Speaker Attribution and Topic Modeling

Jie Heng

DIGS 20006 / 30006, Winter 2019

Instructor: Jeffrey Tharsen

Part 1 Introduction

The task of speaker attribution refers to the annotation of a collection of spoken audios based on speaker identities (Ghaemmaghami et al., 2011). Instead of using the audio data, this study collected the transcripts of politicians' speeches to perform a similar task—given a new transcript of a speech and a trained machine learning model, the machine learning model is expected to identify the speaker's name. We are interested in examining and comparing the accuracy of those models. In addition, this study explores U.S. political parties' similarities and dissimilarities by building topic models using the politicians' speeches. The key words in the topic models might reveal the Republicans' and Democrats' focus and concerns separately. Specifically, this study addresses the following three questions:

1. Which is the machine learning model that has the highest accuracy for the speaker attribution task?

2. What are the topics of democrats and republicans' speeches? Do those key words in the topics for the two parties differ significantly?

The results of this study would expand our understanding on the differences between the two parties and selected politicians. It also contributes to studies in political communication by analyzing the two parties' speeches and providing the hot topics in political speeches.

Part 2 Data and Methods

Data

Data are collected from two sources: (1) Andrew Smith's GitHub repository[1] and (2) American Rhetoric Online Speech Bank[2]. Both sources contain the transcripts of politicians' speeches and the speakers' names. We selected five politicians for analysis. There are three democrats, Barack Obama, Hillary Clinton, and Bernie Sanders, and two republicans, Donald Trump and George W. Bush. Figure 1 shows the number of transcripts of each politician and the number of tokens for all transcripts of each politician. The total tokens for all text data are 581,170.

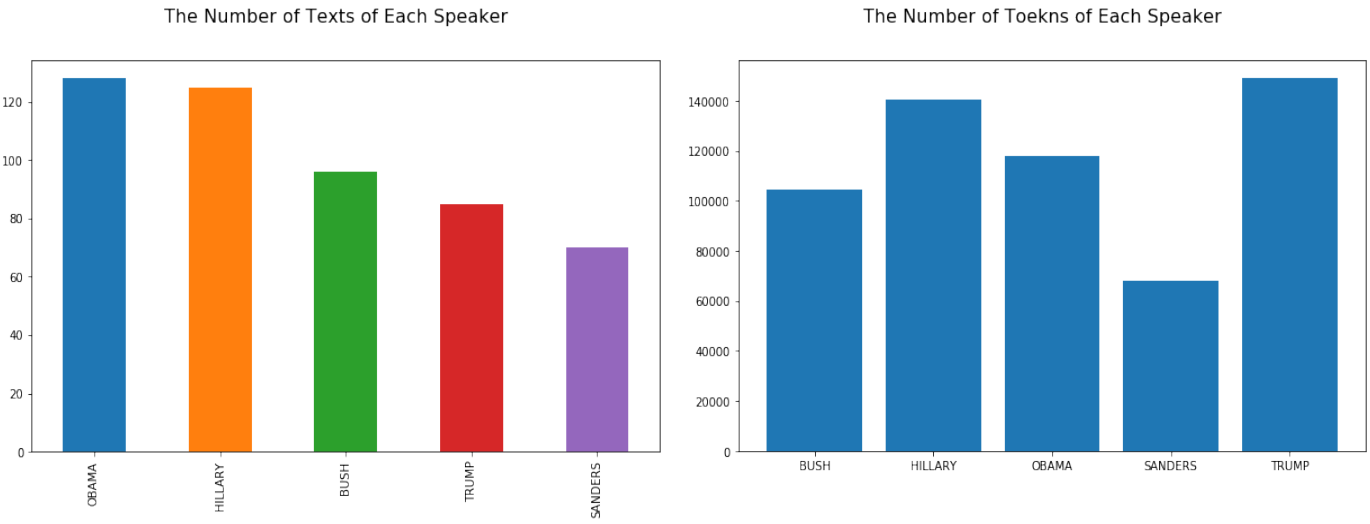


Figure 1. Text Data Information for Each Speaker

The number of texts for each party is demonstrated in figure 2. The transcripts of democrats' speeches are 323 and republicans' are 181.

