INSTRUCTIONS FOR TRAINING DATA

Training data for Chapter 3 was drawn from two sources.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries, 2018-01-16. **https://data.stanford.edu/congress_text**

Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin, and Luke Sonnet (2025). *Voteview: Congressional Roll-Call Votes Database*. https://voteview.com/

For licensing reasons, steps for re-creating the training dataset are provided here in lieu of the data itself.

DATA COMPILATION STEPS

Download the hein-daily.zip file from **https://data.stanford.edu/congress_text**. Extract the speaker map file and speeches file.

Download the NOMINATE dataset for the 114th U.S. House, available at: https://voteview.com/data

Join the speaker map and speeches file on ‘speakerid’

Group the speeches by speaker:

- `E.g.: d = all114.groupby('speakerid').agg({'speech': lambda x: '.'.join(x)})`

Subset the joined file to only the House:

- use ‘chamber’ column and filter to ‘H’ for the House only

To join the NOMINATE file:

- Manually add the ‘speakerid’ value from the text file in a column for each representative in the NOMINATE file.
- Join text and NOMINATE file on ‘speakerid.’
- Optionally, drop unused columns.