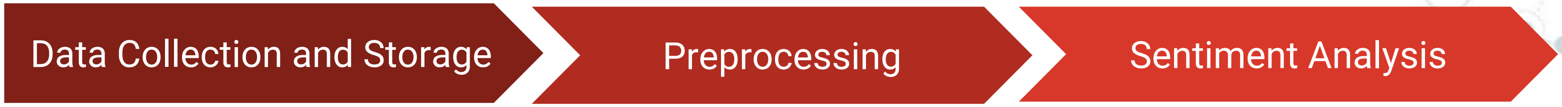
Topic Modeling of Tweets: #ElectionNight #MidTerms 2018

HIRANYA KRISHNA KUMAR

16th Nov 2018

PIPELINE



Data Collection and Storage

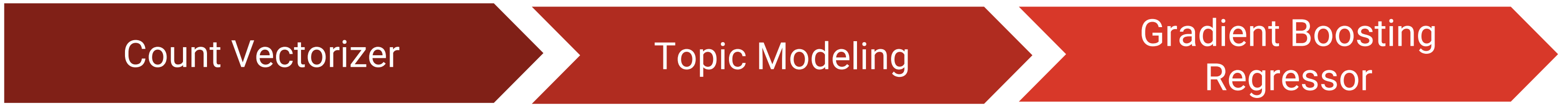
Preprocessing

Sentiment Analysis

Tweepy :Twitter search
API
MongoDB: 70,029 tweets

Clean data, remove stop words , stemming, tokenization

Tweet Polarity, Subjectivity



Count Vectorizer

Topic Modeling

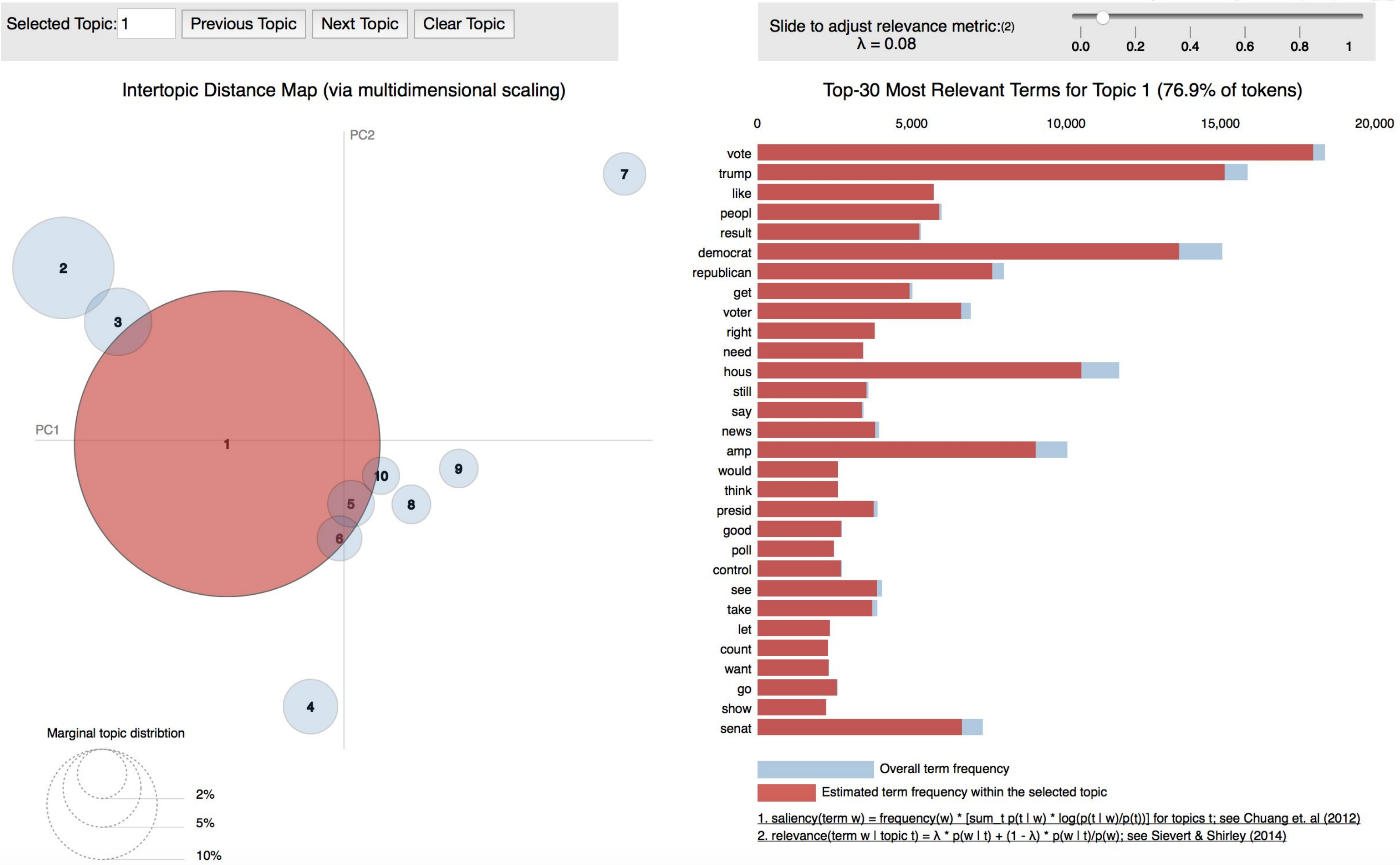
Gradient Boosting Regressor

Term Frequencies , unigrams, bigrams, trigrams

LDA-10 topics
Dimensionality Reduction

Predict retweet counts
Features:
LDA doc_topic probability, polarity,subjectivity