[1] https://github.com/andrewts129/transcript-scraping
[2] https://www.americanrhetoric.com/speechbank.htm

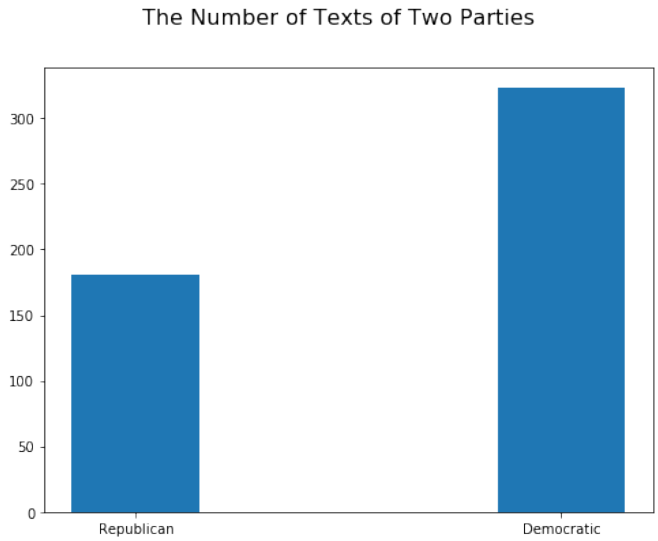


Figure 2. Text Data Information for Each Party

Preprocessing

We utilize the contraction map[3], which could covert the shortened form of one or two words to the actual phrase. For example, using the contraction map, we would able to replace "isn't" to "is not". This process is necessary to prevent the tokenizer turning contracted words to nonsense words, such as "isn" and "t". In addition, we remove the punctuations, stop words provided in NLTK and common words, such as mr., ms, and numbers.

Methods

We first use HCA and PCA to explore the data. Then we apply eight machine learning models on the speaker attribution task (see table 1) and visualize topics using pyLDAvis.gensim.

[3] https://github.com/dipanjanS/text-analytics-with-python/blob/master/Old-First-Edition/source_code/Ch04_Text_Classification/normalization.py

