Research Work Summary: Congressional Records from 104th - 118th Congress

Tianyi Zhang
Dr. Guldi
QTM 499
December 9th, 2023

Introduction

In the ever-evolving landscape of political data science, the Congressional Records project spanning from the 104th to the 118th Congress stands as a monumental endeavor. Covering the years 1995 to 2023, this project not only required sophisticated data retrieval and processing techniques but also demanded a nuanced understanding of political dynamics and legislative processes.

My involvement in this project was diverse and integral, employing advanced skills in web scraping, data management, and text mining. This work allowed me to explore the intricacies of political discourse, providing valuable insights into nearly three decades of legislative evolution and decision-making. This project was not just about data collection; it was an opportunity to bridge the gap between vast data resources and meaningful political analysis, contributing significantly to the field of political data science.

Project Overview

The Congressional Records project, targeting records from the 104th to the 118th Congress, was a groundbreaking initiative in political data analysis. Central to this project was the retrieval of data from the Congress Data API, chosen for its up-to-date and comprehensive legislative information. This choice was strategic, as the Congress Data API offered more recent and detailed data compared to alternative sources like the Stanford dataset. The project's scope involved an intricate process of scraping, processing, and analyzing data to decipher legislative trends and shifts in political rhetoric from 1995 to 2023.

Our work began with the systematic retrieval of Congressional records from the Congress.gov website. This process was not just about collecting data; it was about transforming these extensive records into a manageable and analyzable format. The transformation involved converting vast amounts of data from PDFs into structured text files, organizing them meticulously to preserve their integrity and context. This step was crucial in maintaining the accuracy and reliability of the information.

The heart of the project lay in the exploratory data analysis phase. Here, we explore data to uncover hidden patterns and trends in political communication. This analysis was pivotal in understanding how political discourse has evolved over nearly three decades, providing insights into the legislative process, policy priorities, and rhetorical strategies of different political eras.

By leveraging the power of data science, this project not only contributed significantly to the understanding of political history but also demonstrated the immense potential of using advanced data analysis techniques in political research. The comprehensive nature of our approach, from data retrieval to analysis, ensured that the results were not just informative but also actionable, offering new perspectives into the dynamics of legislative decision-making.

Tasks and Execution

1) **Data Retrieval:** The data retrieval process for this project was notably complex due to its multi-layered nature. Initially, it involved extracting URLs from the Congress Data API, a step that required careful navigation of the API's structure. Following this, the process entailed

downloading PDFs from these URLs, hosted on Congress.gov. This step was challenging due to server constraints and the differences between the API and the main website. Extensive testing and optimization were essential to bypass server limitations and prevent disruptions to the data retrieval program. The need to efficiently manage these distinct steps underlined the project's technical complexity and demanded a high level of precision and adaptability in script development and execution.

2) **GitHub Repository and Documentation Expansion:** The GitHub repository established for this project was a fundamental component in its execution and dissemination. Far from being a mere storage location, it served as a dynamic platform for collaboration and knowledge exchange. The repository's structure was deliberately designed for clarity and ease of navigation, with distinct sections for scripts, datasets, and documentation. The documentation was comprehensive and meticulous, encapsulating not only the project's objectives and scope but also providing in-depth descriptions of each script's functionality. Step-by-step guides were included to facilitate ease of use, ensuring that users from various backgrounds, whether they be data scientists, political analysts, or hobbyists, could effectively engage with the project. This inclusive approach to documentation was instrumental in making the project accessible to a wide audience. In addition to the overall repository documentation, specific resources like the 'data_exploration.ipynb' notebook on my GitHub page play a crucial role in illustrating the practical application of the project's methodologies. This particular notebook is a prime example of the project's commitment to educational clarity, demonstrating the process of text mining and data analysis in a user-friendly, interactive format. By showcasing real-world applications of the scripts and techniques developed for the project, this resource serves as both a learning tool and a practical guide, further enhancing the project's value to the community.
3) **Data Testing and Text Mining Tutorial Expansion:** Rigorous testing was vital to ascertain the integrity and accuracy of the data. This process involved thorough validation against various criteria and external sources. The text mining tutorial developed as part of this project was extensive, covering from basic data cleaning to advanced natural language processing. This tutorial was crafted to be an educational resource, enhancing users' analytical skills. Continuing from this point, the retrieval of data marked only the beginning of an iterative process of data refinement. It was essential to repeatedly review and fine-tune the data, applying insights gained from initial analyses to improve data quality and relevance. This iterative approach was crucial in addressing issues such as generalizability and contamination. For instance, identifying and mitigating biases in the data collection process or addressing discrepancies in text conversion was pivotal in ensuring that the final dataset truly reflected the Congressional records without distortions. This careful attention to data integrity not only enhanced the reliability of the dataset but also ensured its applicability across a range of political and linguistic analyses, thereby significantly contributing to the robustness and versatility of the research.

