Toni Hanrahan

Congressional Topic Analysis – Topic Modeling

**Introduction:**

Both government officials and business professionals need to monitor and react to topics and issues that arise in congressional sessions. New regulations can pose many challenges, and being prepared can provide a advantage. However, listening to, or reading the text of the Congressional Record of all the congressional sessions is not possible for most people.

Monitoring subscription services allow users to sign up to be proactively alerted to relevant legislation and regulation topics. Political scientists and correspondents also use similar monitoring methods to stay up to date. And historians can use this information to track trends over time, or measure the “productivity” of each Congress.

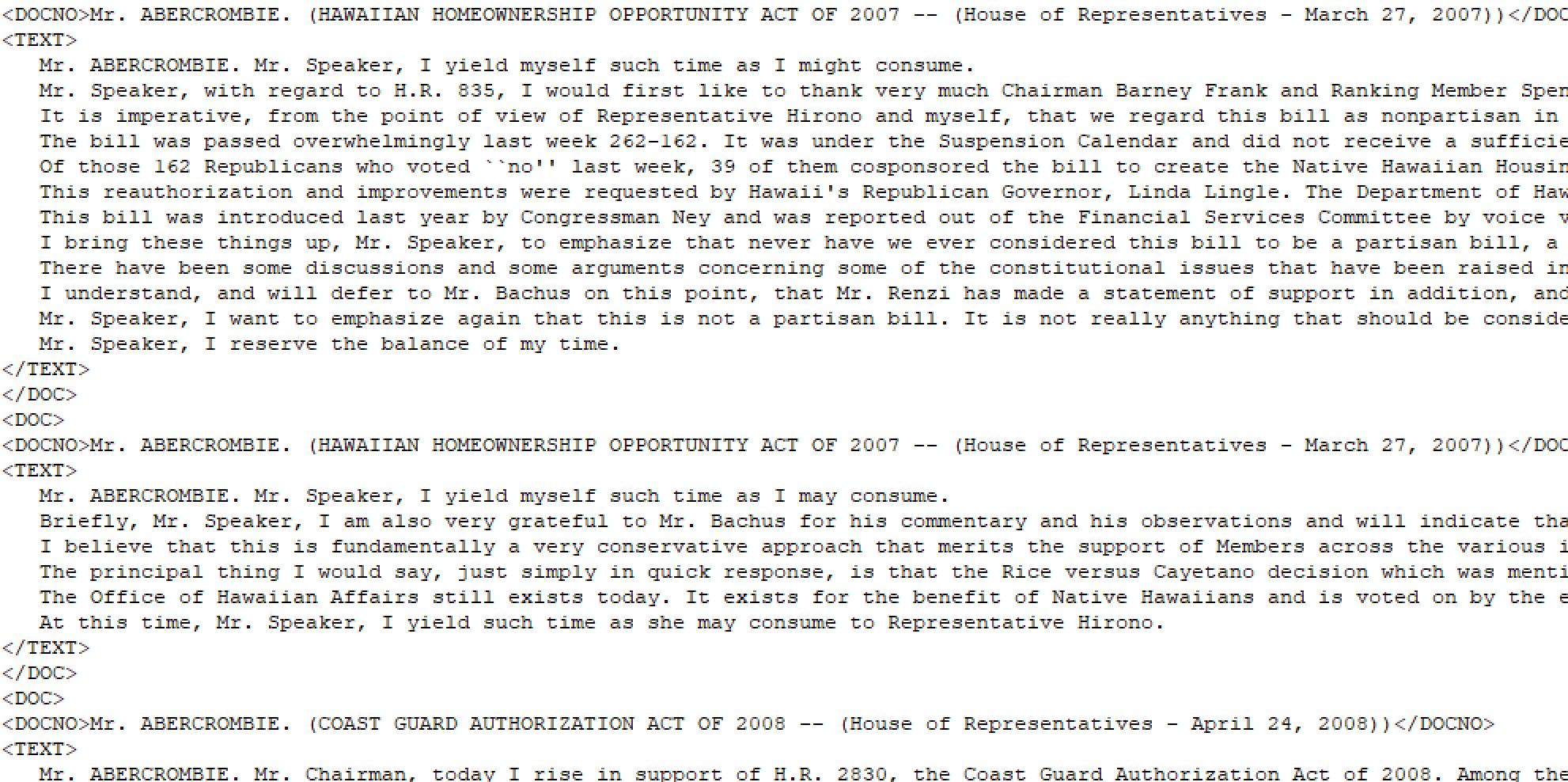
It is fairly common to perform this tracking and monitoring based on the bills and resolution. These are planned in agendas ahead of time, clearly numbered, and introduced in a way that makes it quite easy to track. But, it is far more challenging to try and identify common themes or topics that can span multiple bills. Being able to analyze these higher level topics can provide valuable insights .

