Hart Greenwood

8/19/2020

Montana State Fund Interview Case Study

Unstructured Topical Review of POTUS Speeches, 1797-2016

Objective of Study: Algorithmically model and predict the main topics both by president and by era.

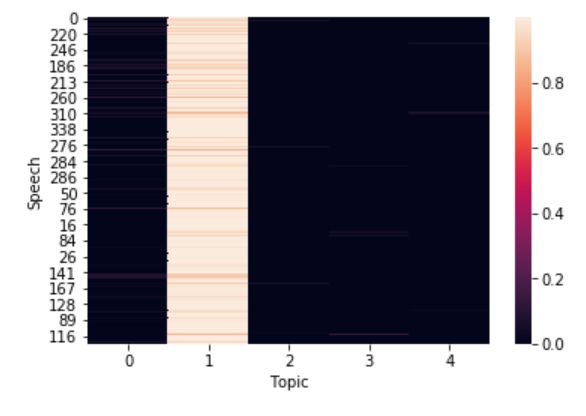
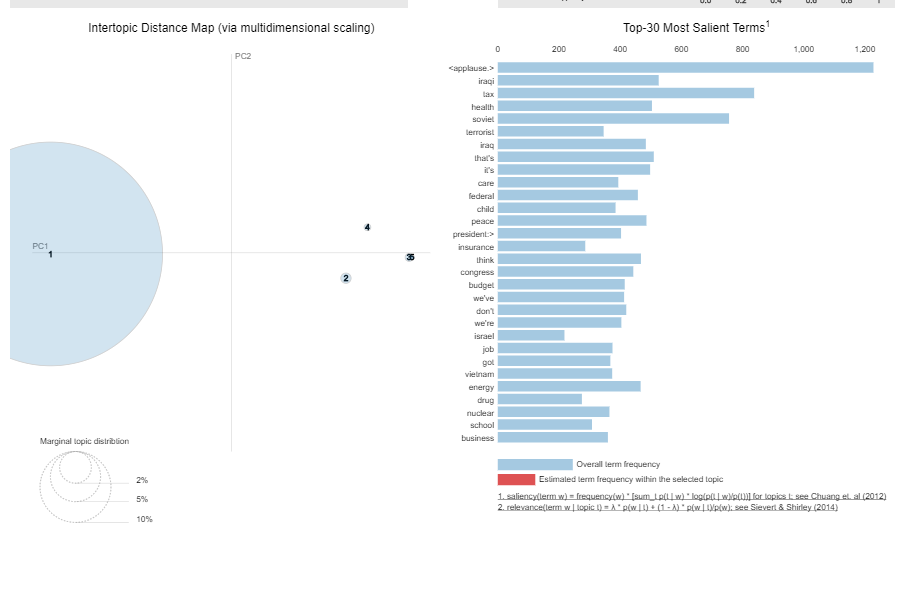
Executive Summary

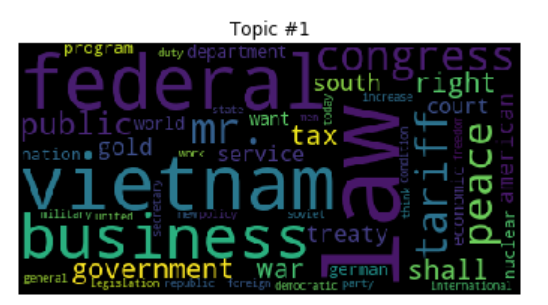
A total of 962 speeches were analyzed via a computational method known as Latent Dirichlet Allocation (LDA). This methodology observed the distribution of frequently used words within a distribution of some number of topics which was dictated by the modeler. The machine then reviewed each document in a corpus and assigned a distribution of each topic to each document. This distribution, along with the key words that make up each topic, was used to infer the topic of each document. This process is known as topic modelling. The documents making up each president’s speeches were grouped by president and era to understand the prevalent topics within each group. Some of the key findings were:

* Assigning a main topic to a president in times of strife was simple as that topic easily surfaced
  + Wilson and Roosevelt almost exclusively talked about War
* A shift in discussion was easily observed between primary presidents, such as Washington, Adams, and Jefferson, and those that followed
  + The topical discussion changed from nation building to the economy and foreign relations
* Within the eras provided, major events generally did not rise to the top, with the exception of the World Wars and Vietnam
  + This indicated that while singular presidents may focus on significant events, such as depressions and pandemics, more commonplace topics, including foreign relations and taxation, dominated conversation across each era.
* Additional modelling and hyperparameter tuning can be done to reduce topical and key word overlap

Additional Information

In order to instantiate an LDA model, the modeler must first load, clean, and process the data. The data was loaded and transformed into a dataframe and then cleaned via the removal of special characters, stop words (commonly occurring words that generally don’t add value), and numeric characters. The text was then lemmatized (word transformation with morphological analysis) which can help improve uniqueness and limit topical overlap. Stemming was also performed in a prior model but lemmatization appeared to produce better results. The processed text was then transformed into a dictionary of words, and the number of occurrences for each word was counted in each document. This count was then vectorized for computational consumption via TF-IDF. This step was crucial to limiting topical overlap. Finally, an LDA model was instantiated using the transformed data. This model was visualized using LDAvis to review Salient Term and topical overlap. Hyperparameters within the LDA model were then tuned to help improve model performance. Given that this is unsupervised learning, inspecting model performance can be difficult. Model performance was evaluated based on topic interpretability and key words with weights greater than 0.000. After model tuning, the topical distribution was visualized with a heatmap and bar chart. These tools simplified topic visualization for a corpus. WordCloud was then used to identify key words in each topic and highlight the most prevalent topic. The distribution of topical results for both presidents and eras can be found in the document [here](https://github.com/hmgreenwood/Professional-Projects/blob/master/MSF_Case_Study/Hart_Greenwood_Case_Study_Results.xlsx). Code for each of the corpus evaluated can be found [here](https://github.com/hmgreenwood/Professional-Projects/blob/master/MSF_Case_Study/Topic_Modeling_of_Presidential_Speeches-1849-1893.ipynb). A sample of the visuals are shown below.





Next Steps and Questions

In the interest of time, the model was tuned on one set of documents. This may result in an appropriate model for that given corpus but does not likely represent a model that best matches another set of data. In order to truly feel confident in a topical pick, the model and topic count should be retuned with every corpus. In addition, bigrams were created, but were not used as they would not work within the framework of the model. It is anticipated that the model would improve with the addition of bigrams and trigrams. Also, the addition of stop words from outside of the NLTK library would also improve the model. Words such as president, federal, and shall all appear frequently in many speeches but do not necessarily contribute to topic selection. Another straightforward step would be to increase domain knowledge. Without clear key words (German, Vietnam, etc) understanding the context of the speech was challenging. With additional historical context, it would be simple to remove additional stop words and arrive at a logical topic pick.