Part 3 Results

EDA

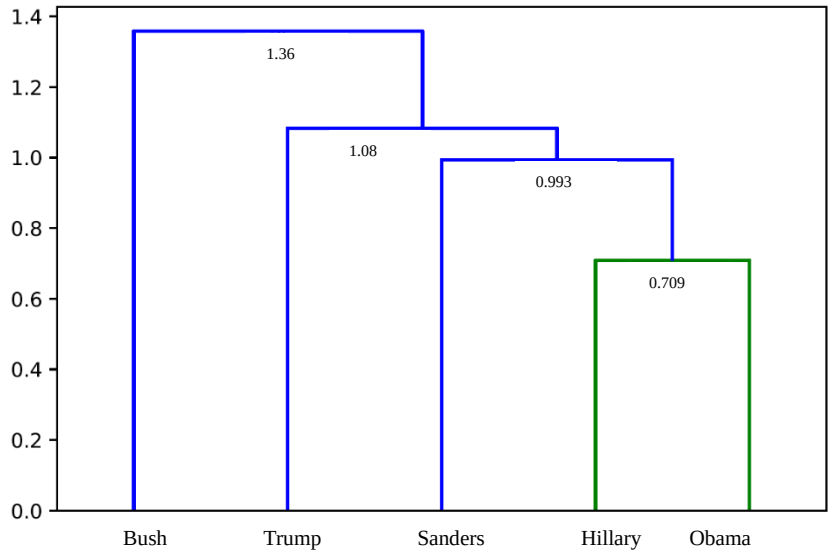


Figure 3. HCA

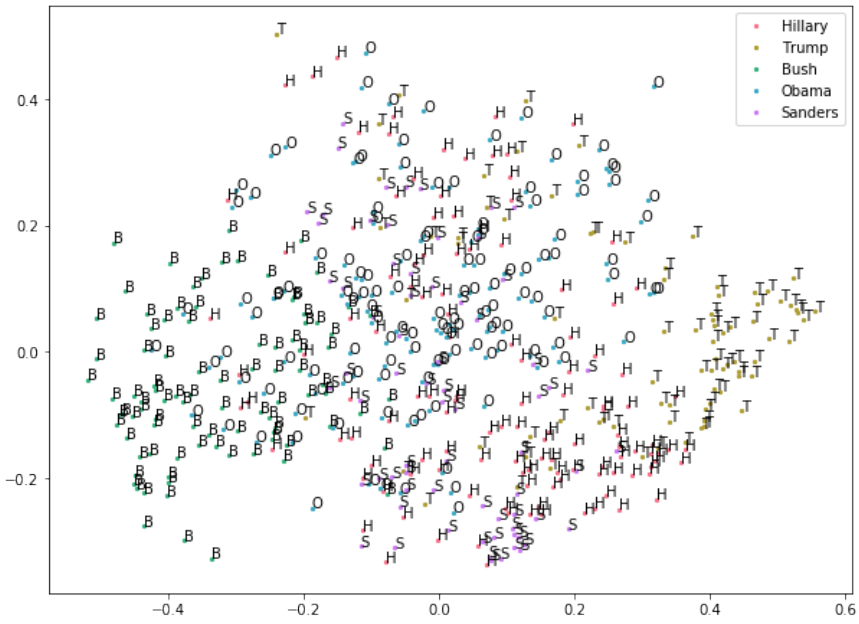


Figure 4. PCA

Figure 3 and 4 shows the HCA and PCA results. They both show an apparent difference between the two parties. Bush and Trump's text data are quite different from the other three

democrats in figure 3 and in figure 4, the two republicans' transcripts are on the far left and the far right, far from the democrats' speeches in the middle. The topic modeling results would reveal this difference as well. However, the two graphs also illustrate the similarities among the Hillary, Obama and Sanders, which might increase the difficulties for speaker attribution tasks.

Speaker Attribution

Model	Accuracy
Markov (k=2)	94.08
SVM	92.76
BoW(word+character)	88.24
BoW(word)	86.27
Naive Bayesian	78.29
KNN	72.55
BoW(character)	84.31
Maximum Entropy	50

Table 1 Machine Learning Approaches on Speaker Attribution

Eight models are utilized to perform the speaker attribution task: the 2nd-order Markov model, support-vector networks (SVM), three bag-of-words models (based on words, based on characters and based on both word and characters), Naïve Bayesian, KNN and maximum entropy. These models are commonly used in text classification and speaker attribution. As the table 1 shows, the Markov model performs best and the SVM model performs fairly well, with more than 90% accuracy; whereas the maximum entropy model performs the worst, with only 50% accuracy.

Although Chomsky (Hale, 2016) criticized the Markov model that in real human language, words influence each other at arbitrary distances in a way that Markov models cannot properly capture, the result shows that Markov model does capture the linguistic features of each speakers. It might because unlike novels and stories, the speeches need to be comprehended in a

short time and do not have many complicated relative clauses, which do not require a long-distance dependency. Thus, the Markov model is able to build linguistic relationships among adjacent words and capture the linguistic features of each speaker.

In most cases, neural networks require a large number of training data to find patterns and achieve good performance on tasks. Given more training data, SVM might have a better performance.