**Analysis and Models:**

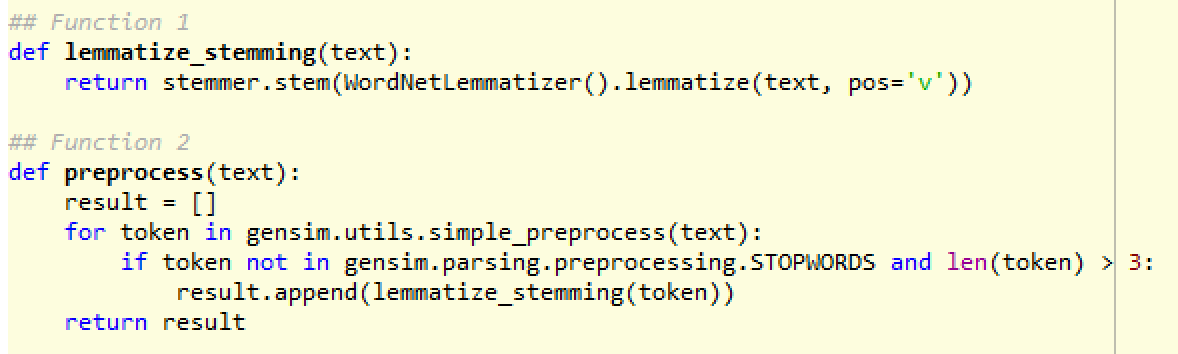
**About the Data:**

This is a large dataset comprised of more than 400 files containing the text of the proceedings from the 110th Congress (House only). The individual test files were saved in separate subfolders labeled as “m” means “male”, “f” means “female”, “d” means “democrat”, “r” means “republican”. For the purpose of this analysis all the text files were combined into one folder in order to be read in as a single corpus.

The data contained a significant number of html encoding characters (e.g. <DOC> </DOC> <TEXT> </TEXT> <b> </b> ) that needed to be removed before it could be read in. In addition, all punctuation and special characters were removed.



The data was read in to a dataframe and contained 345,539 rows and 1 columns. It was then lemmatized and stemming processing was applied.



The processed data was stored in a dictionary and several records were selected in order to view how the stemming and lemming was applied. The most common change seemed to be the removal of the word endings and converting all the words to lowercase. For example, the name “ABERCROMBIE” was reduced to “abercrombi”, “comprehensive” and “comprehension” were reduced to “comprehens”. The data was then converted to a bag of words, and TFIDF (term frequency-inverse document frequency) was applied in order to measure the frequency of each word.

**About the Models:**

The Gensim package was used to perform the Latent Dirichlet Allocation(LDA) algorithm for topic modeling. LDA uses the distribution of dominant keywords to identify the topics. The number of topics was first set to 40, and then run again for 50 topics. The number of passes was kept to 10 for both.

