How Topic Modeling illuminates political issues in the American Progressive Era

Jason Lafita

Department of Cognitive Science, Johns Hopkins University

Computational Intelligence for the Humanities

Professor Craig Messner

May 12, 2025

Abstract

In this project, I discern the political issues key during the American Progressive Era (1870-1920). This era saw increased public involvement in politics and created partisan dynamics seen today. I create multiple topic models over a corpus of speeches from this era to parse out certain topics and visualized topic prevalence in graphical form. The chief result shows that discussions on slavery and government are to be the most common published speeches.

Materials and Methods

My corpus is derived from Hathi Trust. I subset it into the category of "speeches" and filter by date from those published between 1870 and 1920. Then, I filter language to English and place of publication to the United States. This produces a corpus of 913 works, most of which were published in newspapers from that time. I begin with a zipped JSON file consisting of the full text of each document.

From there, I generate a virtual environment and install all needed models from my code. I load in the file and tokenize the full text field of the JSON file in the corpus using the Natural Language Toolkit by sentence, and then by word. Next, I use GenSim to generate a dictionary and a JSON file of subdocuments to be used to create topic models of the data, varying the number of topics. This data is visualized using a bar graph to show which topics are most prominent, Furthermore, the model is evaluated using the coherence score and perplexity.

Procedure

To generate the models, I first use the tokenized zipped JSON file to perform GenSim preprocessing and dictionary modules to create files with a training dictionary and a corpus with each subdocument. To refine the model, I filter the data considered using the parameters of a minimum of 200 subdocuments, no maximum subdocuments, a minimum word length of 3, a

minimum word count of 30, and a maximum proportion of 0.7. These two files are then inputted into the training of the topic model.

I use the generated subdocuments and dictionary to train five topic models using GenSim, varying the number of topics as 10, 20, 30, 40, and 50. For each, the program performs 10 passes and 20 iterations and has a chunk size of 2000. I write the results of the topic model into a JSON file for later analysis. In addition, the model is written into another file for evaluation.

From there, I apply each model to the subdocument file to generate an overall count of the prevalence of each topic. This creates a JSON file with the total count of each topic throughout the dataset. I inspect the counts generated by turning them into one bar graph for each model that maps the presence of each topic.

I evaluate the results of this using the coherence score built into GenSim, which looks at the semantic similarity of the words within each topic to ensure their interpretability. I cross-reference this with the model perplexity. I apply this module to each of the models generated. I output the results for each model into a JSON file. My code takes all of the models at once. Finally, I present these results in a table in markdown form.

Results and Discussion

Based on the evaluation, the model with 10 topics was by far the most effective, as there was a sharp dropoff in coherence score after that. The models do improve in perplexity, but this score is less indicative of human interpretability than the former. I chose to calculate the UMass coherence, since that reflects the use of the topics in relation to one another, which is more likely to reflect the efficacy of the model, but there was still a dramatic dropoff after the 10 topic model.

To analyze the results to glean what issues in political speeches were key, I cross reference the topics printed during the model training step with that of the graphs showing which topics were the most common within each model. The most common topics in each model were one related to legal matters and justice; slavery and what to do with the acquisition of new land; the two political parties and the nation; foreign affairs, specifically in Spain; and general plans and strategies out of 10, 20, 30, 40, and 50 topics, respectively.

I find the differences in most common topics to be incredibly odd. I initially thought that the most common topic would remain relatively similar across models, but that is not the case. The lack of interpretability for the model with 50 topics — where the most common topic contains the words “present,” “ought,” “necessary,” and other hollow terms — backs up the low coherence score for that model.

The topics for the model with 40 topics were also odd. Given the context of the other topics, I found it interesting that speeches on foreign affairs were so prominent in this model, when that was less common in others. The inclusion of Spain as a word within the topic made sense, given the Spanish-American War being a prominent conflict from that era. Also, this model was the only one where there was no clear dominant topic that had far more documents than all the others. The second most common topic was one similar to that of five.

Of all the models, I find that with 20 topics to produce the most coherent results for the purposes of this study. Despite its lower coherence score, the topics were specific enough to parse out the actual political issues being discussed but not too specific that the breadth of the data was overwhelming to look at. The former is an issue present in the model with 10 topics, while the latter appears in all others.

I find that the topic of slavery and new territories being most prominent was unexpected,

seeing as slavery ended in 1865. However, with Reconstruction, as well as the expansion of the states, this result does make sense. The second most common topic from this time seems to be related to government and its structure, containing terms like “Congress,” “constitution,” “states,” and “law.” Speeches of this nature may have talked about the structure of government, since the Progressive Era was a time of much democratic reform, with innovations such as majoritarianism in the House of Representatives and the secret ballot.

My initial thought behind this study was to find which topics politicians may have spoken on to appeal to voters during this era when politics was becoming an increasingly public-facing endeavor. The presence of slavery and progressive reform therefore being the most common topics aligns with that, since slavery and Reconstruction was a very hot-button issue that many Americans had opinions on. The latter was more interesting, since that may reflect speeches from the floor of Congress or in other legislatures being published, as opposed to campaign speeches. However, given the presence of topics around slavery as the largest by a large margin in this corpus, I conclude that this issue was likely the most prevalent from the era.

A limitation of this study was the lack of metadata, which did not allow me to collect prevalence of topics over time; instead, I only could look at mentions over the whole corpus. Furthermore, I did not have information about the context of the speeches — only that they were published between the selected dates. Therefore, I could not narrow down which speeches I used to train my model on further. If I were to repeat this experiment in the future, I would try to bridge these gaps. I would also refine my search further to possibly differ between speeches given by Democratic and Republican politicians to see if there was any difference in the topics that each publicly spoke on.