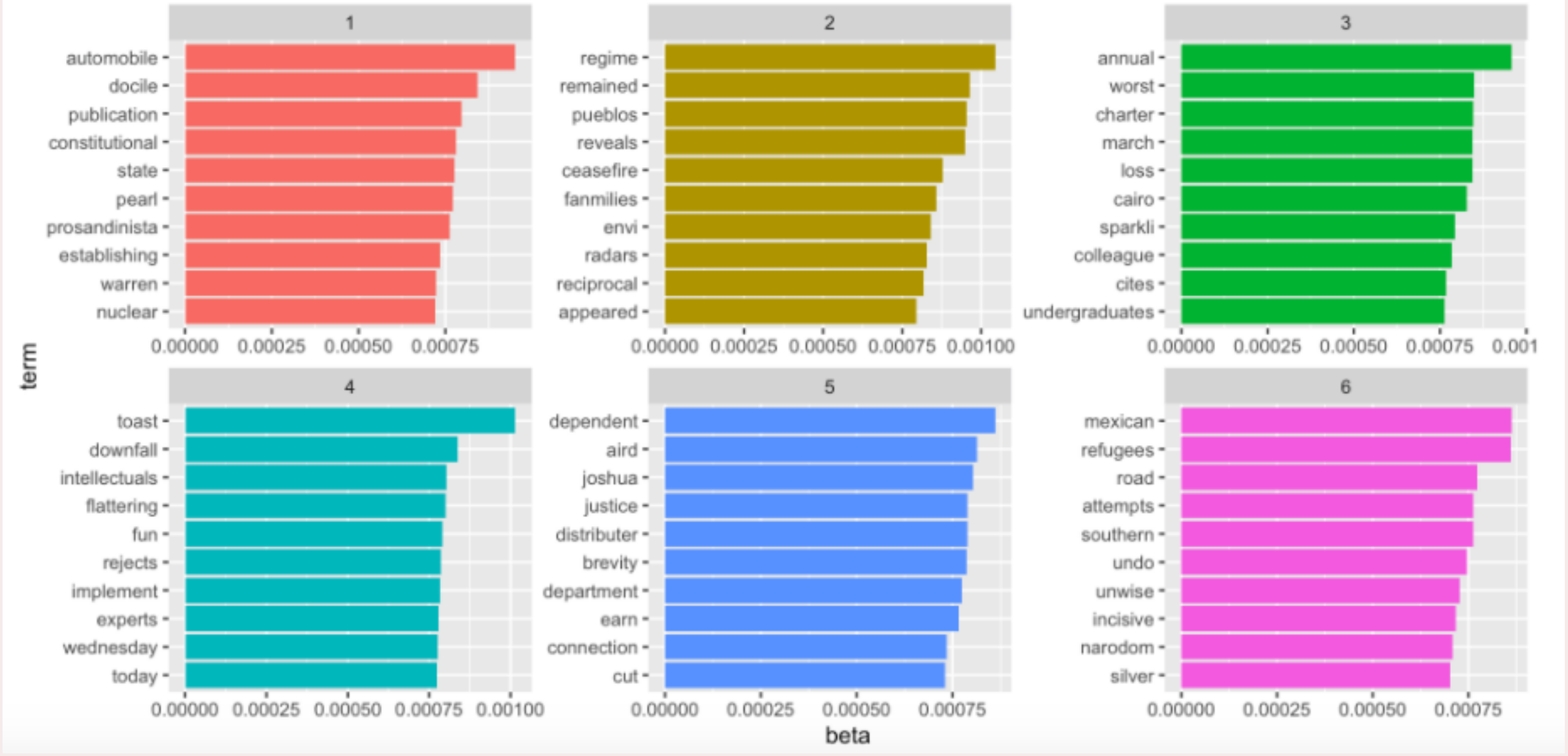

Figure 11. Context Word Topic Modeling over 6 topics

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

- 50 context words before and after the words in our China/Taiwan Dicionary are collected using built algorithms.
- Topic modeling is applied to a corpus of context words over 6 topics to investigate whether specific topics related to China can be found.
- Most of the topics seem to be irrelevant to China. Potenial reasons may be data noises and false positives.