**lda\_model = gensim.models.LdaMulticore(bow\_corpus, num\_topics=40, id2word=dictionary, passes=10)**

**lda\_model = gensim.models.LdaMulticore(bow\_corpus, num\_topics=50, id2word=dictionary, passes=10)**

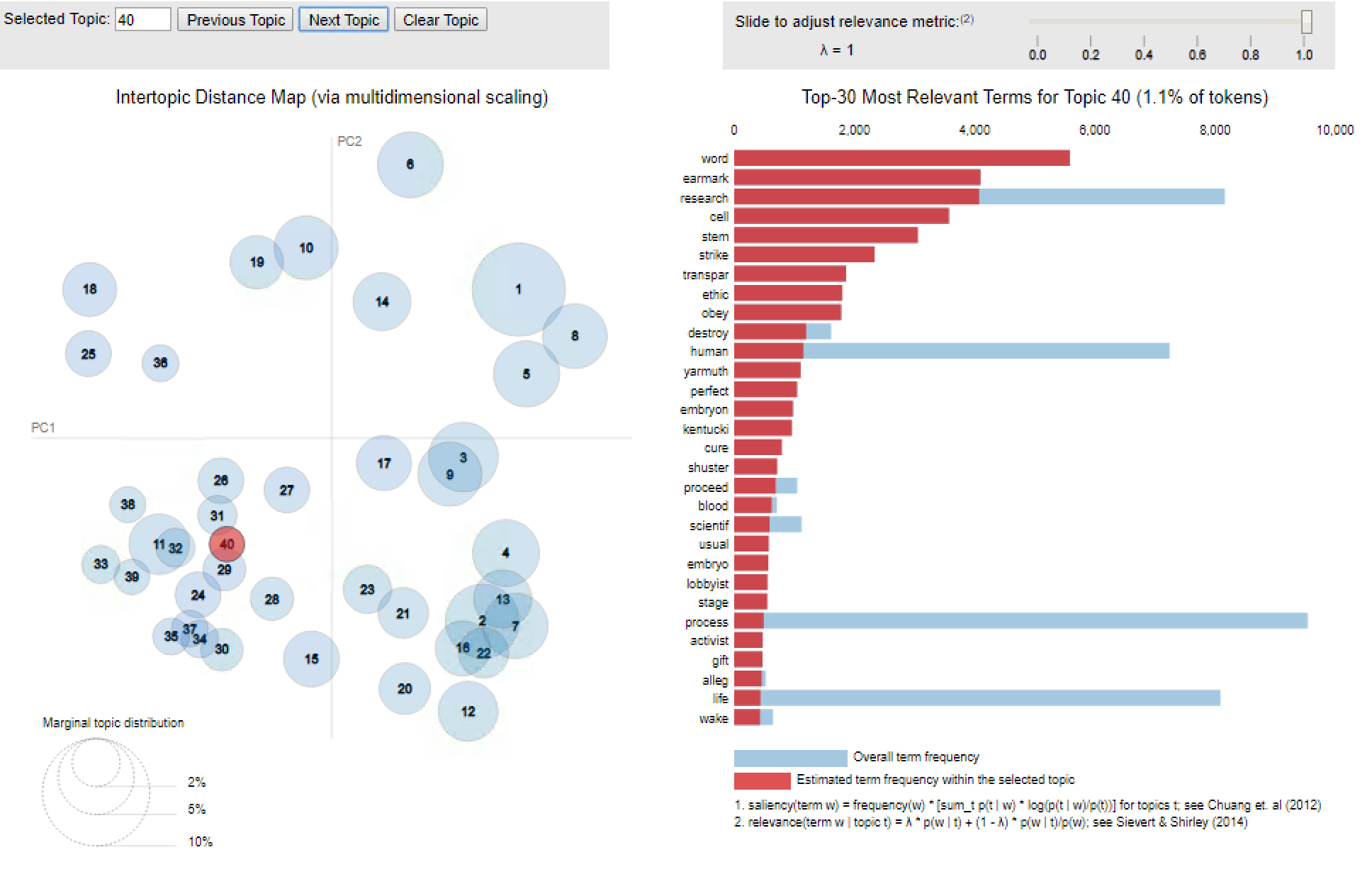
**Results**

Model perplexity and coherence score measure how good a given topic model is.

