Diego Valdes

IST 736

Home Work 7

11/19/2019

**Introduction**

Most data collected today isn’t numbers. Most of the insight needed for businesses to survive and thrive isn’t going to come from tables with easily to translate values. Insight through data is derived from data that can be visual or in text documents. Being able to derive information from this type of data is a capability that separates businesses that will thrive from those that will die.

For any algorithm to be effective in the human world, it must make sense of what humans are discussing. What is suddenly important today? It may be easy to graph and plot numbers, but how does one plot the news of the day? Can topics be a variable on a heat map? Can partisanship be explained with a normal distribution? Can reactions from a debate be ranked and be part of a linear model? Well, while the answer to most of these questions is no, a better answer may be, not yet.

Topic modeling is a crucial part of the predictive analysis driving decisions today. Most of the data being generated today is not numbers that can easily be broken up into third normal form. It’s unstructured, ugly, and filled with text that can range in subject and feeling. Data is used to make sense of the world, but if that trend is to continue, and greater insights and capabilities obtained, deriving the meaning behind our words is the next step in understanding.

**Analysis**

**About the Data**

The data is a collection of speeches from the 110th Congress (House of Representatives). The collection is divided into subfolders based on sex and party affiliation, creating 4 subfolders (male democrats, male republicans, female democrats, female republicans). Each speed is labeled with the representative delivering the speech, the topic, and the date, and followed by the text. For the purpose of analysis, all of the folders were combined into one folder

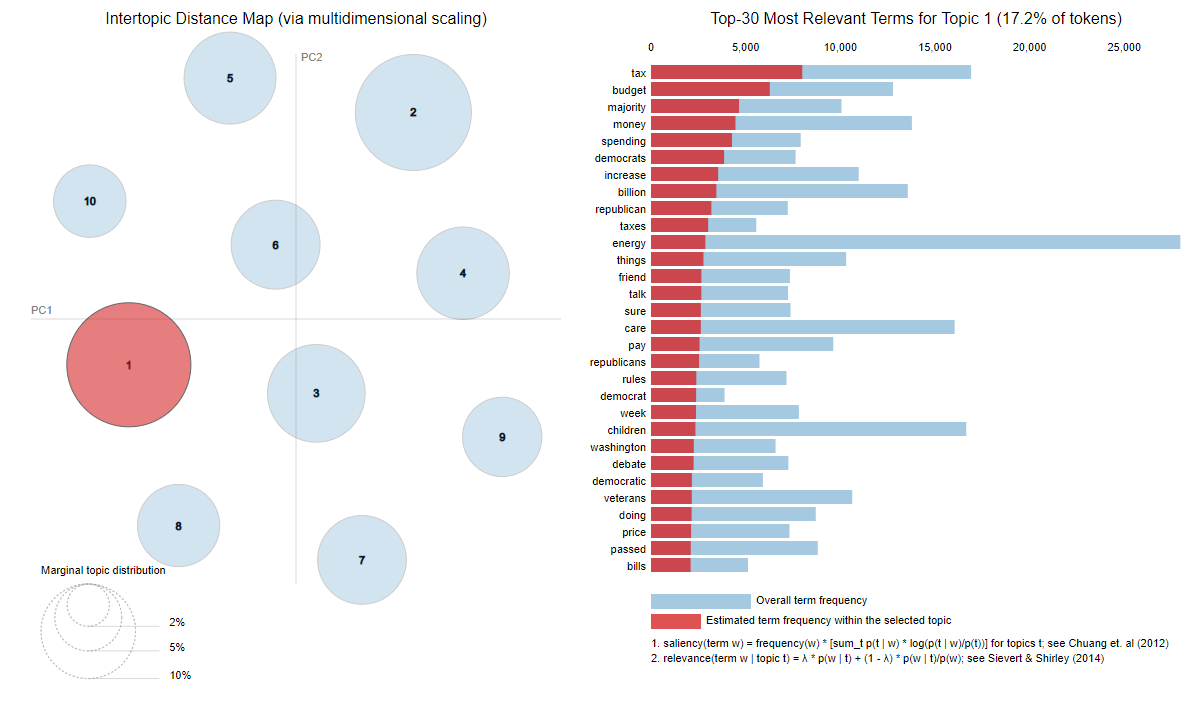
**Models**

Multiple countvectorizers were implemented to run through and LDA model. Parameters were adjusted and/or removed to find the best model. Once a general set of parameters inclusion was decided upon, they were adjusted with each creation of the model and the results compared to previous results for further adjustments. These were mainly in the max/min document frequency and number of features.