EDA

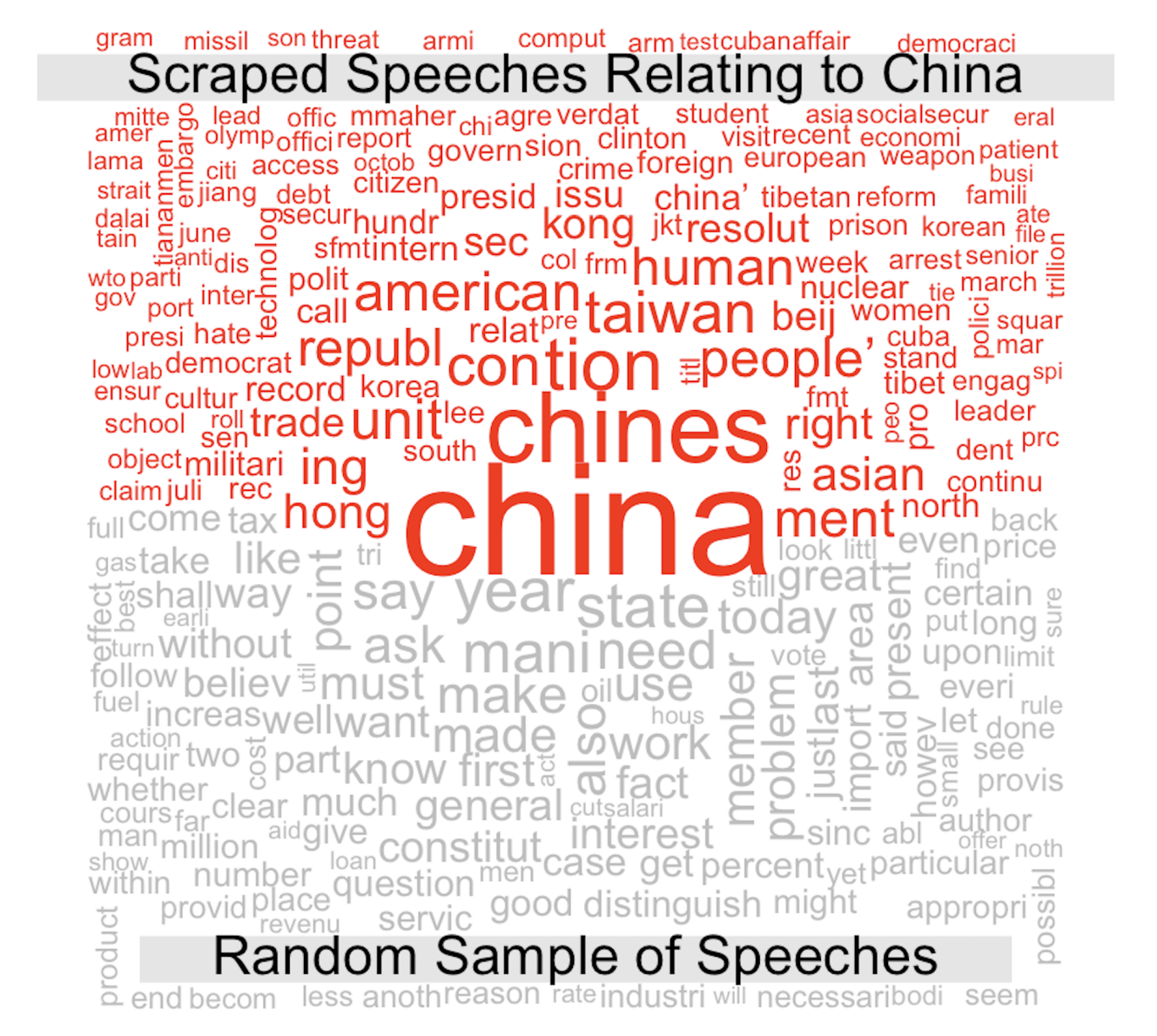

Figure 1. Comparison wordcloud showing most frequently used words from two datasets.

Top: a sample of Senate speeches scraped from the internet that were manually ensured to be about the People's Republic of China. Bottom: a random sample of Senate speeches.