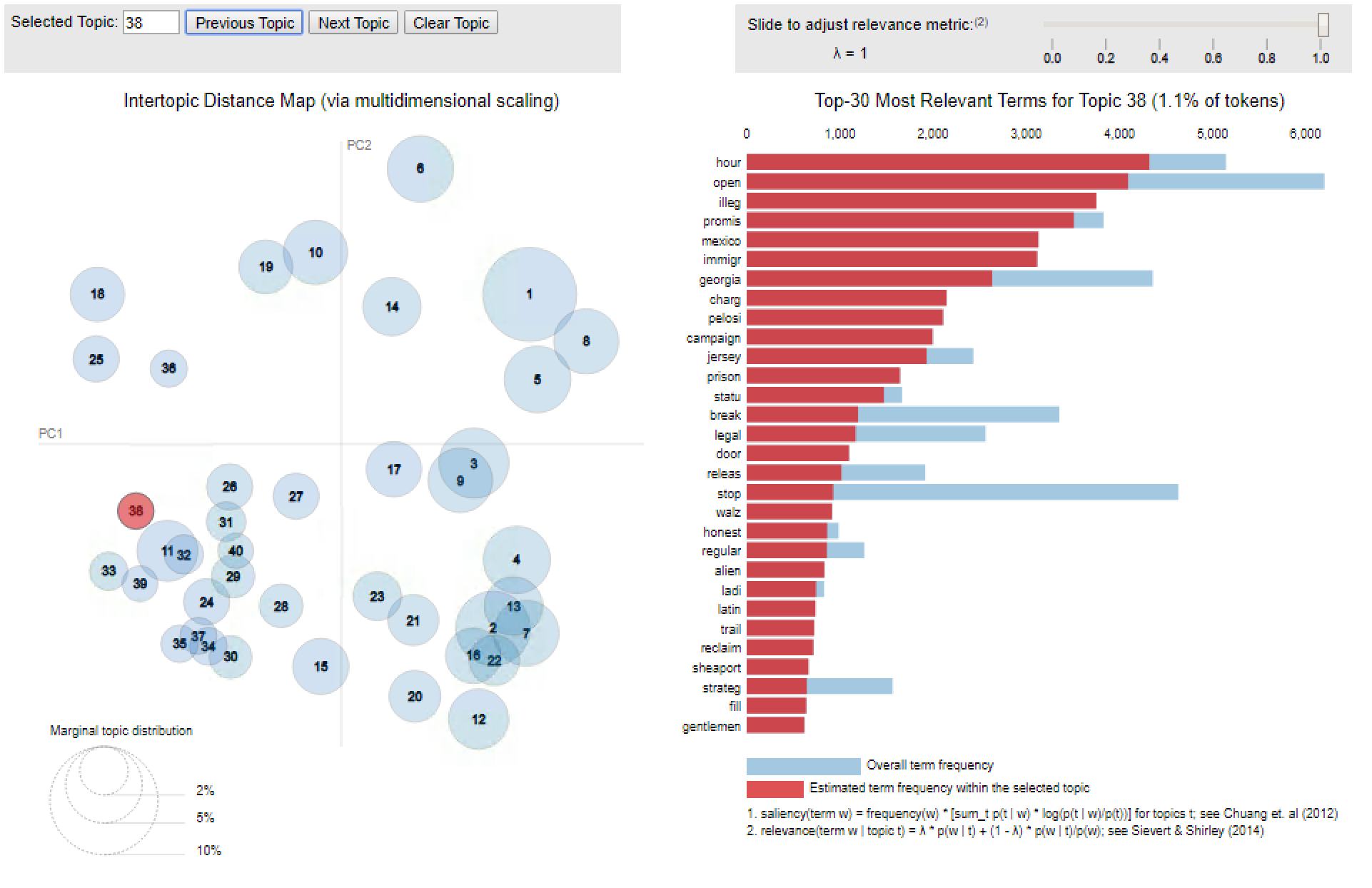
**Model 1 (N = 40):**

Perplexity: -8.876177929432412

Topic #40 appears to be