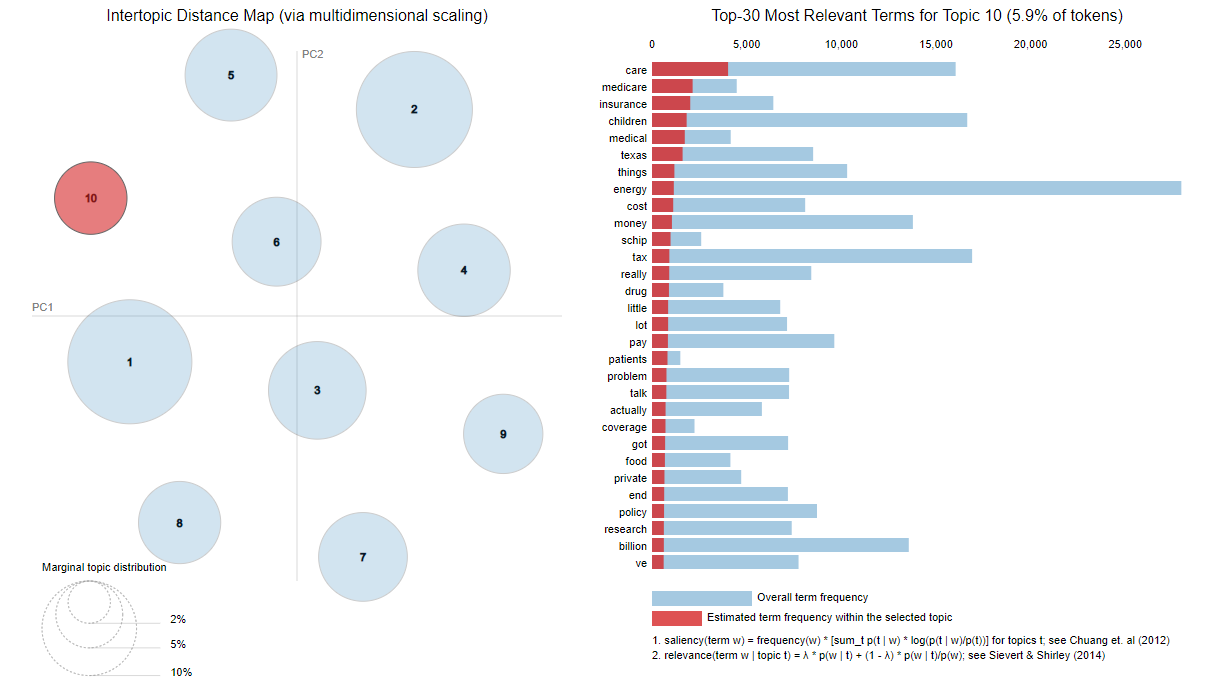
The LDA model also went through multiple adjustments based on results. Most of these adjustments came with the number of topics. Although Congress addresses 40-50 topics, having this value produced duplicate topics. Having a topic between 10 and 20 seemed to be the sweet spot, as there may be an occasional duplicate, but some topics that were absent when the variable was less than 10 were included.