- Senate speeches after 1995 can be found online, manually indexed by topic
- Speeches in our data set (all before 1995) are not indexed
- 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 "Beijing" or "Mao", and search for occurrences of such words and the words around it, AKA context words

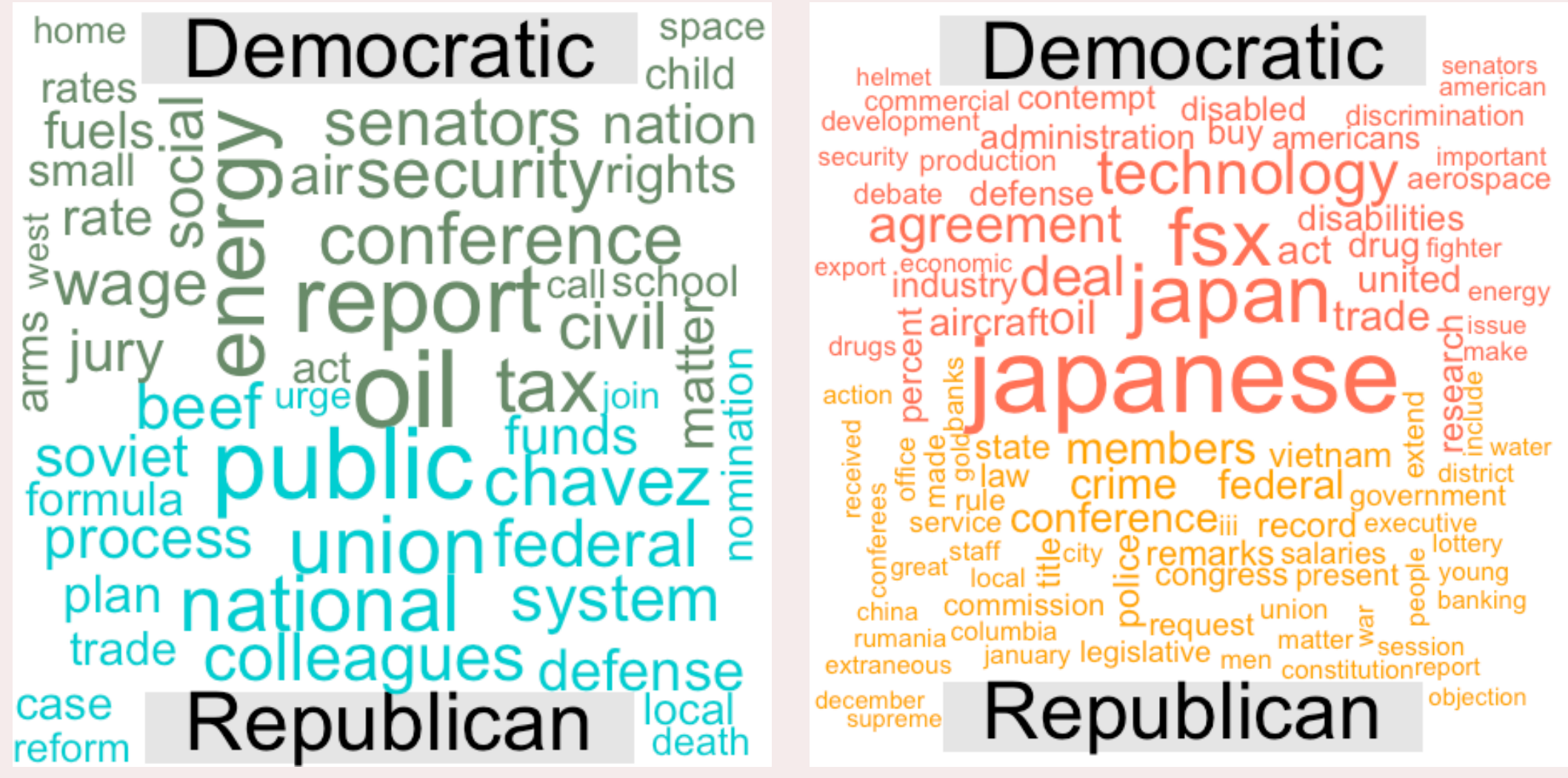

Figure 2 & 3. Comparison wordcloud showing most frequently used words from each Party.

