Luke Moles (lmm8fb@virginia.edu)
DS 5001
11 May 2021

Exploratory Text Analysis of Early Twentieth Century American Political Science Articles

Introduction

The early 20th century represents an important time throughout the world. A large number of impactful events occurred during this period that had long lasting effects. This is true for the United States as well, where many international and domestic events became well discussed occurrences during their time. In particular, many events centered around politics, or influenced them significantly. The goal of this work is to analyze a text corpus related to American political science during this time period in order to investigate trends in politics and society.

The dataset was obtained from the Digital Scholars Workbench [1], and it is composed of full text JSTOR articles. All texts come from the Annals of the American Academy of Political and Social Science with political science category tags. The years of publication range from 1900 to 1924, with slightly greater density in the later years. For each article, the recorded features include author, title, publication date, and categories. Articles can have multiple categories, but all texts used in this work contain a political science category tag.

Methodology

The articles were first processed into important tables that would facilitate analysis. These included a library table with all relevant metadata for each article, a token table table organized with the hierarchy of the texts (document, paragraph, sentence, token), and a vocabulary table that contained all tokens used in the corpus along with important features of these terms.

Next, term frequency - inverse document frequency (TFIDF) scores with max scaling were calculated at the document level for later use. These were grouped with hierarchical clustering that used Ward linkage and Euclidean distance to examine important clusters. Principal Component Analysis (PCA) was also performed on these scores to analyze important latent structures in the data. The top 10 principal components by explained variance were retained, and loadings were calculated. After this, Latent Dirichlet Allocation (LDA) was performed on the articles to find 10 latent topics within the dataset, as well as the distributions for the generative process. Word embeddings were created with word2vec pretrained embeddings. These were analyzed in conjunction with the topics from LDA. Following this, sentiment analysis was conducted for the articles using the Bing sentiment lexicon.

Analysis

Using max scaling TF with sum aggregation, the top terms across the corpus by TFIDF included *peace, popular, strong, modern,* and *respect*. These seem to reflect the time period well. Although World War I occurred during this time, it may have only encouraged the use of words and themes related to peace. Similarly, *modern* may be expected due to the relatively quick progress in science and industrial technologies at the time, especially leading up to and during the war.