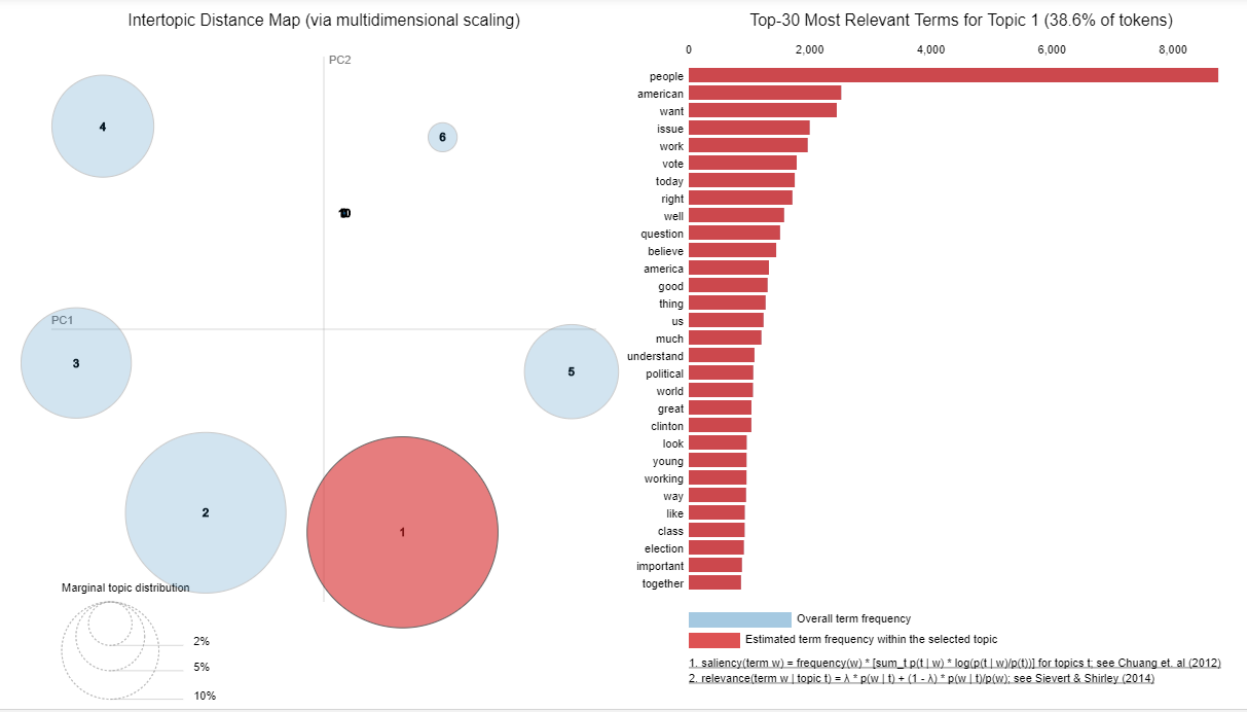
Topic Modeling

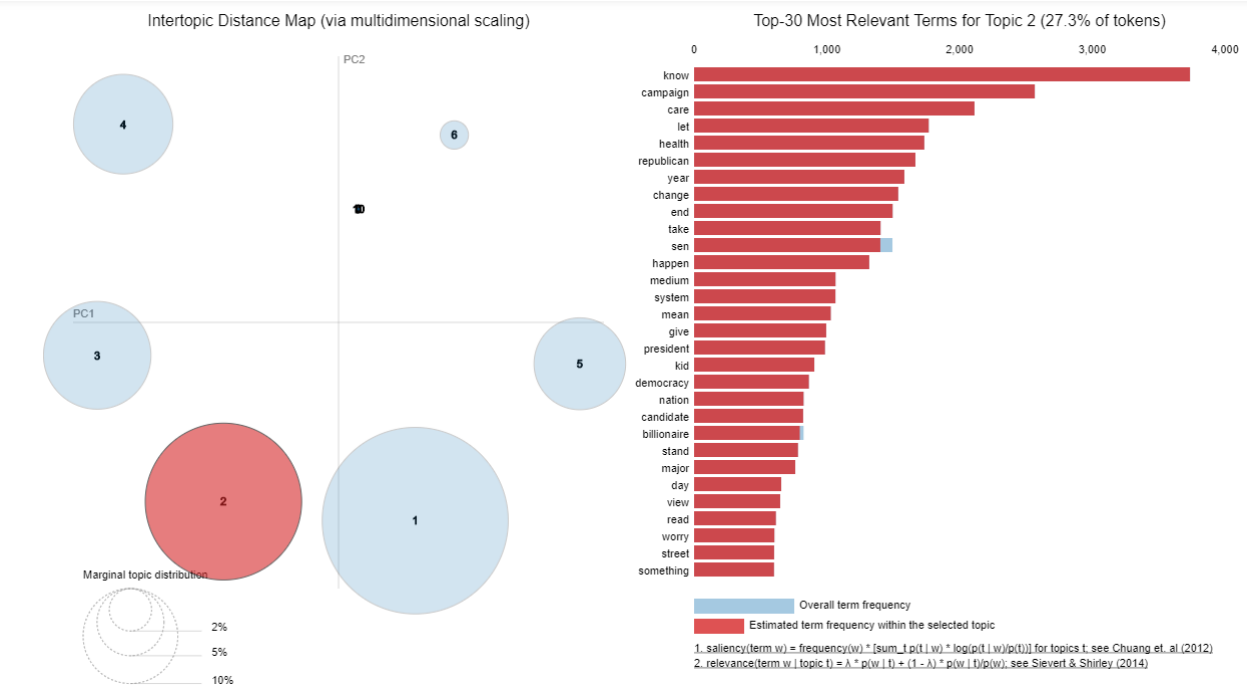
The Appendix A and B present four popular topic for each political party separately. The Democrats are more likely to talk about healthcare, employment and women. The Republicans, however, pay more attention to immigrants (words such as immigration, border and wall) and family issues. In general, both models address presidential elections. The results accord with our knowledge of Republicans and Democrats that Republicans tend to take a more conservative attitude. The topics of republicans are significantly influenced by Trump. His slogan ("make America great again") and his emphasis on immigration are reflected in the topic models.

