Appendix D: How to Use TMHA in a U.S. History Course

While the data used for TMHA comes from British history, the code, concepts, and interpretive methods taught in this book can be directly translated to American history. We suggest the exercises in these chapters are ideal entry points for a U.S. history course and for researchers working with the United States Congressional Records, American newspapers, reform pamphlets, presidential addresses, legislative hearings, or other corpora of American political discourse.

The workflow for text mining U.S. corpora follows the same core logic demonstrated throughout this book: first, decide whether you aim to perform exploratory data analysis–where the goal is to scan broadly across the archive to surface patterns–or whether you will begin instead with a clearly stated research question that already specifies the issues of interest. Based on that purpose, select an appropriate unit of analysis (e.g., speeches, records, debate sections). Segment the archive into meaningful temporal intervals (for example, by year, decade, session, or administration). Once the archive is temporally structured, employ computational analysis to trace how language changes over time.

In this section, we describe how this same workflow can be used in an American History course—especially one focused on the U.S. Congress and the modern Congressional Record.

Subsetting the Data

Throughout the book, Hansard was divided into small, computationally legible chunks so that each dataset could be run on student laptops. The same principle applies here: rather than requiring students to ingest two centuries of U.S. Congress at once, you can create manageable, "bite-sized" corpora of specific dates or more narrow thematic domains (for example: Reconstruction, New Deal legislation, Civil Rights era, War on Drugs, Federal environmental policy, etc.). We recommend taking roughly ten-year slices for 19th and early 20th century congressional text (when the daily volume is lower), and shifting to five-year slices once you reach the 2000s and contemporary era (when the number of tokens, documents, and debate days increases dramatically).

Structuring the Data

To use a dataset like the Congressional Record or an American historical newspaper archive directly with this book without edits, the dataset should be placed into a CSV format. To put a dataset into CSV format, export the cleaned text out of your database, text mining environment, or scraping pipeline as a table where each row represents one speech unit. A row-level speech unit can be defined as: one speaking turn, one paragraph, one debate section, one article, or one floor contribution. The key is consistency. Once the units are standardized, write them out to a CSV file using the tools in your tool chest.

Some tools you might use include spreadsheet tools (Excel, LibreOffice Calc, Google Sheets), scripting tools such as R (readr::write_csv) or Python (pandas.DataFrame.to_csv()), OCR pipelines like Tesseract for converting scanned PDFs into machine-readable text, you can use ChatGPT to help you write code to do this. For example, describe your dataset or, if you are still looking for the precise wording, give ChatGPT a screenshot of your data. Even command-line utilities like awk and sed are powerful for batch text cleaning if you are working at scale or on a Linux server. Any combination of these tools is valid—digital humanities encourages working with the tools that best match your context, your workflow, and your comfort level. What matters most is that the final CSV is consistent and legible so that the downstream visualization and modeling code can operate reliably.

Running Code

Once you have a CSV, you can now use the code in this book as-is by renaming the dataset's columns to match the naming conventions from the chapters. For example, make sure the main text column is simply called "text", and if relevant, make the debate category column simply "debate". Alternatively, you can lightly modify the code to point to your dataset's existing column names instead. Both approaches work. The advantage of renaming the dataset columns is that you standardize everything once and the book code runs without modification. The advantage of modifying the code instead is that students can see (and practice) adapting research code to handle slightly different archival schemas—an extremely valuable applied research skill in the real world of historical data wrangling.

Asking Questions

After the dataset is aligned, you can ask parallel interpretive questions: what were the major political debates in your period of U.S. Congress? For the 19th century, might the analytic focus be on slavery or civil rights? For the 20th century, is the analytic focus on debates around abortion or the war on terror? In the 21st century, perhaps these debates focus around climate change or AI governance? Selecting topical boundaries and comparing decades against one-another becomes a central interpretive activity. Students can pull windows before and after turning points to examine how the language of within the corpus changes over time—just as we did with Hansard.

For example, in a unit on the history of voting rights, students could construct a corpus from 1870 to the present and examine debates around suffrage expansion and restriction—from Reconstruction, to Jim Crow, to the Voting Rights Act, to post-Shelby County. They could then use the same visualization strategies (bar-charts, scatterplots, heatmaps) to evaluate whether the patterns of language match what the historiography predicts, or whether computational reading surfacesunexpected turning points that historians may not yet have fully theorized.

Conclusion

Interoperability is a core strength of this book. The workflows demonstrated in each chapter are research instruments that can be applied to many other corpora. The ability to move from Hansard to the Congressional Record–without reinventing the pipeline each time–can teach students how to generalize methods. Or, the book can enable researchers to move directly to analysis to compare political cultures and build cumulative insight across corpora, not just within a single national case.