Project Proposal

1. Research questions and the standard for "success"

Speeches serve as crucial data for understanding the agendas of political figures. While political speeches have emerged as critical data in political science, TAD-based studies of speeches of Chinese leaders remain scarce. Lim et al. (2025) conducted a preliminary exploration of Xi's agenda using *The Database of Xi Jinping's Important Speech Series*, identifying 25 topics and illustrating the temporal trends of their proportion. However, since this study's utilizes a variety of overlapping sources, duplicate documents may introduce biases in the estimated topic proportions. Therefore, this paper aims to apply **supervised machine learning** to identify unique speeches within the database and then classify them using **structural topic model** to better understand the shift in Xi's agenda over time.

Furthermore, although the study by Lim et al. (2025) explores changes in Xi's agenda, these insights are solely derived from the temporal trends in estimated topic proportions, lacking a nuanced discussion of how these changes are reflected in semantic shifts. Therefore, this paper aims to apply **word embedding** to map semantic shifts in key terms, such as "reform," over time.

Finally, the study by Lim et al. (2025) focuses solely on the speeches of Xi, lacking a comparative perspective. This study seeks to incorporate the speeches of Xi's predecessors, such as Hu Jintao and Jiang Zemin, to enable a comparative analysis of different leaders. Specifically, I intend to compare the **cosine similarity** between key terms like “reform” and related concepts across different leadership periods. This analysis aims to examine **Whether Xi Jinping’s skepticism toward liberalizing market and empowering private enterprises has led him to strengthen centralizing control over China’s economy** (Leutert, 2018; Friedman, 2023).

If we observe that Xi’s concept of “reform” is semantically farther from “poverty right” and “rule of law” compared to his predecessors, while being closer to terms such as “party,” “regulation,” and “leadership,” this would provide support for the above hypothesis.

2. Data sources and availability

As I demonstrated in **Assignment 1**, I have already applied web scraping to collect all speeches from *The Database of Xi Jinping’s Important Speech Series*. Therefore, obtaining Xi’s speech data is not a major concern.

For the speeches of the two previous leaders, I plan to use the published collections: *Selected Works of Hu Jintao* (three volumes) and *Selected Works of Jiang Zemin* (three volumes). Although there is no existing corpus, PDF versions are accessible online, allowing me to convert them from PDF to TXT and then construct the corpus.

3. The text-as-data methods

I plan to apply supervised machine learning, STM, word embedding, and cosine similarity in my analysis. The specific implementations of these methods are detailed in Section 1 (Methods are bolded for emphasis).

Reference

Friedman, J. (Nov 2, 2023). The Maoist Roots of Xi's Economic Dilemma. Foreign Policy. https://foreignpolicy.com/2023/11/02/china-economy-xi-socialism-growth-consumption/

Leutert, W. (2018). Firm Control: Governing the State-owned Economy Under Xi Jinping. *China Perspectives*, *2018*(1-2 (113)), 27–36. https://doi.org/10.4000/chinaperspectives.7605

Lim, J., Ito, A., & Zhang, H. (2025). Uncovering Xi Jinping's Policy Agenda: Text As Data Approach. *Developing Economies*, *63*(1), 9–46. https://doi.org/10.1111/deve.12418