Left: comparison word cloud for speeches without words in our naive dictionaries. Right: comparison word cloud on the speeches with China/Taiwan related words according to our naive dictionary.

Topic Modeling & Time Trends

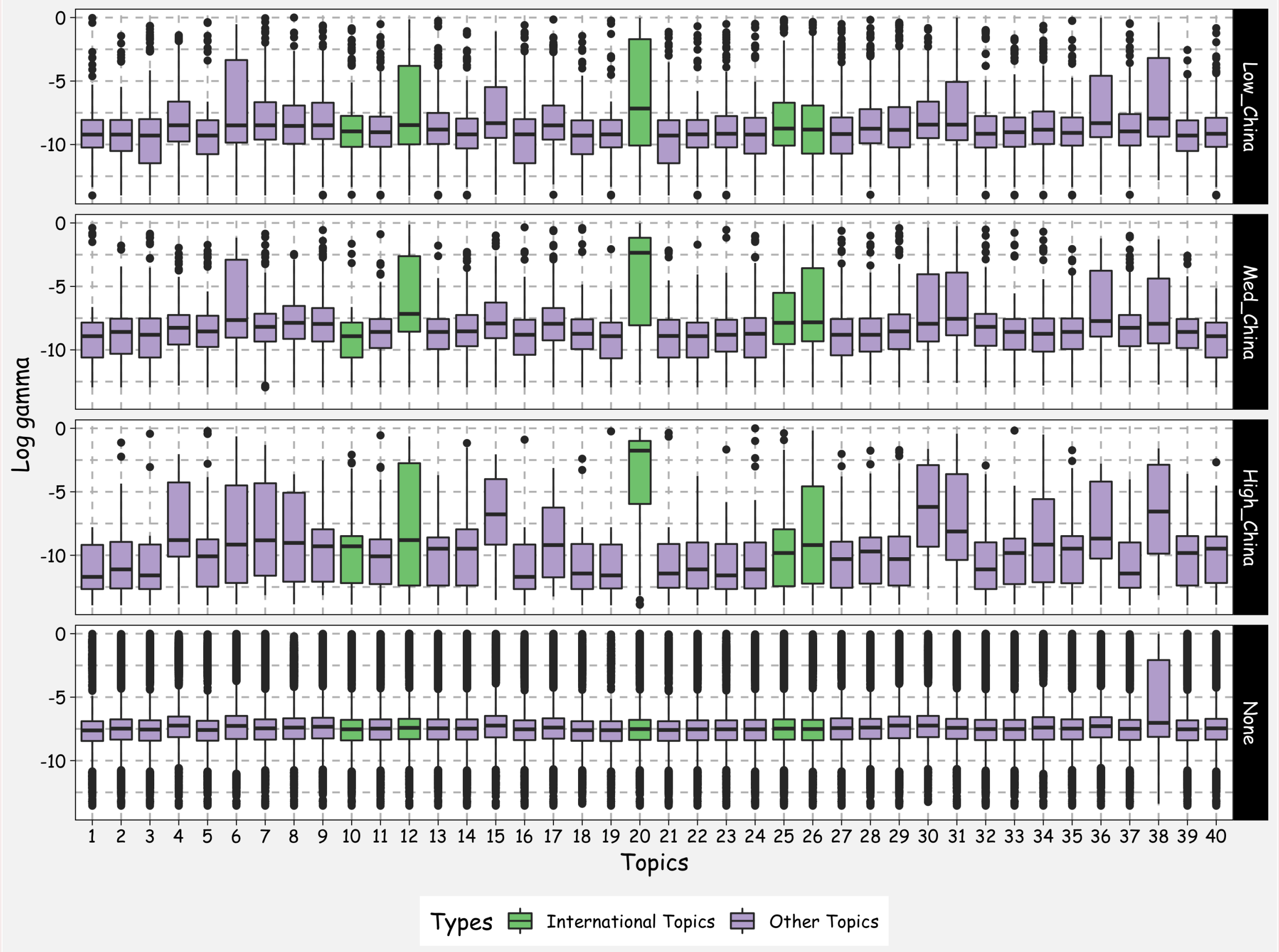

Figure 6. Distribution of log gamma values for different topics faceted by group

- "Low", "Med", "High", and "None" categorizes speeches based on the total mentions of words directly associated with China in the naive dictionary.
- By inspection of the top words shown on Figure 8, 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 values for the "Low", "Med" or "High" groups are discussing international things. They may be false positives even though they mention China.