Interpretations and Insights into the Datasets

1) **Custom Stopwords:**

In the process of text analysis, one critical step is the refinement of stop words - words that are filtered out before or after processing text. Despite utilizing the Natural Language Toolkit (NLTK)'s predefined list of stop words, it was observed that the list did not capture all non-informative words within the context of our dataset - the Congressional Records. This

necessitated the creation of a custom stop words list, a task that underscores the subjectivity inherent in textual analysis.

Defining Stop Words Standard stop words lists, such as the one offered by NLTK, typically remove high-frequency words that add little semantic value to the analysis. However, these lists may not be fully aligned with the unique context and content of specialized datasets like legislative records. Therefore, a more tailored approach is required to address the specific lexicon used in congressional discourse.

Our approach to customizing the stop words list entailed a thorough review of the dataset to identify words that, despite their frequency, did not contribute meaningful insights. This involved: Parsing the dataset to determine the frequency of each word. Manually scrutinizing the list of frequent words to identify those that were contextually irrelevant or generic. Reflecting on the specific analytical goals of our research to decide which words were unlikely to yield valuable information. This led to the development of a customized stop words list, which included terms such as 'mr', 'state', 'year', and 'new', among others that are prevalent in legislative language but may not be indicative of the content's analytical value.

Determining what constitutes a stop word is an iterative and subjective process, which must be balanced against the research objectives. Researchers must employ their understanding of the dataset and domain expertise to refine this list. As the analysis progresses and more is learned about the data, it is common practice to revisit and adjust the custom stop words list accordingly.

2) Word Cloud

The word cloud is a visual representation of the most frequently used words within a collection of Congressional Records from the 104th to the 118th Congress. In a word cloud, the size of each word is proportional to its frequency of occurrence in the text.

- **Prominent Themes:** The larger words like "science," "public," "health," "record," and "budget" indicate these are prominent themes within the Congressional Records. These could represent key

areas of legislative focus such as public health, science and technology policy, fiscal planning, and record-keeping.

- **Legislative and Administrative Focus:** Words like "commerce," "transport," "veteran," "affairs," and "government" suggest discussions around commerce and trade, transportation infrastructure, veterans' affairs, and broader government operations.
- **Policy and Management:** Terms such as "balance," "credit," "tax," and "fund" likely point to financial aspects of policymaking, including discussions on tax credits, budget balancing, and funding allocations.
- **Geographical References:** The presence of state names ("West Virginia," "North Dakota," "Rhode Island," "South Carolina") implies that some discussions were focused on state-specific issues or that representatives from these states were particularly active or mentioned in the records.
- **Political Process:** The words "absence," "quorum," "congress," "senate," and "roll" refer to the procedural aspects of the legislative process, such as attendance, voting, and the roles of different chambers.
- **Contextual Relevance:** The word "least" is relatively large, which could indicate its frequent use in discussions or debates, perhaps in phrases like "at least" or "least of which."
- **Diversity of Topics:** The variety of words and the lack of a single dominant term suggest a wide range of topics and issues are covered in the Congressional Records.
- **Potential Noise:** Some words in the cloud, like "record," "example," and "however," might be common across many documents and not necessarily indicative of specific discussions. These could be considered for stop-word filtering if they do not provide unique insights into the content of the records.

This interpretation helps identify the focal points and nuances of legislative discourse within the time frame covered by the Congressional Records. It is important to note that while word clouds provide a quick visual summary, they lack context and the ability to convey the complexity of discussions, which would require a deeper textual analysis.

3) Topic Modeling with latent Dirichlet allocation (LDA)