Figure 1: Hierarchical Clustering

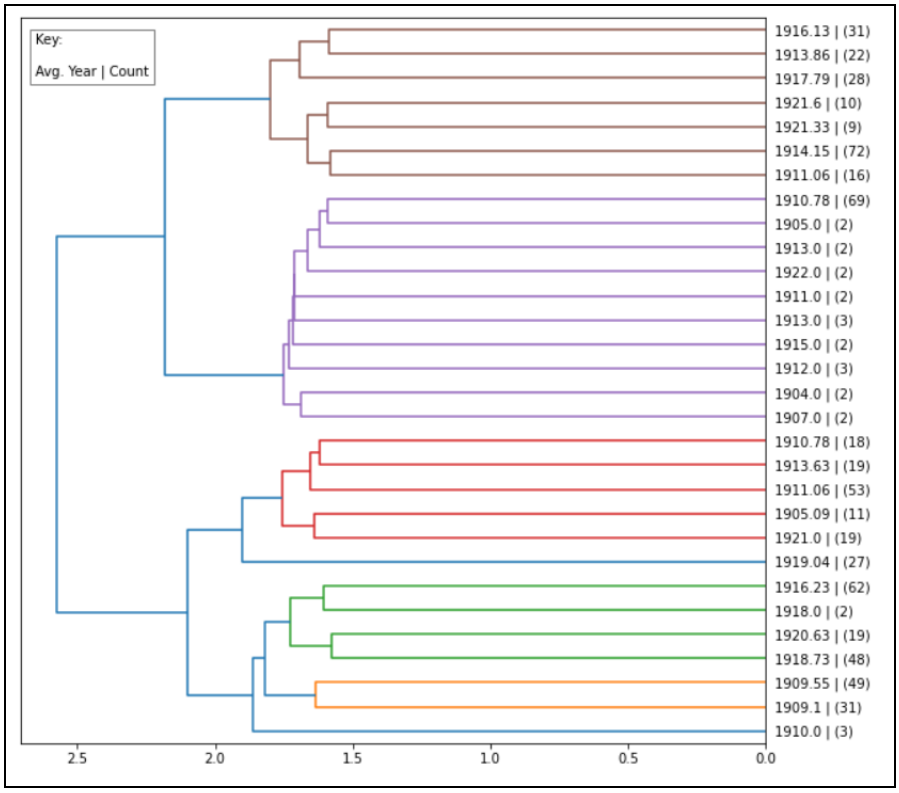


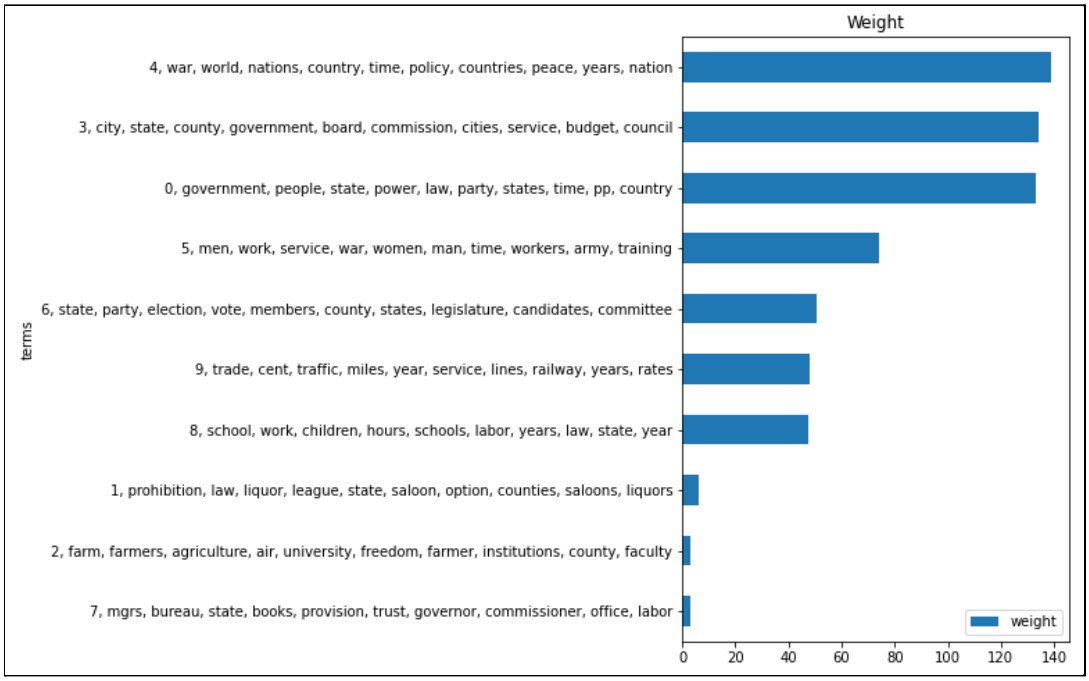
Figure 2: PCA

The hierarchical clustering performed on these aggregated values was able to identify a large group of possible outliers. These are the purple groups in Figure 1, and they mostly consist of two or three articles each. There is also a cluster of 80 articles with an average publication year of 1909, possibly corresponding to a significant event. Lastly, the green cluster in Figure 1 appears to connect multiple articles from the end of the 1910's. Generally, the clusters showed that there may be some trend in topics over time, even given the relatively short timespan of the corpus. This is further investigated in the following methods.

The most common categories in the corpus were "Political science - Government" and "Political science - Military." PCA showed that the first principal components are highly effective at distinguishing between these categories.

Unlike clustering, PCA was not successful at measuring the date of publication. This suggests that most of the variability in articles during this period is accounted for by domain, not time.

Figure 3: LDA



Topic modeling with LDA revealed important information about the latent groupings of documents. Figure 3 represents the aggregated weights for each of the ten topics, as well as the top ten words from each topic based on the calculated distributions. We observe that the most heavily weighted topics are related to government, administration, and military. An important distinction can be made between two military topics. One topic (4) relates to war as it is seen politically by governments on an international scale, while the other topic (5) contains words tying it to the human element with words such as *men, women, work, service,* and *training*. There are also topics related to trade, education, and agriculture. Lastly, an interesting topic revealed here is associated with prohibition, which was first enacted around 1920, but discussed before then. We can further strengthen these latent topics with word embeddings.