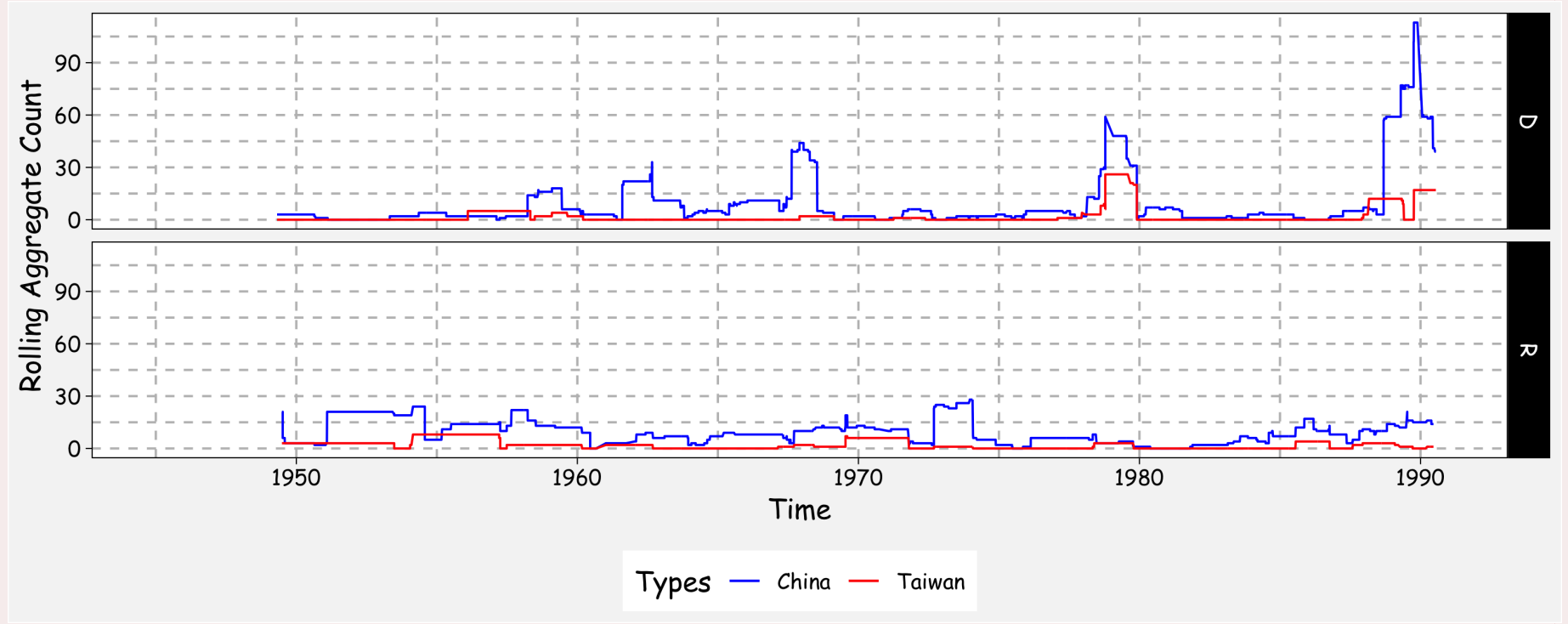

Figure 7. Rolling Average of China and Taiwan Counts for Parties across Time

- According to the time trend overall, Democrats are more likely to mention China and Taiwan.
- The apparent spikes in the graph are 1959, 1962, 1968, 1979, and 1990 for Democrats and 1950-1960 and 1974 for Republicans. These spikes are also significant as they don't correspond to spikes in the counts of total sampled speeches in different years.