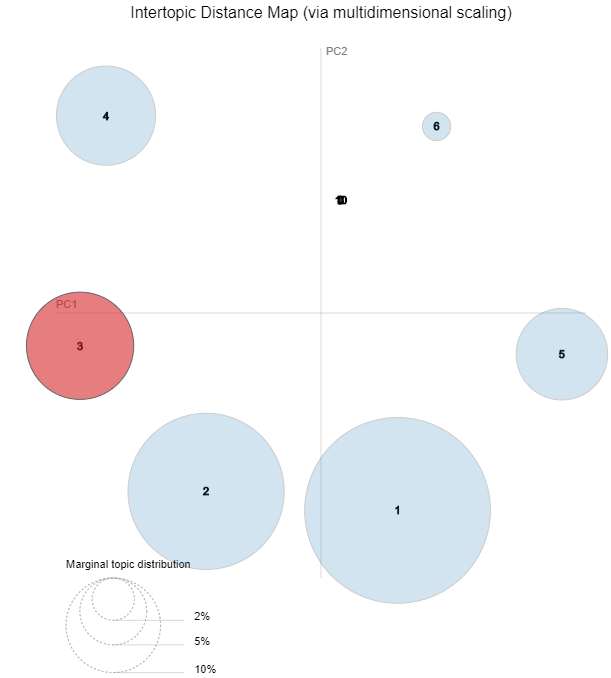
Part 4 Discussion

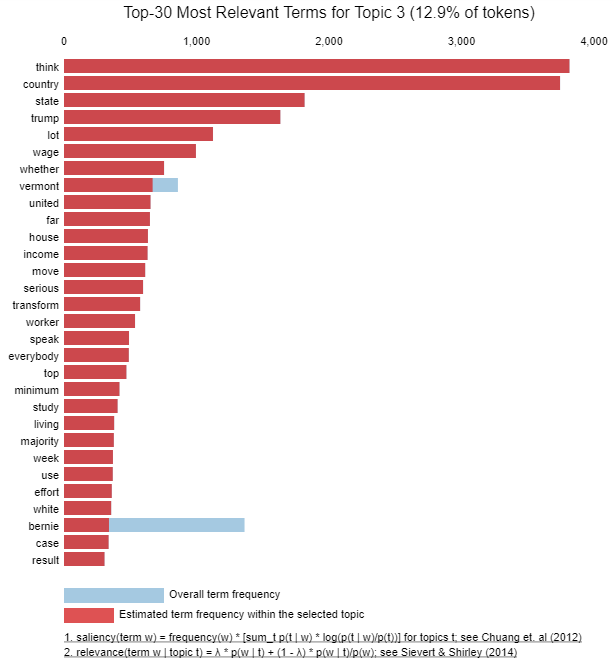
The results show that the Markov model and SVM performs best for the speaker attribution tasks. And the differences between the two parties are quite obvious in their topics of speeches. Future studies could include speeches in different decades to study the changes of topics in politicians' speeches and analyze changes in popular topics for the two parties.