GENERAL TOPIC -HOUSE, SENATE ELECTIONS



Selected Topic: 1

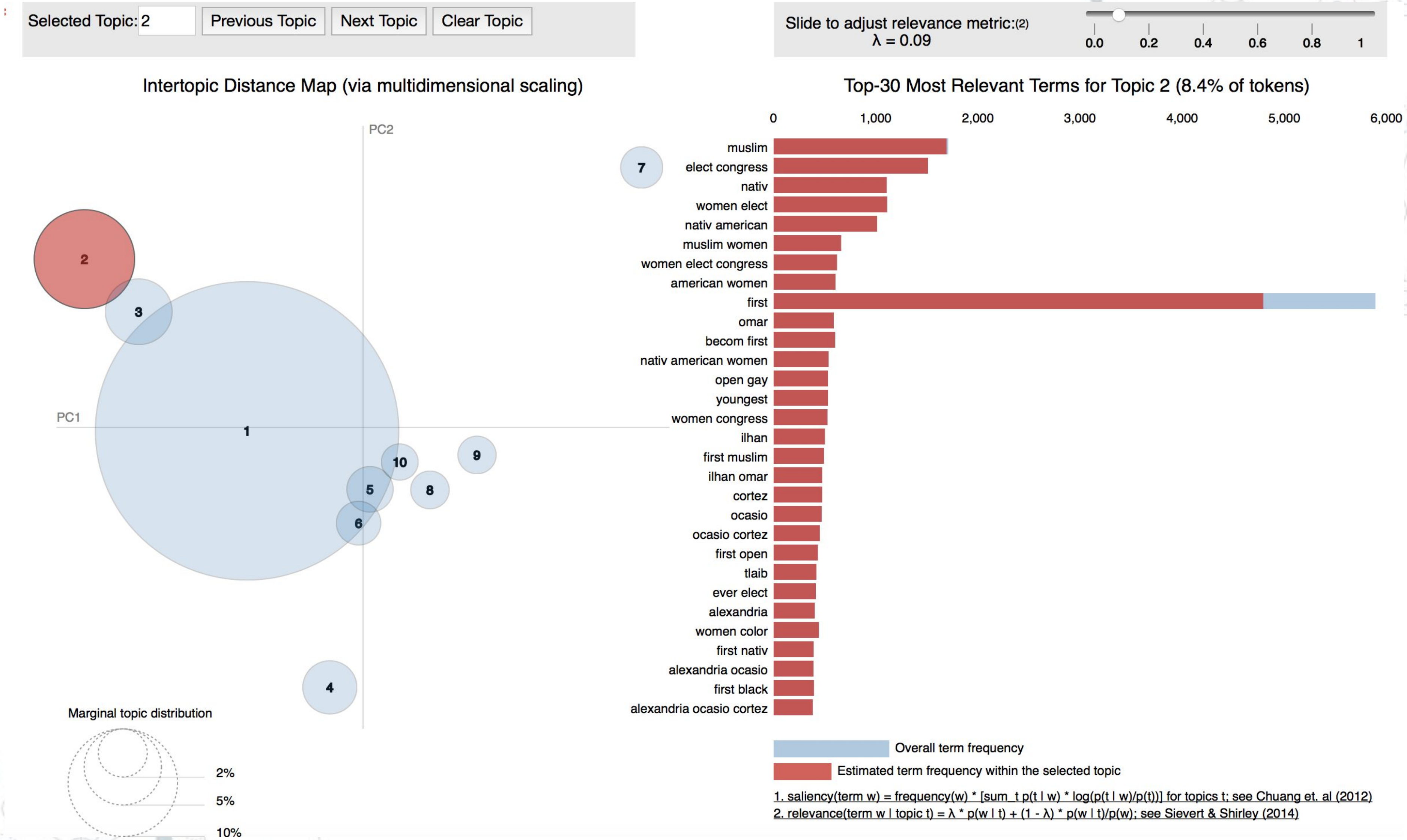
Slide to adjust relevance metric:(2)
λ = 0.08

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 1 (76.9% of tokens)

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)

HISTORIC WINS



Selected Topic: 2

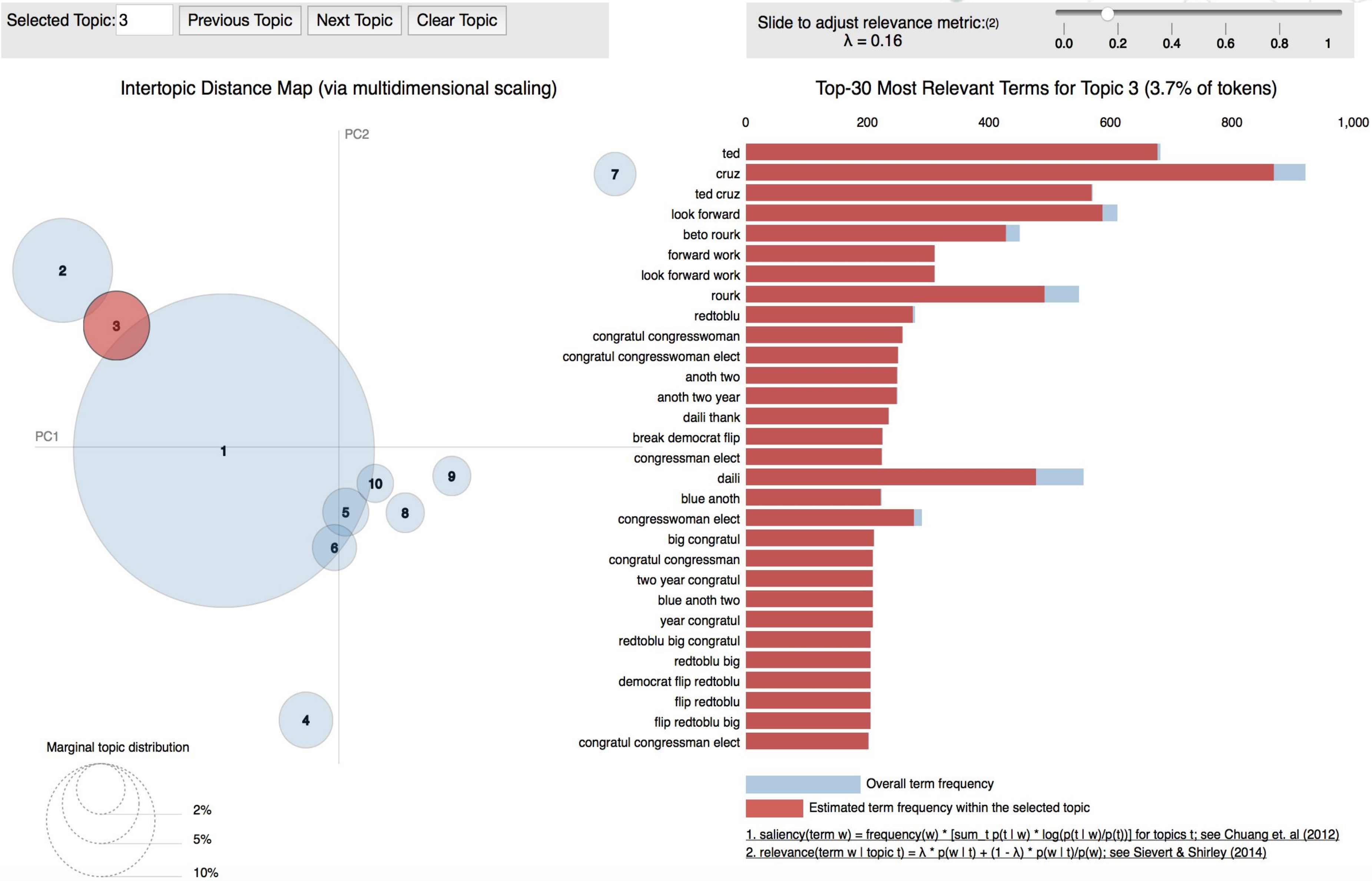
Slide to adjust relevance metric:(2)
λ = 0.09