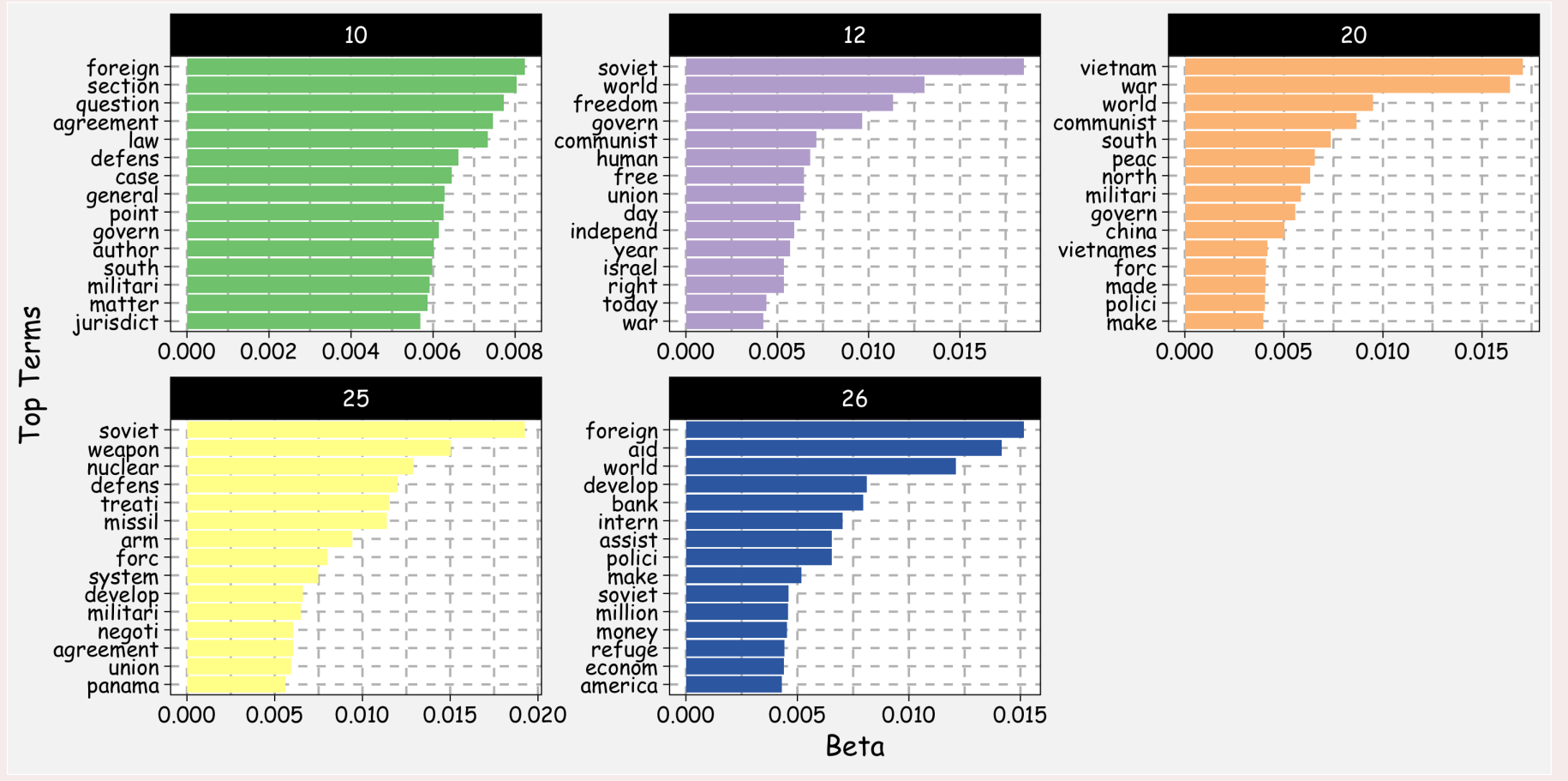

Figure 8. Top Words for Selected Topics

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

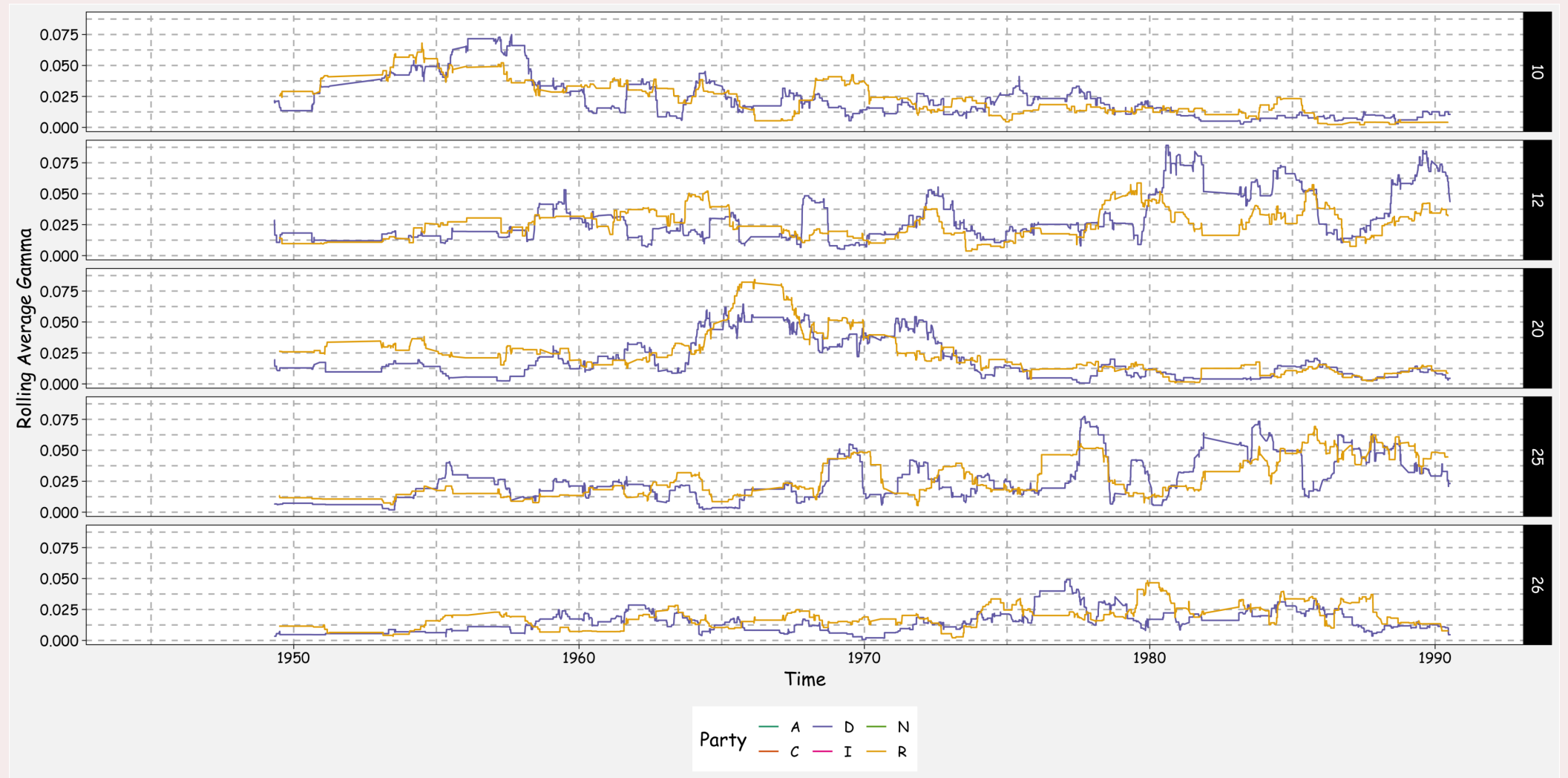

Figure 9. Rolling Average of Gamma Values for International Topics across Time

- Gamma values indicate how likely speeches' content are related to specific topics.
- Most trends are similar for democrat and republican parties, only with some obvious differences for topic 12 (related to Soviet and communism) after 1980, which is more likely discussed by democrats.

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.