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.

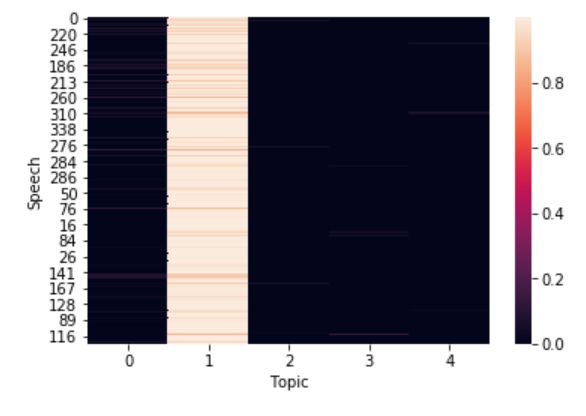
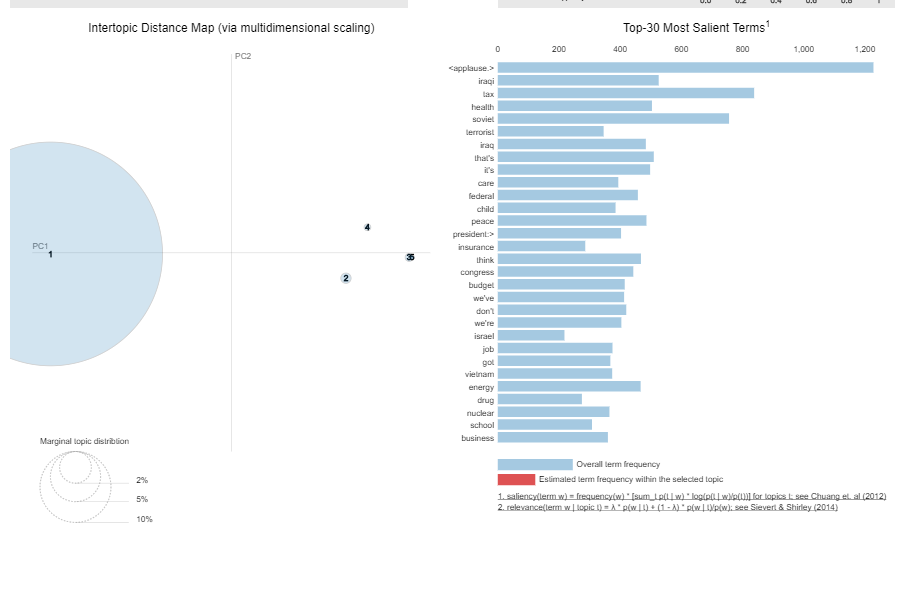
Executive Summary

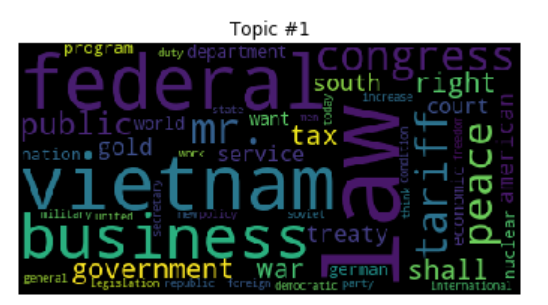
A total of 962 speeches were analyzed via a computational method known as Latent Dirichlet Allocation (LDA). This methodology observes the distribution of frequently used words within a distribution of some number of topics which is dictated by the modeler. The machine then reviews each document in a corpus and assigns a distribution of each topic to each document. This distribution, along with the key words that make up each topic can be 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 learnings are as follows:

* Assigning a main topic to a president in times of strife is very simple as that topic easily surfaces
  + Woodrow Wilson and Roosevelt almost exclusively talked about War
* A shift in discussion is easily observed between the first set of presidents and those following
  + The topical discussion changed from nation building to the economy and foreign relations
* Over the given time frames observed, major events generally did not rise to the top, with the exception of the World Wars and Vietnam
  + The economy and public affairs dominated most conversation, generally speaking
  + This indicates that while large events, such as depressions and pandemics may dominate our minds, they are not the events that make up most of our lives
* 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 transformed 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. I found this step crucial in 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 the LDA model were then tuned to help improve model performance. Given that this is unsupervised learning, inspecting model performance can be difficult. I was looking for topics that were interpretable and that had key words with weights greater than .000. After model tuning, I visualized the topical distribution with a heatmap and bar chart. These tools made it easy to visualize the most likely topic for a corpus. I then used a WordCloud to visualize the key words in each topic. This allowed me to make a judgement call on 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). A sample of the visuals are shown below.





Next Steps and Questions

In the interest of time, I tuned the model with the first set of data that I was reviewing. 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, I should retune the model and topic count for each set of documents. In addition, I created bigrams, but was unable to get those bigrams to work within the framework of the model. I would anticipate model improvement with the addition of bigrams and trigrams. I also only used the library selection of stop words and it is evident that I could easily improve performance by adding additional stop words such as president, federal, and shall. Another straightforward step would be to increase domain knowledge. I found that 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.