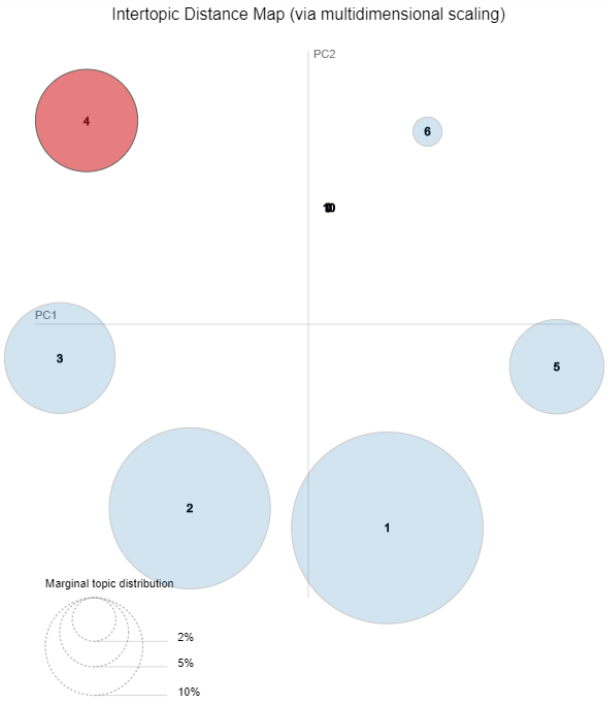
Appendix A Topics for Democrats

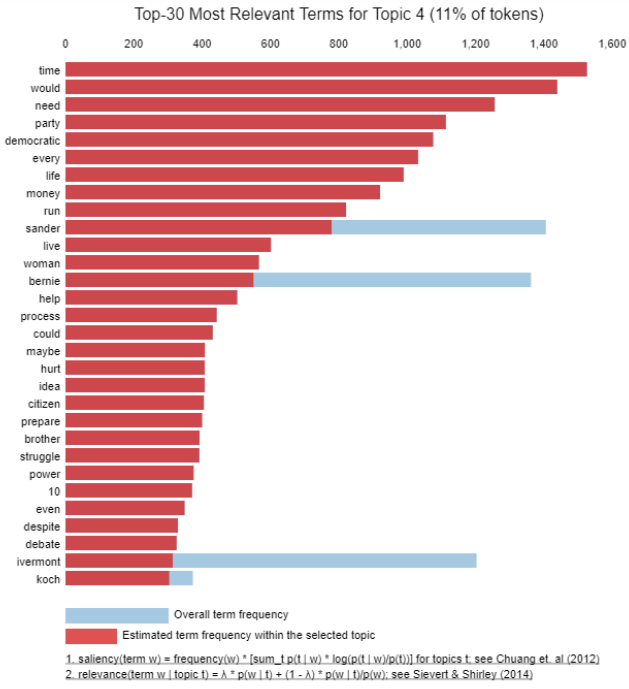




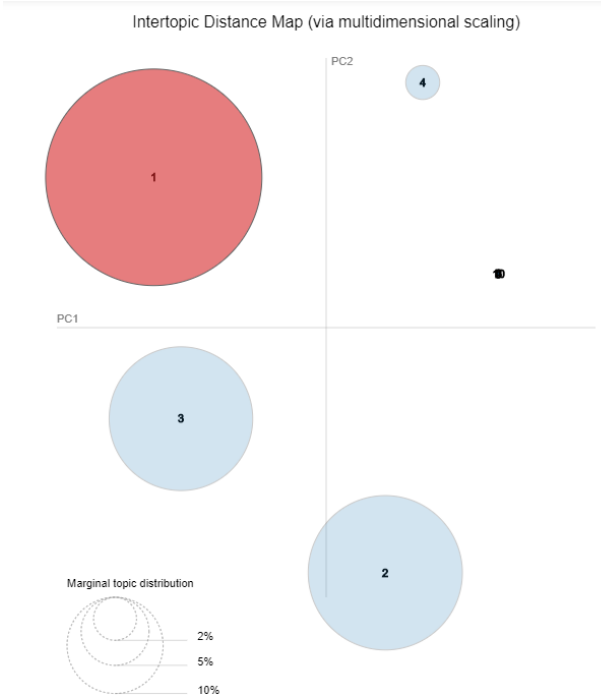

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
6
3
5
2
1
Marginal topic distribution
2%
5%
10%