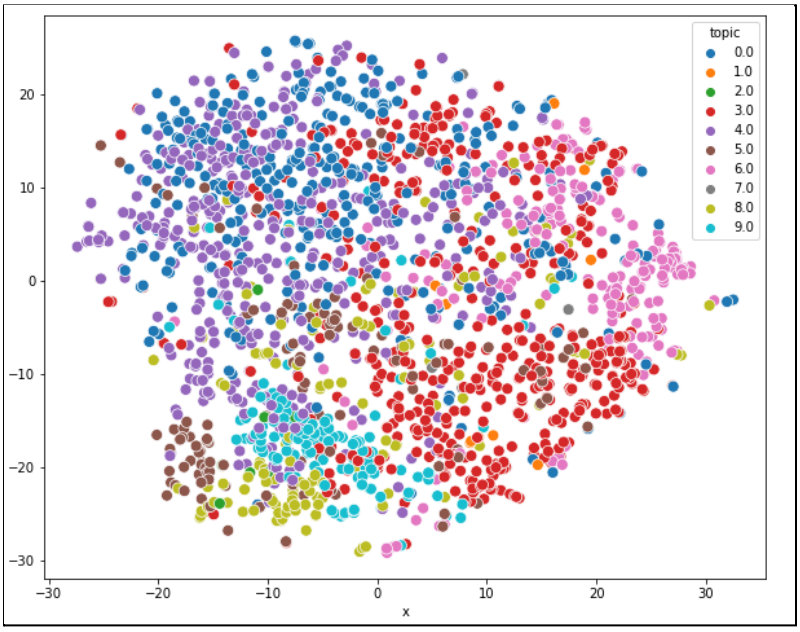
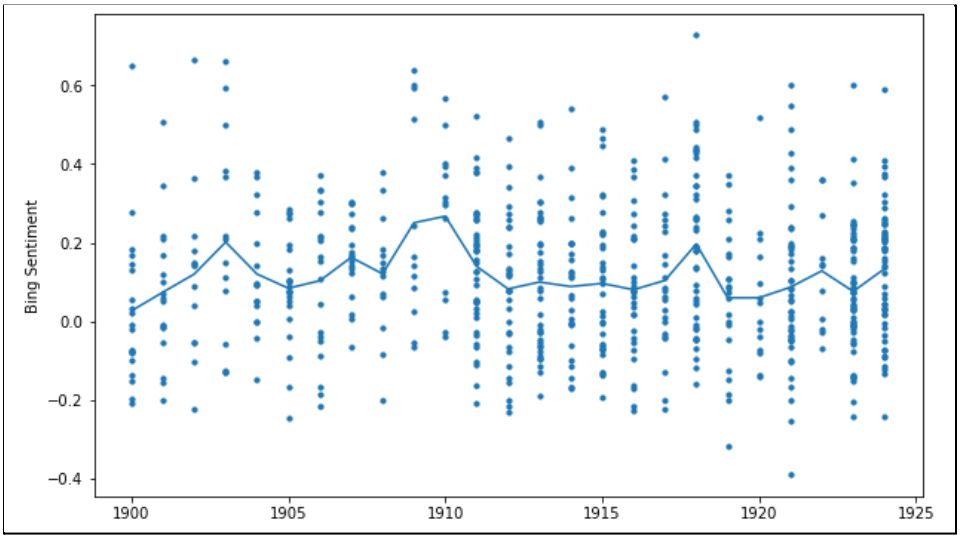


Figure 4: t-SNE of Word Embeddings

The t-SNE plot in Figure 4 shows that word choice tends to align with these topics.

To examine further trends in time, the Bing sentiment lexicon was used. Average sentiment was calculated for documents over every year in the corpus, and the results indicate that, although there was a great deal of variability, there were clear peaks and valleys in overall sentiment. For example, a peak occurred immediately at the end of World War I, but this was followed quickly by a sharp drop in sentiment. This drop may have been caused by either the Spanish Influenza Pandemic, or the economic and cultural instability associated with war recovery period. An interesting observation from this data is that the highest sentiment occurred during the years 1909-1910. Cross-referencing this with the topic model shows that most of the articles were related to topic eight, which focuses on education and labor. This may suggest that these years were especially promising in these fields. They also correspond to the first years of the Taft presidency, so there may exist a connection to his platform and administration. When examining average sentiment by topic, we observe that the highest sentiment tends to correspond to these categories of labor and education. The agriculture topic also scored particularly high. Meanwhile, the lowest topic was prohibition. Between the two topics associated with war and military affairs, the one related to the human efforts (service, training, work) had higher average sentiment, while the government and administration topic of military affairs had low sentiment.

Figure 5: Avg. Sentiment over Time



Conclusion

The analyses suggest that both time and domain, unsurprisingly, have a significant effect on political discourse. Time is especially important for mapping the articles to major world events, even across multiple domains. By searching for latent topics and groupings, more detailed information could be gained about the data. This revealed that topics of education, labor, and agriculture were addressed with favorable sentiment, while war and government administration were less favorable. As mentioned, this may relate to the state of society during this period, or it may be a result of the actions of the government at the time. We also observed that the topic of prohibition was significant enough to be detected as a latent feature, and that it was viewed more negatively than other topics when measured by Bing sentiment scores. Although this data seems old, it may still be relevant and comparable to other time periods. To

extend this work, it may be meaningful to look for similar topics or trends in other time periods, or in other nations.

References

[1] Constellate. Dataset Builder. Online: https://tdm-pilot.org/. Accessed May 1, 2021.