

# Case Study – Modeling the Topics of Presidential Speeches

**DS 4002 Spring 2023 Graham Twente**

**Submission format:** PDF document, source code

## **Individual Assignment**

**General Description:** This assignment will have you create a topic model to describe U.S. presidential speeches over time. Submit to Canvas your managerial report and the link to your GitHub Repository.

**Prerequisites:** A basic programming course, understanding of topic modeling, and a growth mindset.

**Why am I doing this?** The goal of this assignment is to learn how to implement a topic model to answer a desired research question and to communicate findings in an applicable workplace scenario. You are expected to work individually, but you may develop your problem-solving skills by using a variety of resources available to you including those contained in this project folder and others available online.

- Course learning objective: synthesize extant research on a topic
- Course learning objective: write (and alter if necessary) a research question and a hypothesis to guide research efforts.
- Course learning objective: develop analysis plan to answer defined question.
- Course learning objective: prepare findings for communication to manager.

**What am I going to do?** Produce a managerial report for the White House Director of Speechwriting as if you held a data analyst role supporting the speechwriting efforts of the office prior to an upcoming State of the Union Address. This report should include an executive summary, motivation and context, hypothesis, research question, analysis, and conclusion sections. You will also assemble materials used in the project in a GitHub repository. Importantly, you are responsible for choosing a topic modeling method, creating the model, training the model, and analyzing the results of the process.

- 4–6-page managerial report document – to communicate project contents and findings to the White House Director of Speechwriting.
- GitHub repository – to provide resources like code, data, and references.

## **Constraints:**

- You are given a dataset containing speeches from U.S. presidents.
- You are only required to identify and describe the major topics contained in these speeches.
- Your model should treat the set of speeches as a large set or corpus (i.e., do not train the model on each speech individually).

- You are free to subset speeches after training the model by president, time, etc. to identify frequencies of overall topics or other trends by subgroup as you see fit.

### Tips for success:

- Take a moment to carefully read and understand the prompt and provided materials. Be sure your model addresses the prompt at hand.
- Be creative – there is no ‘best’ way to do this assignment, make decisions that make sense to you, identify, and explain those decisions.
- Choose appropriate topic modeling methods – there are many different topic modeling methods available, each with its own strengths and weaknesses. Consider factors such as size and complexity of the dataset, the number of topics you want to identify, and the desired coherence of generated topics.

**How will I know when I have succeeded?** You will meet specifications by following the criteria in the following rubric:

Spec Category	Spec Details
Managerial Report	<p>Goal: This step represents a report that you will submit to your manger to brief them on the main approach, analysis, and conclusions of your research project.</p> <p>Your submission should include the following:</p> <ul style="list-style-type: none"> <li>• <b>Formatting:</b> 4-6 pages, PDF format</li> <li>• <b>Executive summary:</b> Use short sentences to outline the overall goal of this project and the work you completed.</li> <li>• <b>Motivation and Context:</b> In two paragraphs, synthesize relevant research and current knowledge about presidential speech contents.</li> <li>• <b>Hypothesis:</b> Write a one sentence hypothesis that is quantifiable, testable, and related to the motivation and context.</li> <li>• <b>Research Question:</b> Write a one sentence question that your project aims to answer.</li> <li>• <b>Analysis:</b> Explain your analysis methodology such that a peer could replicate your procedure without prior knowledge of the project, include any exploratory plots of the presidential speeches dataset, and highlight any changes made to the hypothesis,</li> </ul>

	<p>research question, or model that occurred over the course of the project.</p> <ul style="list-style-type: none"> <li>• <b>Conclusion:</b> Summarize your key learnings from this project, describe challenges that you faced, how you addressed these challenges, and next steps for future research on the topic.</li> </ul>
GitHub Repository	<p>Goal: Explain the goal of the project and main learnings</p> <p>Your GitHub repository should include the following:</p> <ul style="list-style-type: none"> <li>• README.md: Create a landing page for readers of your project to describe the contents of the repository including code, data, and references in IEEE style.</li> <li>• SRC folder: Submit a commented script in a programming language of your choice that executes your methodology (e.g., data cleaning, exploratory data analysis, model building, model testing, and analysis).</li> <li>• LICENSE.md: Select an appropriate license for the repository, the MIT license is likely appropriate.</li> <li>• DATA folder: Place data used for the project in this folder.</li> </ul>

Acknowledgements: This rubric is designed with help from rubrics created by Professor Loreto Alonzi in DS 4002 and based on research by [Streifer & Palmer \(2020\)](#).