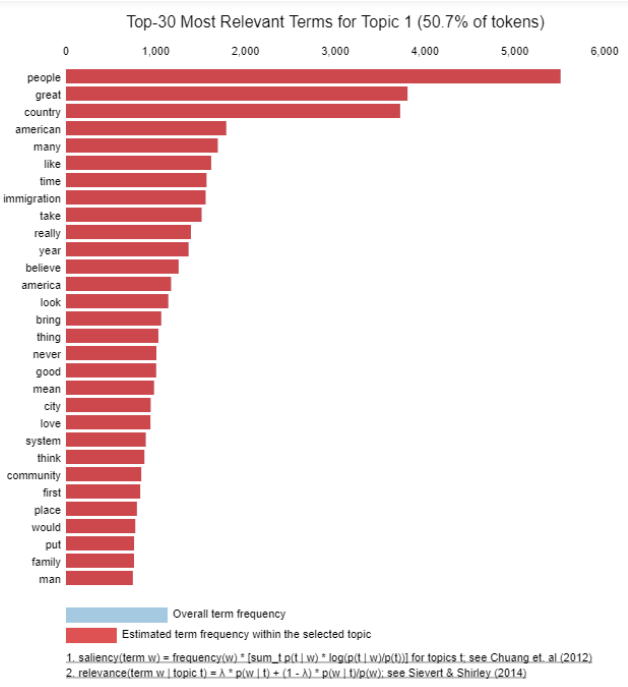

Top-30 Most Relevant Terms for Topic 3 (12.9% of tokens)
0
1,000
2,000
3,000
4,000
think
country
state
trump
lot
wage
whether
vermont
united
far
house
income
move
serious
transform
worker
speak
everybody
top
minimum
study
living
majority
week
use
effort
white
bernie
case
result
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

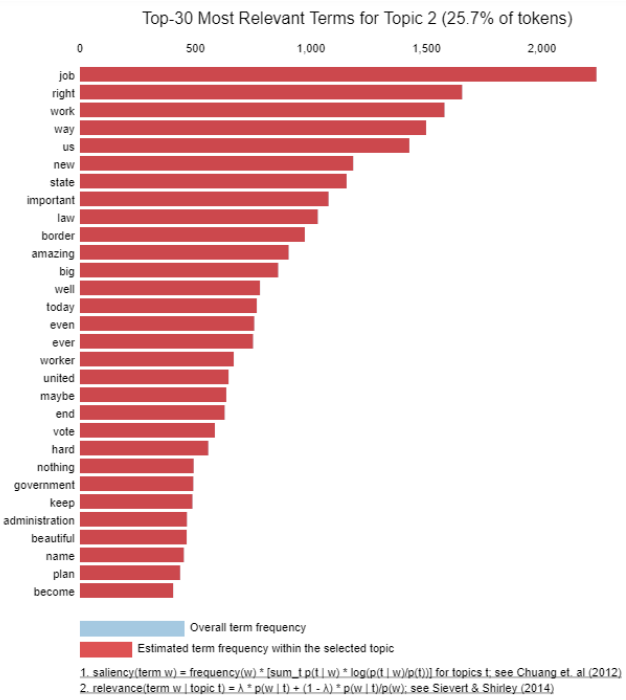

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
6
3
5
2
1
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 4 (11% of tokens)
0
200
400
600
800
1,000
1,200
1,400
1,600
time
would
need
party
democratic
every
life
money
run
sander
live
woman
bernie
help
process
could
maybe
hurt
idea
citizen
prepare
brother
struggle
power
10
even
despite
debate
ivermont
koch
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