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 2 (8.4% of tokens)

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)

DEMOCRATS FLIPPING HOUSE, TEXAS SENATE



Selected Topic: 3

Slide to adjust relevance metric:(2)
λ = 0.16

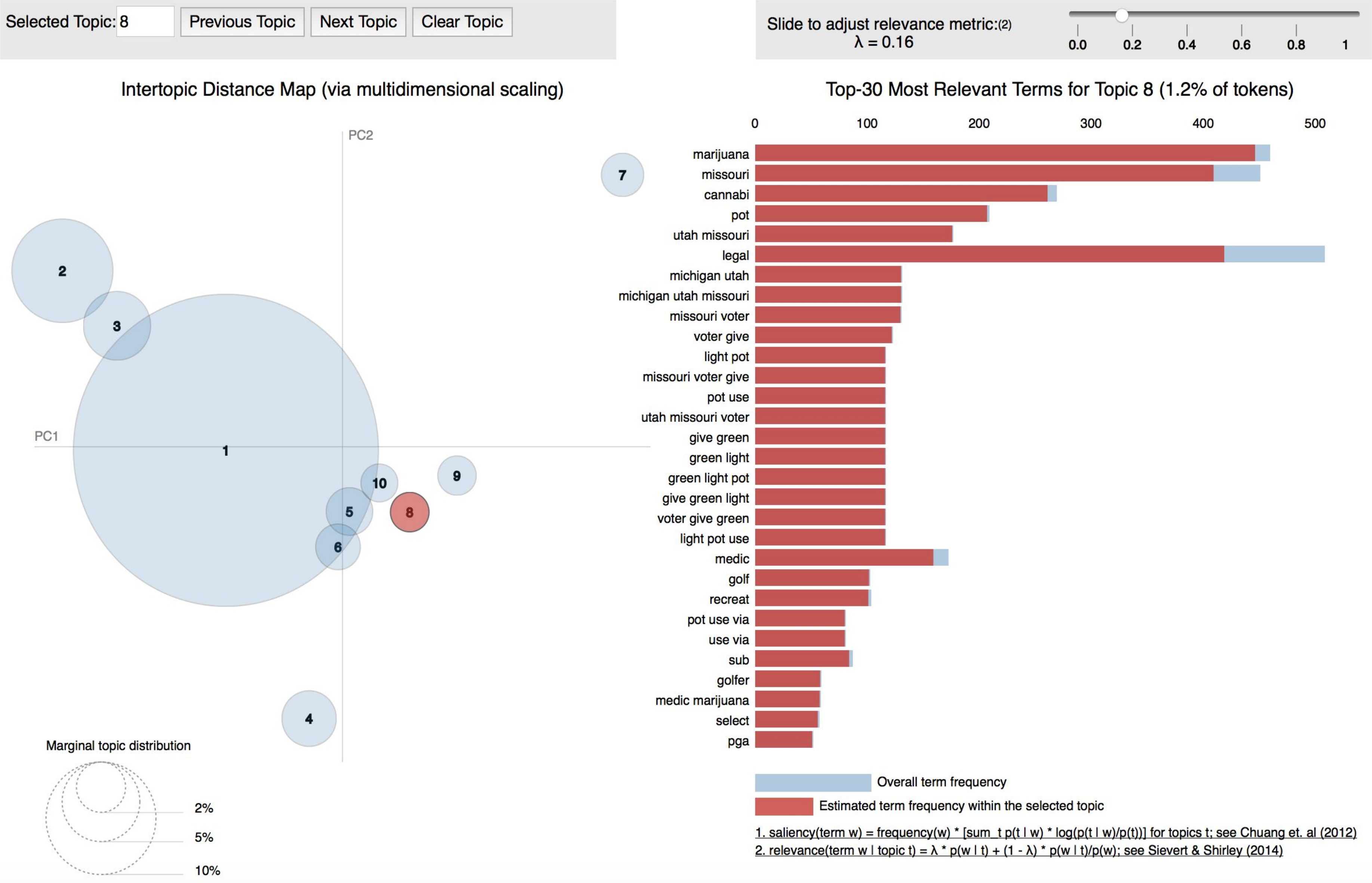
Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 3 (3.7% of tokens)

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)