about stem cell research



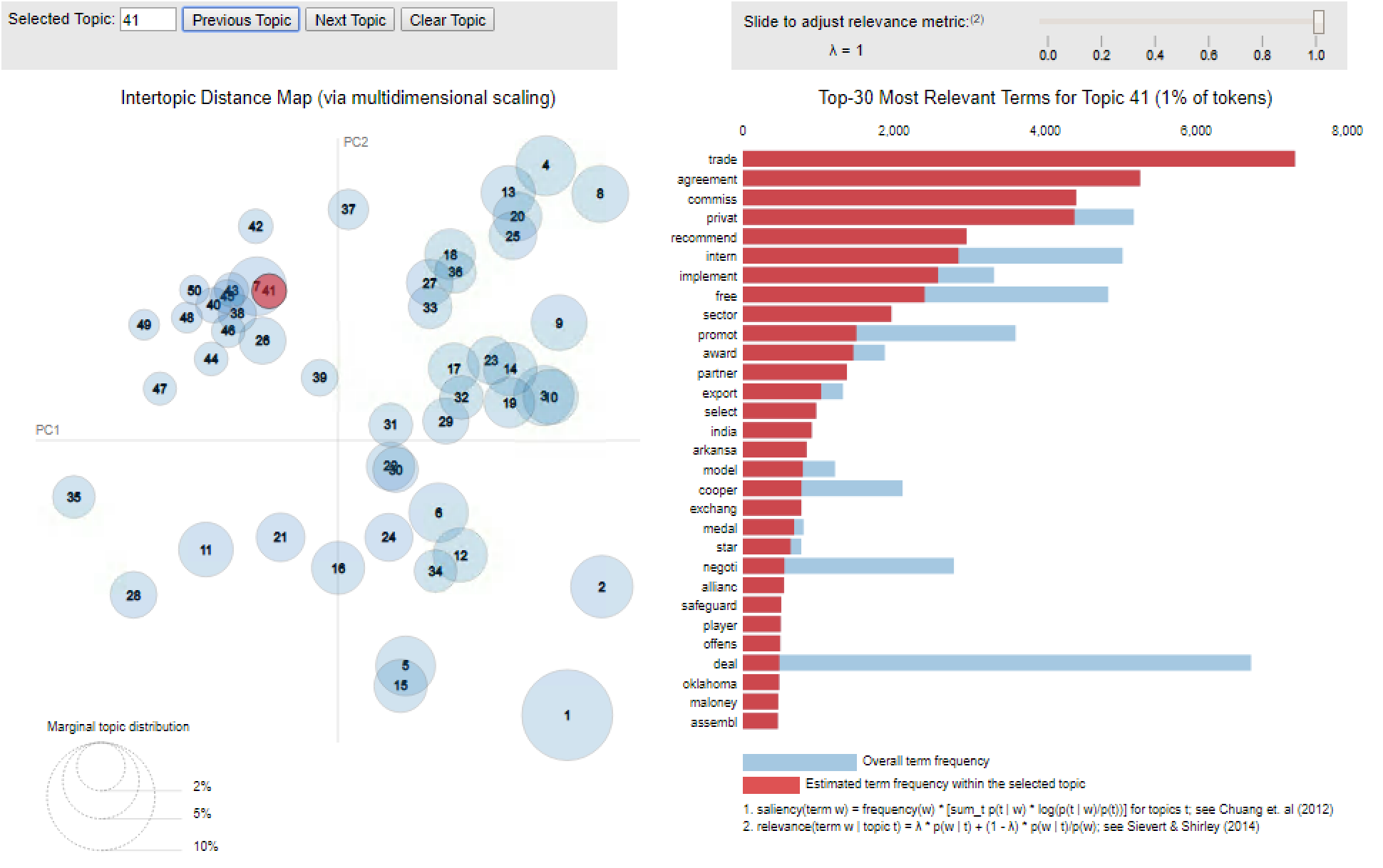
Topic #38 appears to be about illegal immigration