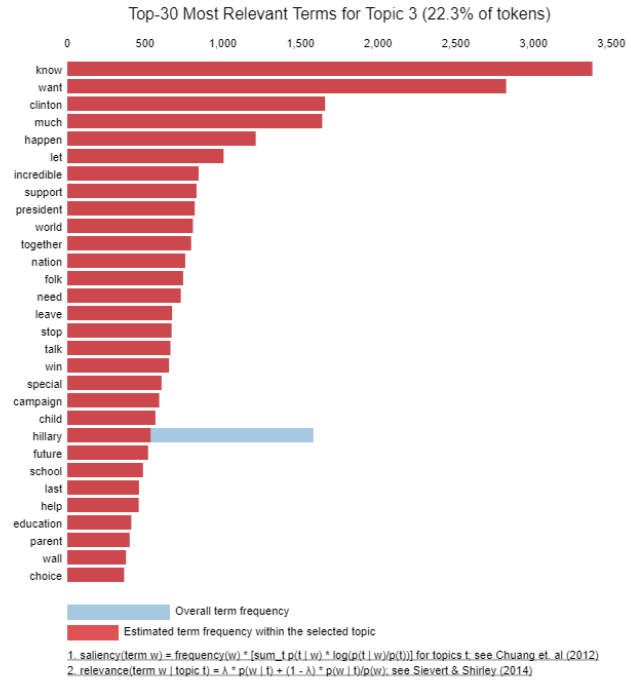
Appendix B Topics for Republicans

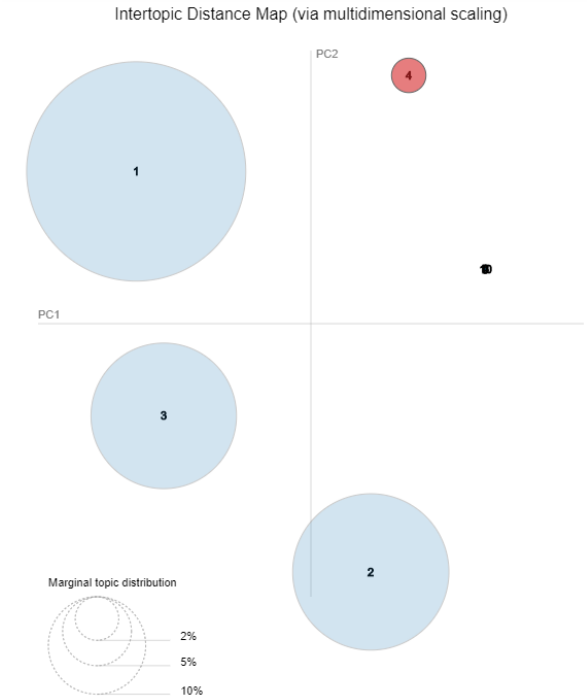


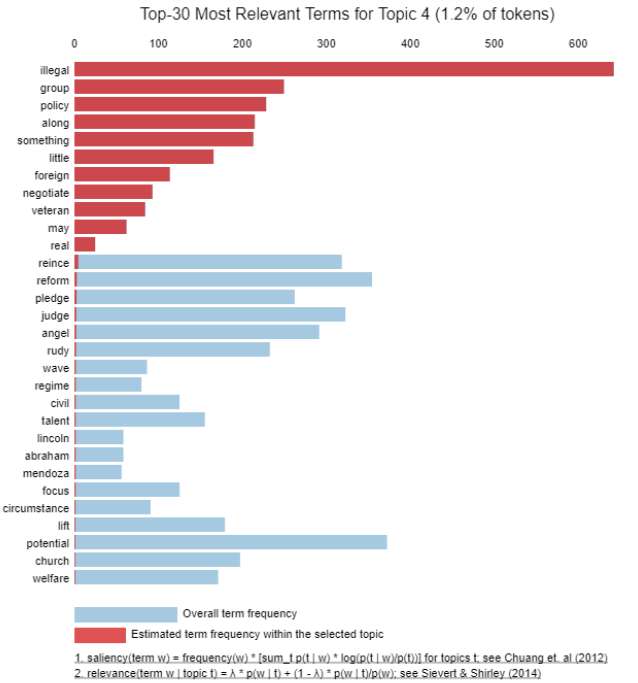





Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 3 (22.3% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
3,500
know
want
clinton
much
happen
let
incredible
support
president
world
together
nation
folk
need
leave
stop
talk
win
special
campaign
child
hillary
future
school
last
help
education
parent
wall
choice
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 4 (1.2% of tokens)
0
100
200
300
400
500
600
illegal
group
policy
along
something
little
foreign
negotiate
veteran
may
real
reince
reform
pledge
judge
angel
rudy
wave
regime
civil
talent
lincoln
abraham
mendoza
focus
circumstance
lift
potential
church
welfare
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Reference

Ghaemmaghami, Houman, David Dean, Robbie Vogt, and Sridha Sridharan. "Extending the task of diarization to speaker attribution." In *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

Hale, John. "Information-theoretical complexity metrics." *Language and Linguistics Compass 10*, no. 9 (2016): 397-412.