STATES LEGALIZING MARIJUANA

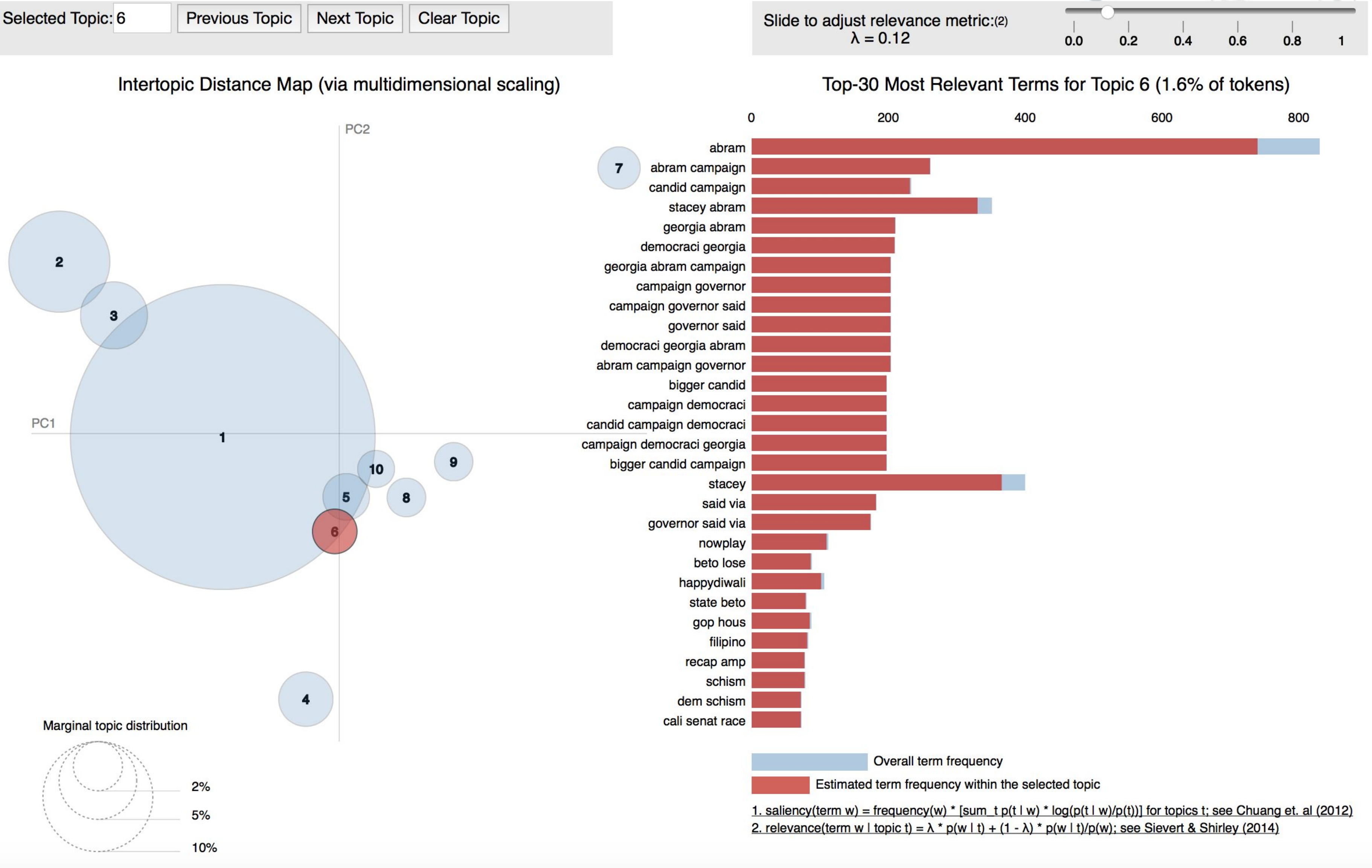


Selected Topic: 8

Slide to adjust relevance metric:(2)
λ = 0.16

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)

GOVERNOR ELECTIONS IN GEORGIA



Selected Topic: 6

Slide to adjust relevance metric:(2)
λ = 0.12

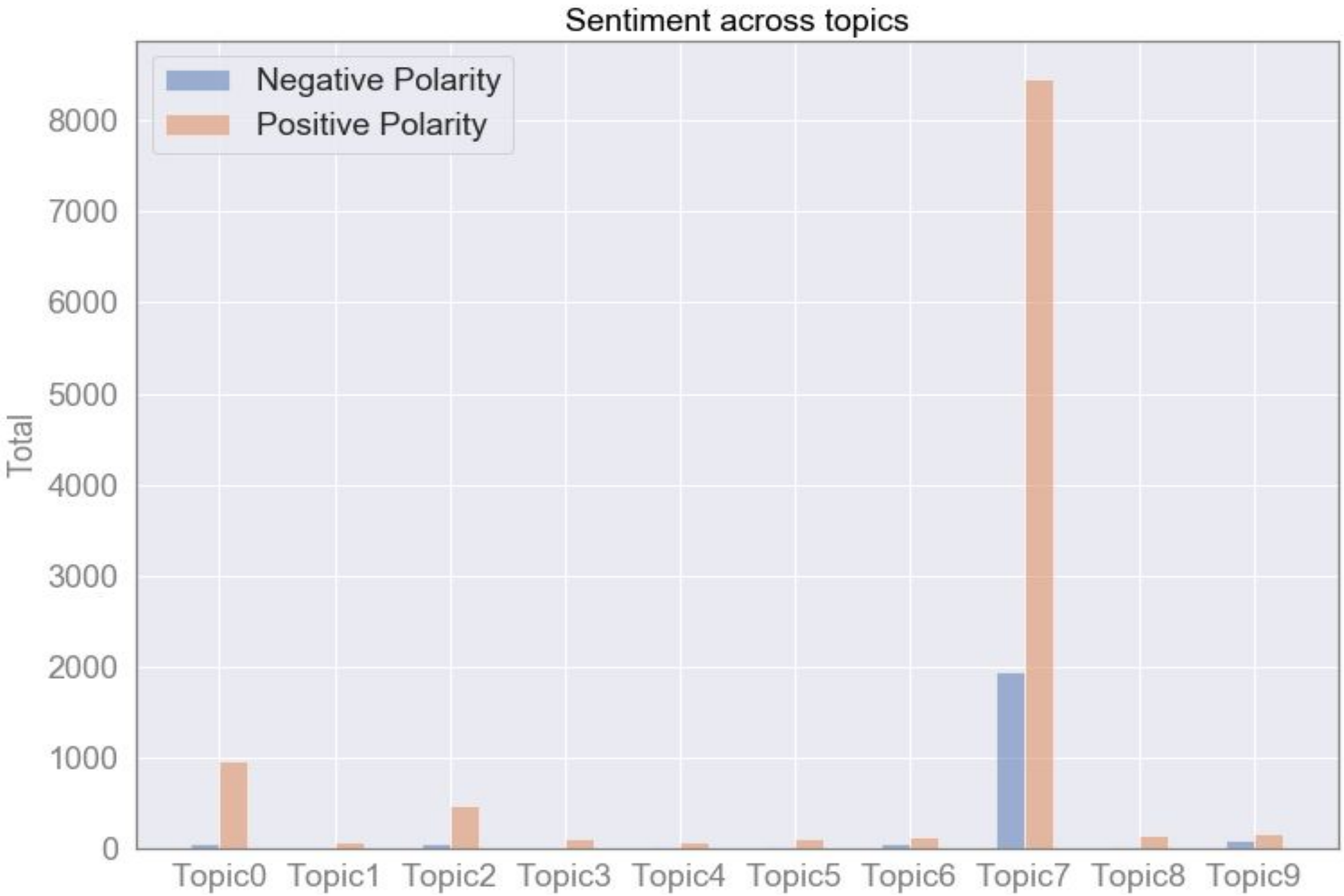
Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 6 (1.6% of tokens)