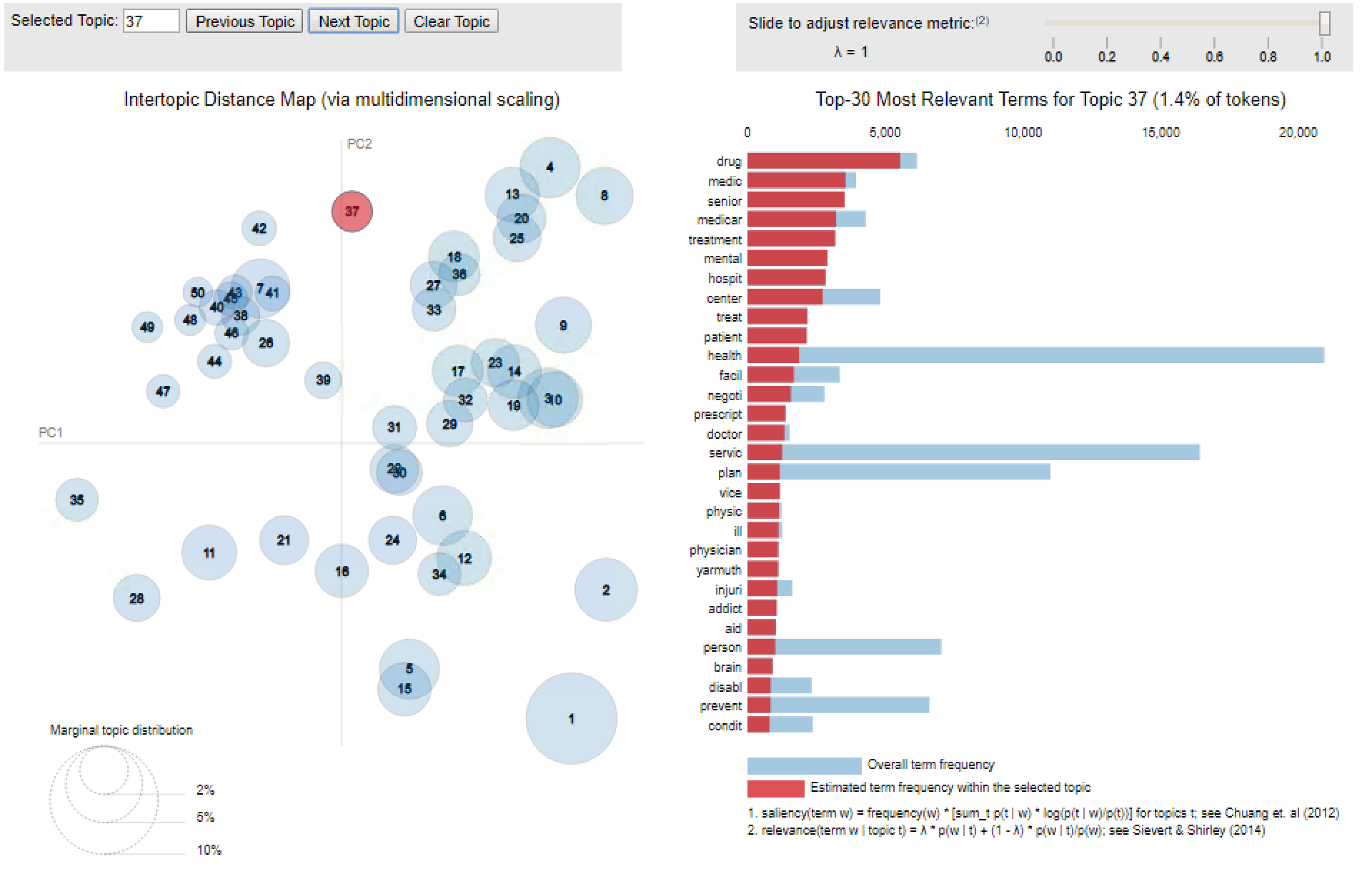
**Model 2 (N = 50):**

Perplexity: -9.28111710535581

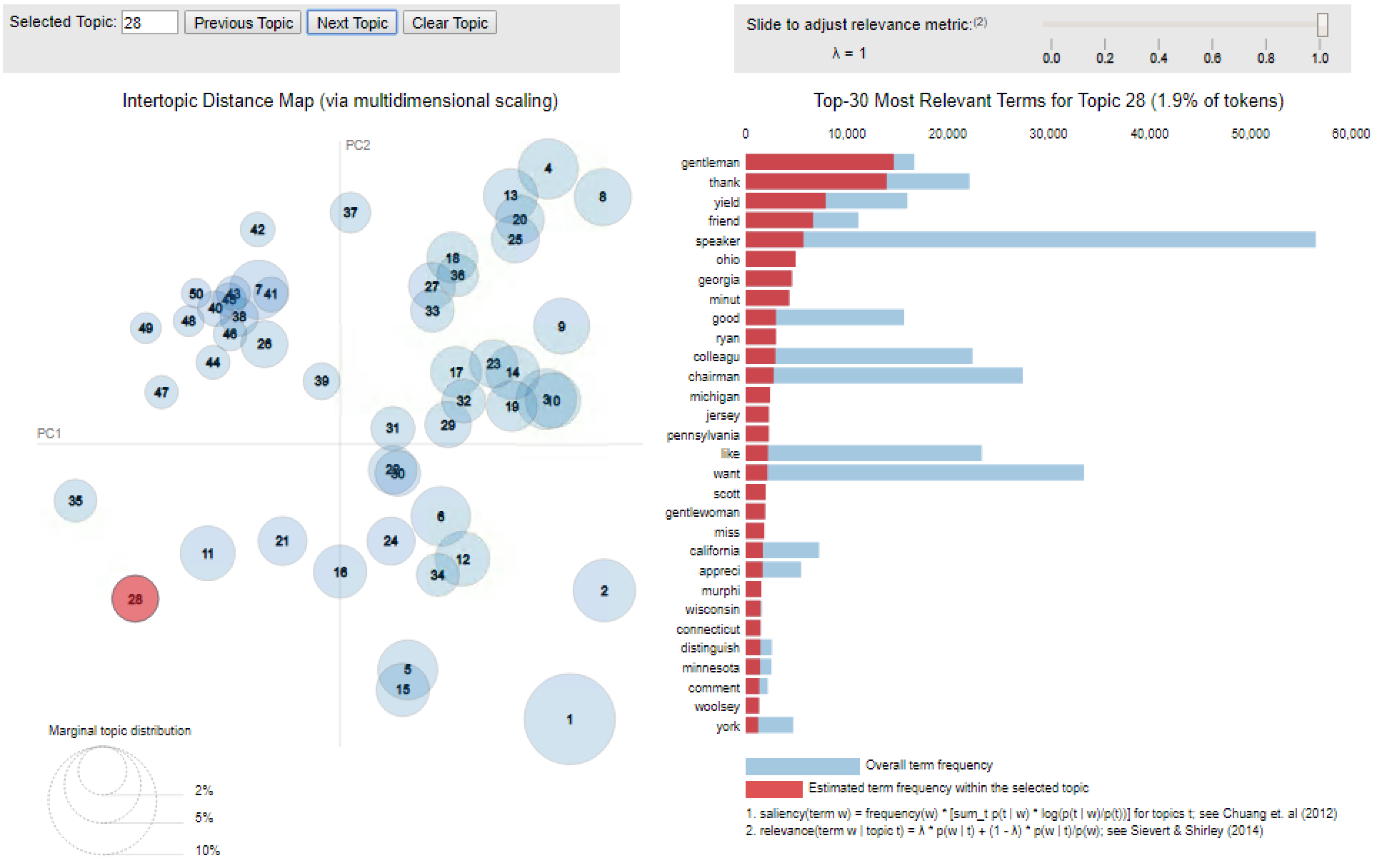
Topic # 41 appeared to be about trade agreements.



Topic #37 appears to be about Medicare healthcare.



However, several topics, such as # 28, appeared to be somewhat random, or simply a group of filler phrases or comments. “Gentlemen”, “thank”, “yield”, “friend”, “speaker” are all somewhat related terms, but don’t really seem to be a cohesive topic.



**Model Comparison:**

Setting the number of topics to 40 appeared to create more cohesive topics. When the number of topics was set to 50, there appeared to be several “extraneous” topics that did not seem to be cohesive. The model run set to 50 topics also had a larger perplexity (-9.28111710535581) than the run set to 40 topics (-8.876177929432412).

**Conclusions**

Being aware of the key issues people are talking about, and recognizing their problems and opinions is highly valuable to government officials and business professionals. Unfortunately, it is difficult and laborious to manually read through such large volumes and compile the topics. Automatic algorithms can automatically extract the most common topics people are discussing from large volumes of text.

The process of identifying these key topics from congressional proceedings could provide a tremendous advantage. Not only can it provide a more summarized view of the bill and regulations under discussion, which is often difficult to discern from just the title of the bill itself, but it can be used for a variety of other analytic purposes. And, using data science techniques offers a scalable and efficient method for uncovering insights.