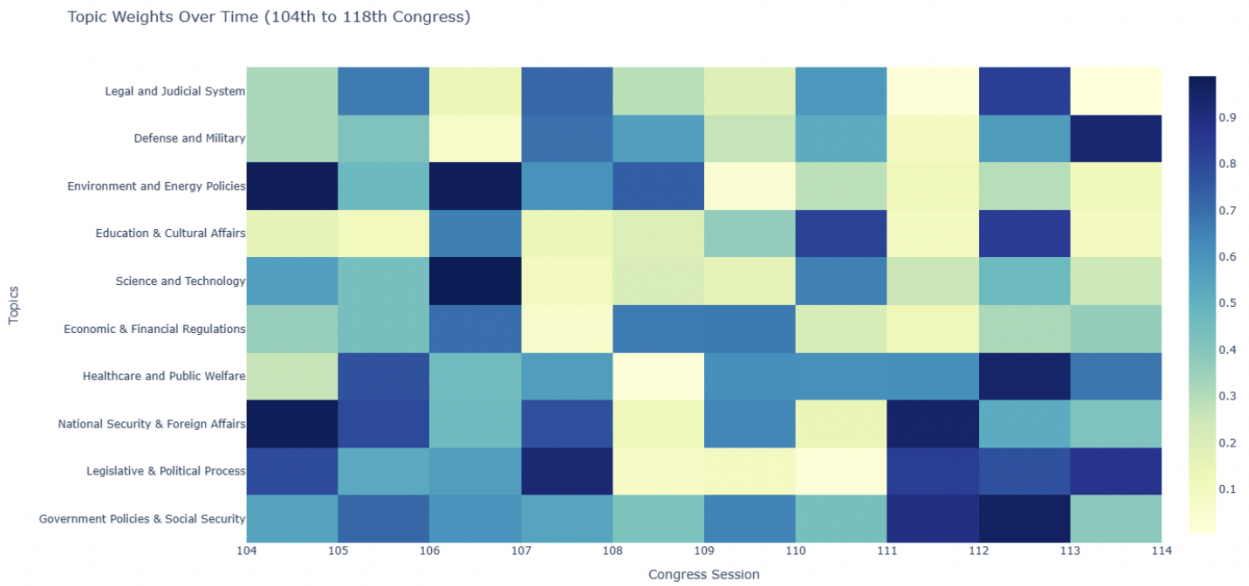
The results are generated from a latent Dirichlet allocation (LDA) analysis or a similar technique, which is used to discover abstract topics within a text corpus. Each topic is represented as a distribution over words.

- **Topic 0 - Public Funding and Health:** This topic seems to focus on public funding for various projects, with emphasis on work, health, and family protection. There's also a note on transportation and job creation.
- **Topic 1 - Family and Veteran Affairs:** This topic also centers on work and funding, but with a specific focus on family health, protection, and veteran affairs. Energy and school funding appear as well, alongside budget considerations.
- **Topic 2 - Healthcare Finance:** Dominated by health-related terms, this topic likely discusses healthcare funding, financial aspects of health care like tax and credit, and the impact on families and job payments.

- **Topic 3 - Military and Security:** While work and funding are again prominent, there's a distinct presence of military and protective services. The term "trump" suggests discussions related to or during the Trump administration.
- **Topic 4 - Energy and Technology:** Funding and work are persistent themes, but here the focus shifts towards military, energy projects, foreign affairs, and technology.
- **Topic 5 - Social Budgeting:** Topics of budgeting, tax, and public funding are prominent, with a social angle touching on family, children's welfare, and social issues.
- **Topic 6 - Economic Policy:** This topic is heavy on economic terms like tax, jobs, and budget, with an indication of political discourse involving Republicans.
- **Topic 7 - Uncommon Terms:** The terms here are not meaningful; they may represent noise, artifacts from text processing, or extremely rare discussions in the dataset.
- **Topic 8 - Education and Local Projects:** School funding is significant here, along with health, work, and public funding. Local energy and protection issues are also mentioned.
- **Topic 9 - Energy and International Affairs:** Fund allocation and health are recurrent, but there's a strong component of energy policy and employment, with references to international contexts like Iraq. In these topics, recurring themes like "fund," "work," "health," and "public" suggest these are core subjects in the Congressional Records.

The differences among topics likely reflect different aspects or contexts in which these terms are used (e.g., domestic policy, economic discussions, healthcare, military, and education). The presence of specific terms like "trump" and "iraq" in some topics can pinpoint the discussions to particular time periods or geopolitical focuses. Topic 7 stands out as an anomaly, which may require further cleaning of the data or adjustment of the model parameters.

4) Topic Weights Over Time in U.S. Congressional Sessions



The heatmap visualization represents the changing focus and priorities of topics discussed in the U.S. Congress across different sessions, from the 104th to the 118th. By examining the color gradient, we can infer which issues dominated legislative discussions during each session and observe the temporal trends and shifts in attention. For example, if we see a darker color in the "Healthcare and Public Welfare" row during a particular Congress session, it indicates that there was a substantial focus on healthcare and public welfare issues during that time. Conversely, a lighter shade in the same row for a different session would suggest that the topic was not as prominent in the congressional agenda.