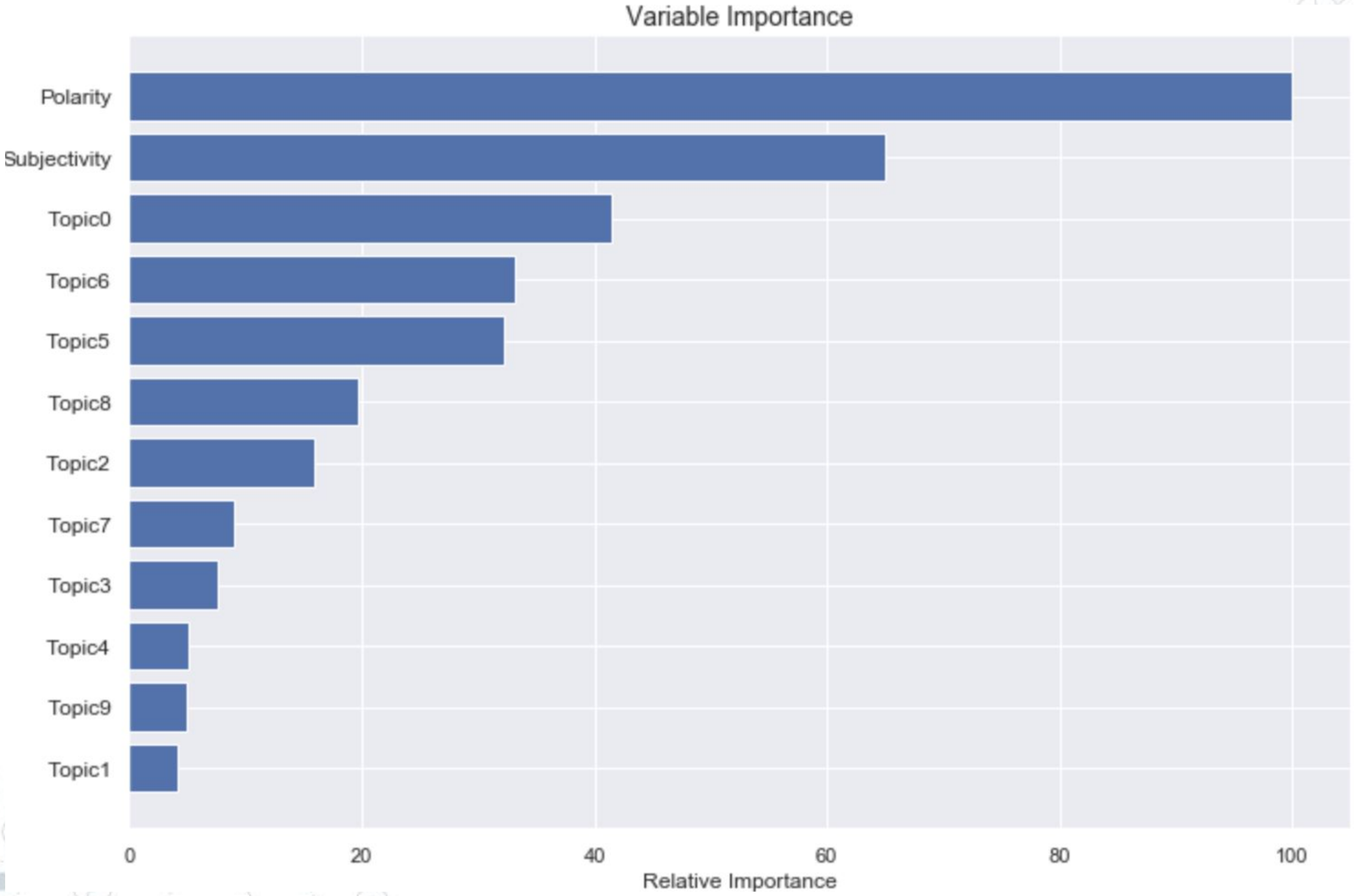
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