**Results**

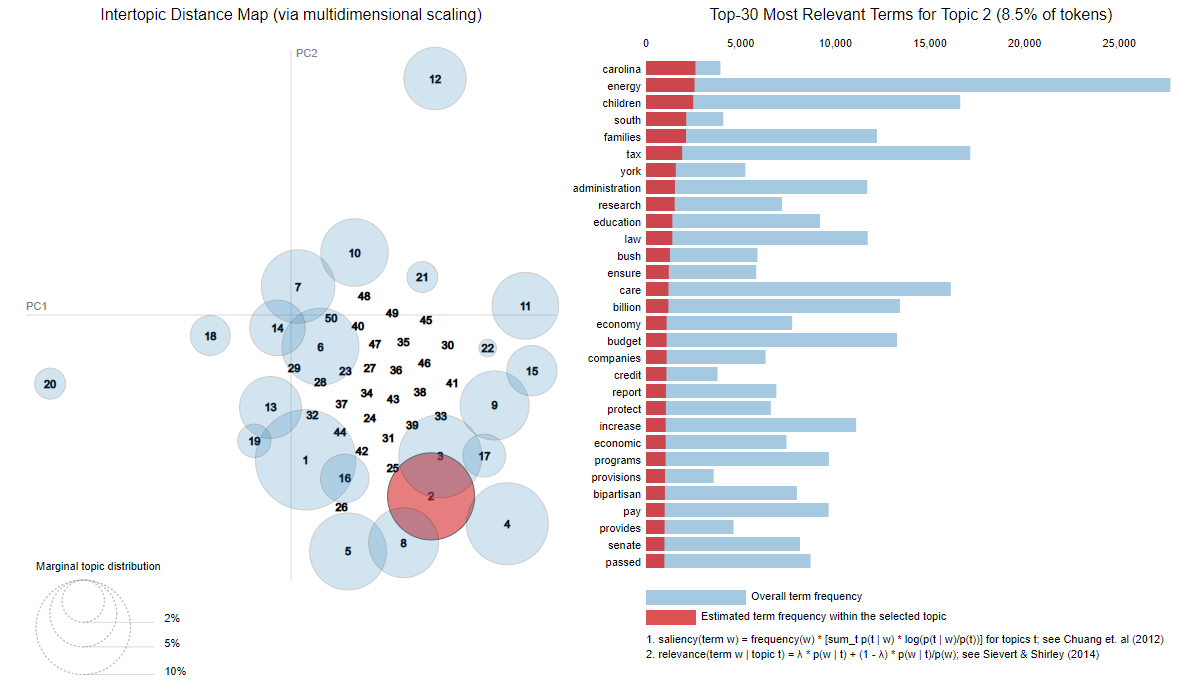
Through all the models many of the same topics appeared. If the number of topics exceeded 10, some topics appeared more than once, ie: energy, taxes, children, healthcare. Looking at the top 30 most salient terms in a topic describes how the terms describe the topic. As an example, topic 1 seems to be about taxes and government spending.



The smallest topic of these 10, to much of the dismay of many Americans is healthcare. This is easily determined by looking at the relevant terms.



If using a larger number for topics, it becomes less clear on deciphering the topics based on the key words. Using 50 as a parameter is a good example of this as the topic isn’t clearly defined.



**Conclusion**

The nature of language makes it difficult to analyze and derive meaning. No better example can be found that examining text messages, tweets, and Facebook posts where the literal reading of the sentence is not the true meaning. A complicated example would be the use of the most notorious ‘N’ word in the English language. How would a computer determine when the usage is friendly vs offensive when Americans have been struggling with it for over a hundred years.

The LDA model was very good at deriving the topics of the speeches analyzed. Regardless of parameters, the same general topics continued to appear. Closer examination of the top words in each topic are proof of the effectiveness. If the results of the model were used in a gameshow, where the most relevant words were used to describe a topic, humans would validate the results.

These results are promising, as the most relevant words do fit into a topic, they are not simply a collection of random words. The key in this dataset was keeping the topics concise and adjusting the model to eliminate topics that were being duplicated, which may have also led to bigger topics being watered down. It may be appropriate for certain topics to be broken up, ie: Veterans affairs can have sub topics, but for this analysis, an example of this wasn’t found.