Several insights can be drawn from this heatmap:

- **Government Policies & Social Security:** The varying intensity of colors across sessions implies fluctuations in the focus on government policies and social security, which could be attributed to changes in the political landscape or responses to specific events (e.g., economic crises or policy reform debates).
- **Legislative & Political Process:** A consistent pattern of color intensity suggests a steady discussion on the legislative and political process, likely due to the ongoing nature of legislative work and political activities within Congress.
- **National Security & Foreign Affairs:** Darker shades in certain periods could reflect times of international conflict, diplomatic events, or security concerns that necessitated more in-depth discussion and legislative action.
- **Healthcare and Public Welfare:** If darker colors are observed during sessions that coincide with healthcare debates or public health emergencies (like a pandemic), it would imply intense legislative activity surrounding these issues.
- **Economic & Financial Regulations:** Variations in color depth might correspond with economic cycles, financial crises, or significant regulatory reforms, indicating that economic and financial issues were more or less pressing in different sessions.
- **Science and Technology:** The intensity here could reflect Congress's engagement with innovation and technology-related policy, possibly influenced by technological advancements or industry lobbying.
- **Education & Cultural Affairs:** The focus on education and culture may ebb and flow based on domestic priorities or social movements, with darker colors indicating times of significant educational reforms or cultural debates.
- **Environment and Energy Policies:** Darker colors might correlate with periods of environmental crises, shifts in energy policy, or global climate agreements, indicating a heightened legislative focus on these topics.
- **Defense and Military:** This could show times of war, military engagement, or defense spending debates, with more profound colors indicating periods of intensified defense-related legislative activity.
- **Legal and Judicial System:** Changes in color intensity could reflect Congress's focus on judicial appointments, criminal justice reform, or significant legal debates.

Overall, the heatmap serves as a tool to understand not only the shifting priorities of Congress over time but also to infer historical contexts that might have influenced these changes. It can be a starting point for

more in-depth analysis of each topic's underlying causes for gaining or losing prominence in legislative discussions.

Relation of Tasks to the Larger Project

The tasks completed in this project were not just individual contributions but integral parts of a cohesive whole. The data retrieval process, for instance, was more than just a preliminary step; it was the cornerstone of our analysis, ensuring that all subsequent investigations were grounded in accurate and thoroughly vetted information. This meticulous approach to data collection and refinement was essential in guaranteeing the validity of our analyses, which were predicated on the reliability of this foundational data.

The role of the GitHub repository and its comprehensive documentation was paramount in fostering an environment of transparency and collaboration. These elements were critical in a project of this magnitude and complexity. The repository served a dual purpose: it acted as a chronological record of the project's evolution and as a dynamic, accessible resource for the broader research community. This aspect of the project exemplified our commitment to open-source principles and knowledge sharing.

Furthermore, the development of educational resources, such as the text mining tutorial and the 'data_exploration.ipynb' notebook, was a significant endeavor. These resources translated the project's methodologies into practical, real-world applications, thereby broadening the impact of our work. They served as valuable tools for those looking to apply similar techniques in their research, demonstrating the practical utility and adaptability of our methods.

Advancement of Career Goals

This project was a catalyst in advancing my career goals. It honed my technical expertise in data retrieval, processing, and analysis, particularly within the context of large-scale, complex datasets. The skills developed in managing intricate APIs and voluminous datasets have become indispensable assets in my data science toolkit.

Moreover, the project enhanced my competencies in documentation and knowledge dissemination. These skills are vital for leadership and innovation in research and development sectors. By creating and sharing educational resources, I not only improved my abilities in pedagogy and communication but also positioned myself as a thought leader in the field.

The project's contribution to my professional portfolio is substantial, showcasing my proficiency in navigating and resolving real-world data challenges. This experience has been instrumental in broadening my career prospects, opening new avenues in research, data analysis, and project management. My involvement in this project has not only solidified my standing in the realm of political data science but also expanded my opportunities for future ventures in data-driven research and analysis.

Appreciation of Collaborative Networks and Resources

We are grateful for the invaluable support from the Congress.gov GitHub page[1], the U.S. Government Publishing Office (GPO)[2], and the Stanford Congressional Record resources[3]. Their dedication and resources were instrumental in our research efforts.

Conclusion

This project was not just about data analysis; it was a comprehensive learning experience that allowed me to grow professionally. Working on such a large-scale and impactful project has improved my skills in data science, collaboration, and problem-solving. The opportunity to connect with esteemed scholars and leverage valuable resources has greatly enriched the project and my professional journey.

[1] We are grateful for the invaluable support from the Congress.gov GitHub page, which played a crucial role in maintaining and publishing congressional records, aiding our data collection efforts.
[2] The U.S. Government Publishing Office (GPO) provided essential resources that greatly enhanced the depth and accuracy of our analysis, particularly in terms of detailed data on U.S. Congress legislators.
[3] The Stanford Congressional Record resources, including parsed speeches and phrase counts, offered insightful methodologies for metadata extraction and data preparation. Their well-documented preprocessing pipeline was instrumental in guiding our approach to handling complex datasets.