SENTIMENT ANALYSIS

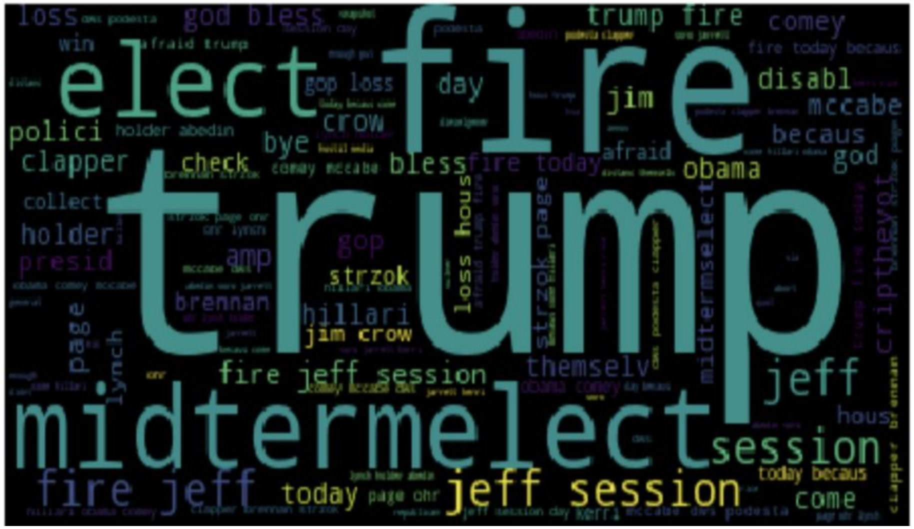


GRADIENT BOOSTING REGRESSOR FEATURE IMPORTANCE TO PREDICT RETWEET COUNTS

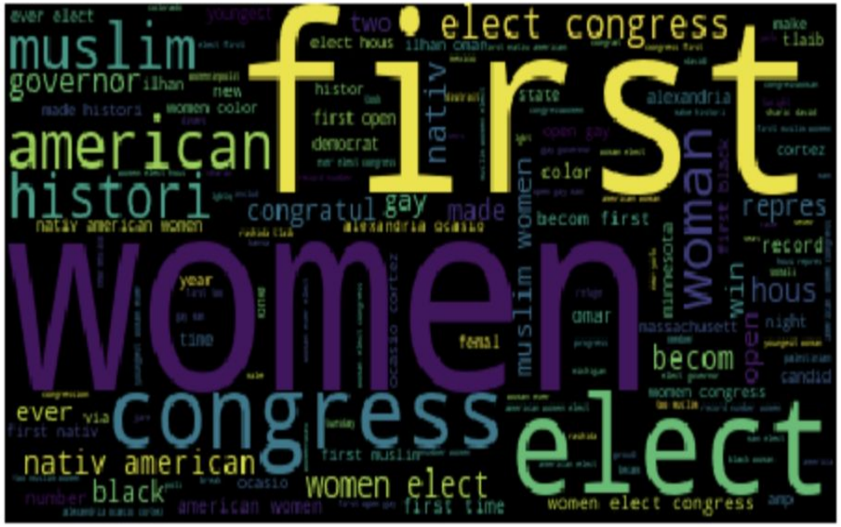


WORD CLOUD OF TOPICS

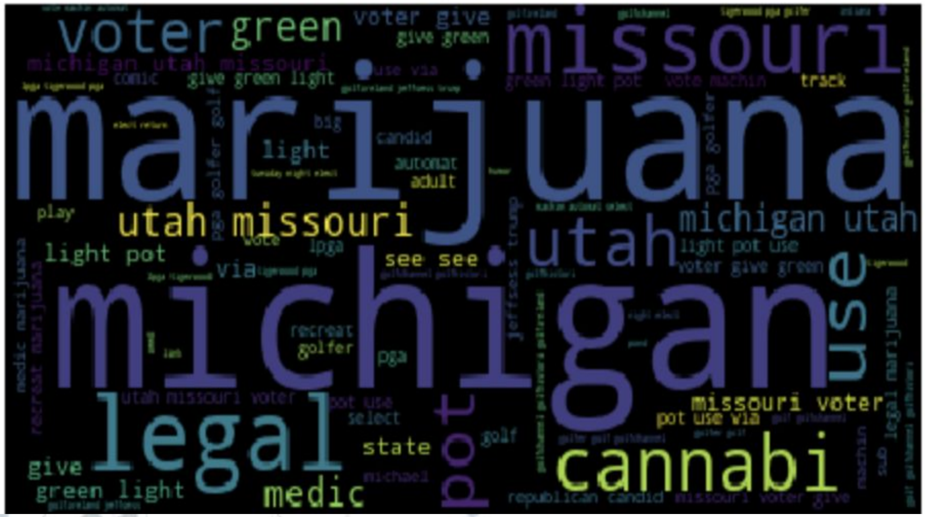
TOPIC 6-TRUMP



TOPIC 0-HISTORIC WINS



TOPIC 5-STATES VOTING TO LEGALIZE MARIJUANA



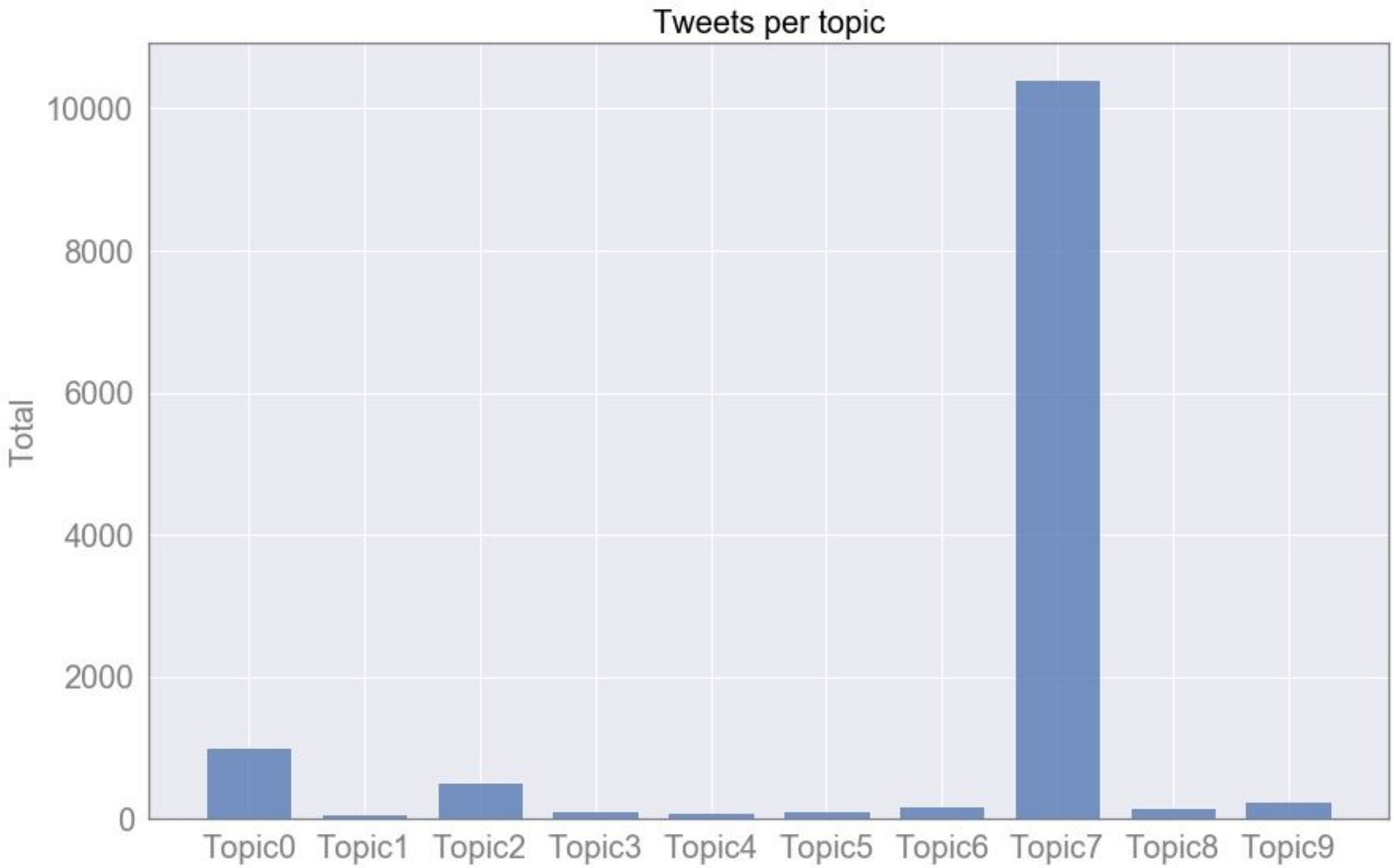
FUTURE WORK

1. Use a different pipeline like word2vec and k-means clustering to identify topics and most frequent words in tweets
2. Gather more tweets and use guided LDA to get subtopics within a topic
3. Tune hyperparameters like doc_topic_prior,evaluate_topics frequency,word_topic_prior
4. Build a flask app to predict a topic for a tweet on midterm elections

Thanks!

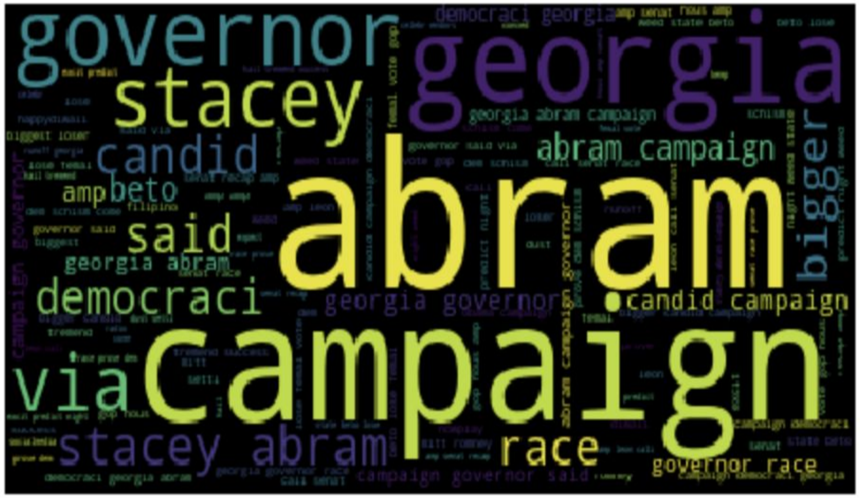
Any questions?

FUTURE WORK

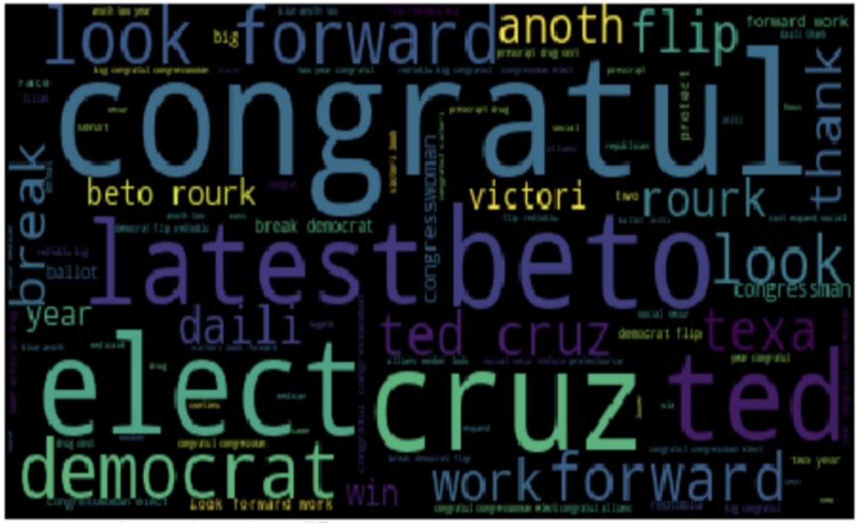

Tweets per topic
Total
10000
8000
6000
4000
2000
0
Topic0
Topic1
Topic2
Topic3
Topic4
Topic5
Topic6
Topic7
Topic8
Topic9

WORD CLOUD

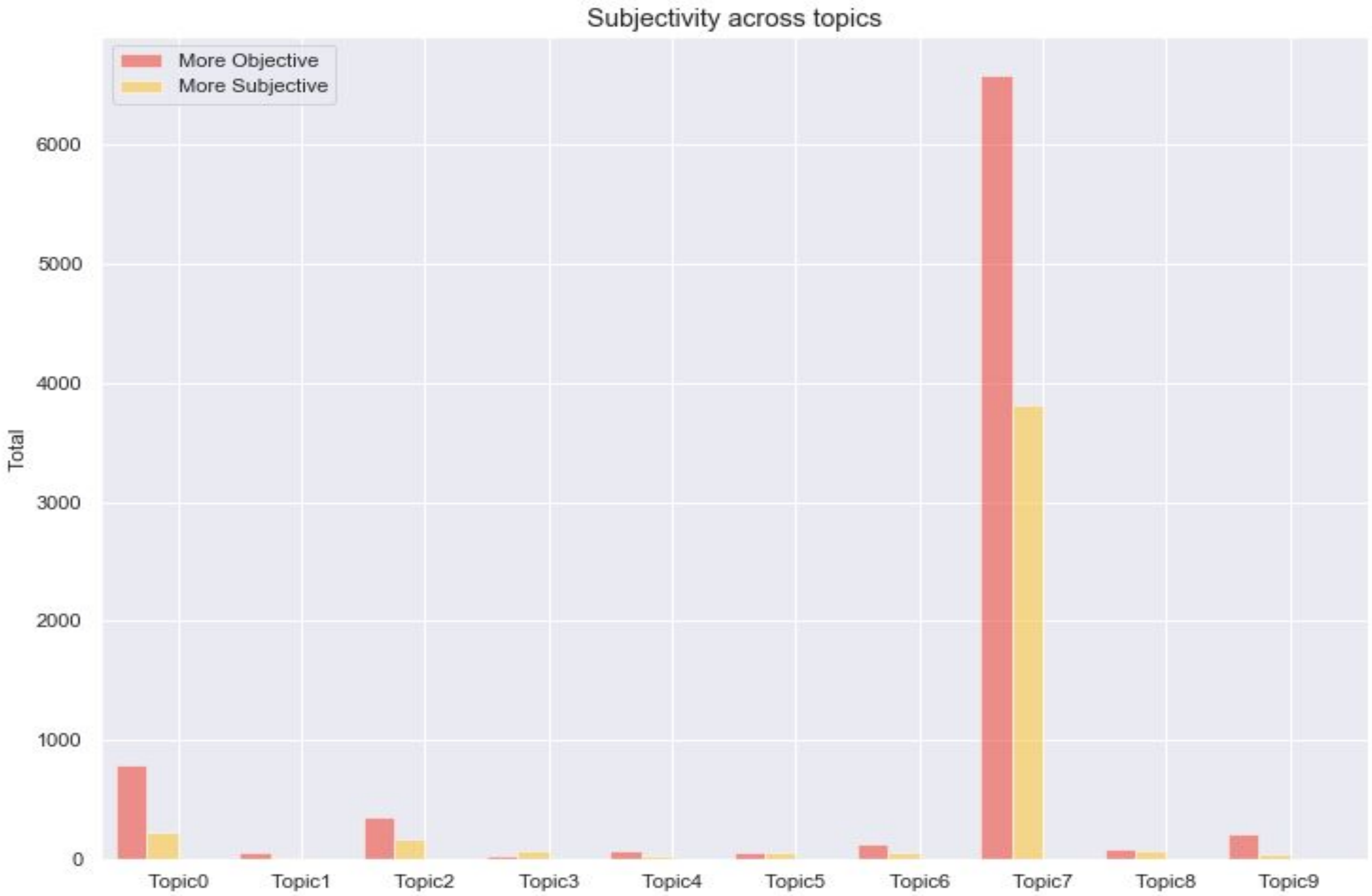
Topic #8



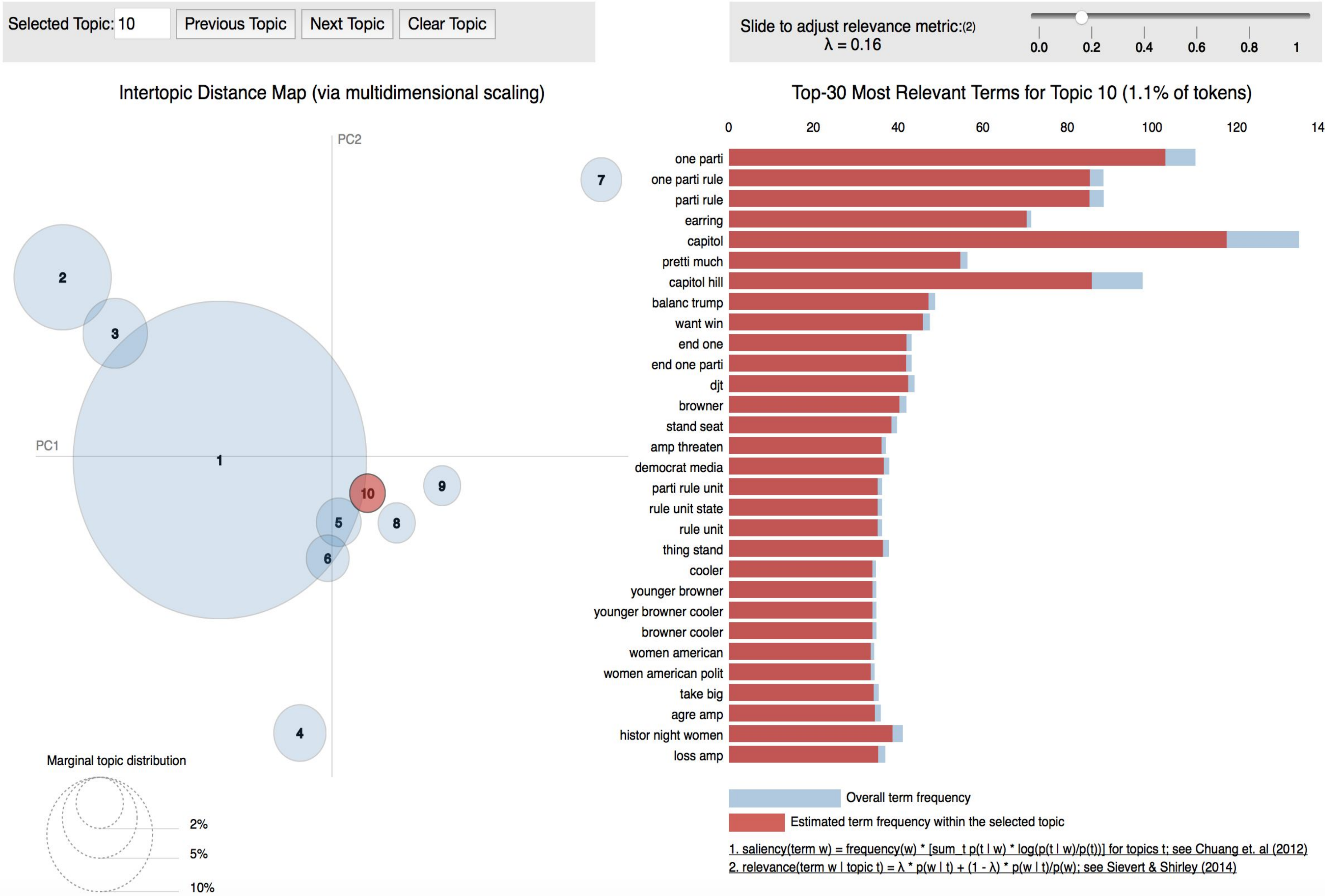
Topic #2



SENTIMENT ANALYSIS



TOPIC ABOUT ONE PARTY RULE ENDING



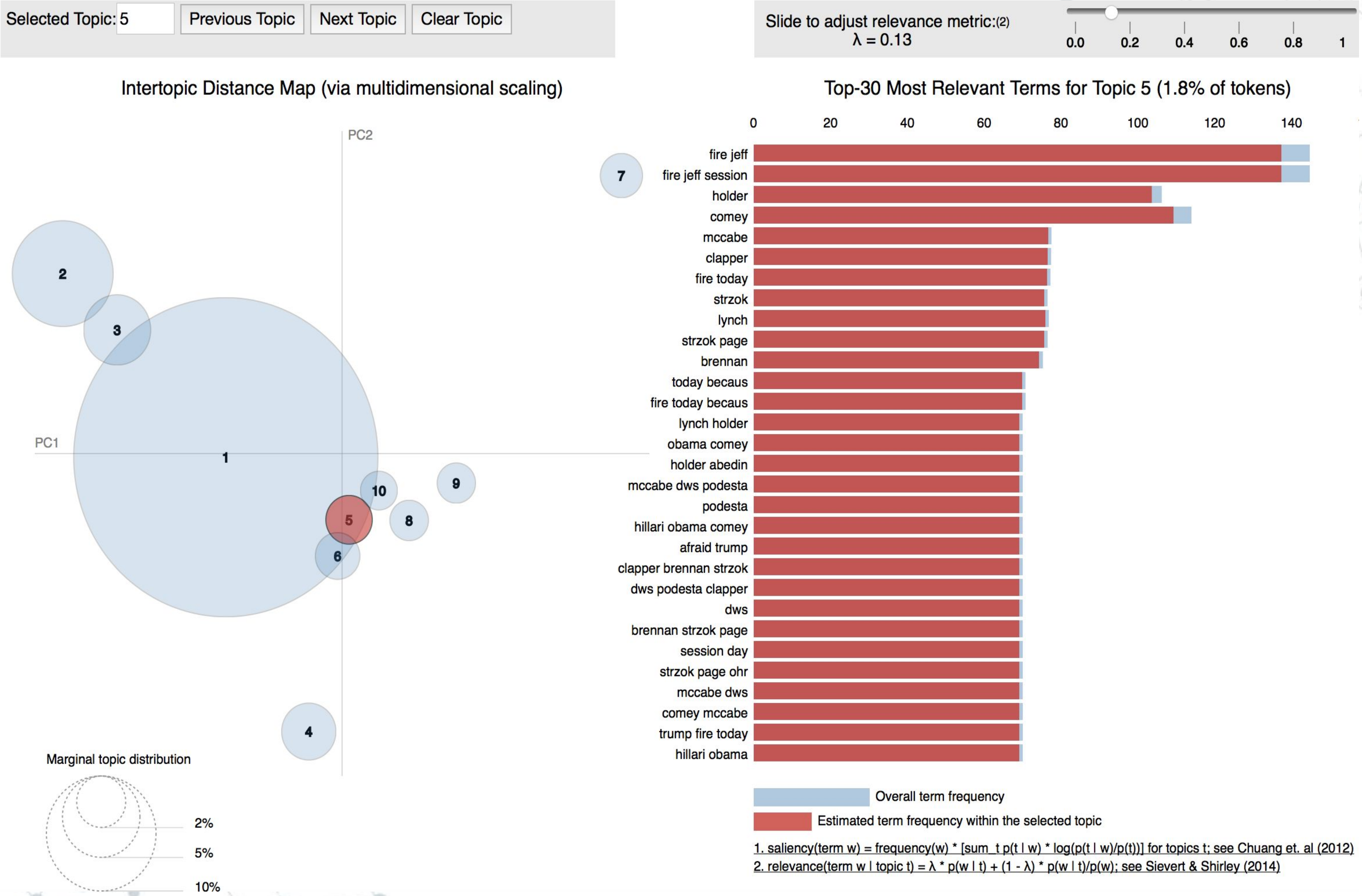
Selected Topic: 10

Slide to adjust relevance metric:(2)
λ = 0.16

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)

TOPIC ABOUT TRUMP FIRING JEFF SESSIONS



Selected Topic: 5

Slide to adjust relevance metric:(2)
λ = 0.13

🟦 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)