Topic Analysis on Obama Briefings

Overview

The goal of the project is to discover the abstract topics that occur in a collection of Obama briefings to check which topic Obama mentioned the most.

Client

Social media like facebook, youtube, Instagram and ect: Topic analysis can be used to analyze which topic people most interested at that time. For example, topic analysis can be used to find which topics people post the most at the certain period of time.

Data

Obama’s briefings were found in the white house webpage. We use the data file at github, which is the download from whitehouse homepage, “ https://www.whitehouse.gov/briefings-statements/ ”. The data file can be found at “ https://github.com/mahmoud/briefings “

Briefings.jsonl : 1778 Obama’s briefings from 2009 to 2016.

Data Cleaning

- Punctuation : punctuation such as “a”, “!” , “?”, ets were removed at each contents of briefings using “string.punctuation” provided from Python. After removing punctuation, 54.3% was reserved.

- Unicode implementation: To have words encoded using Unicode encoding, I encoded text type to utf-8 and remove any numbers and also change every words to lower cases.

Data Preparation

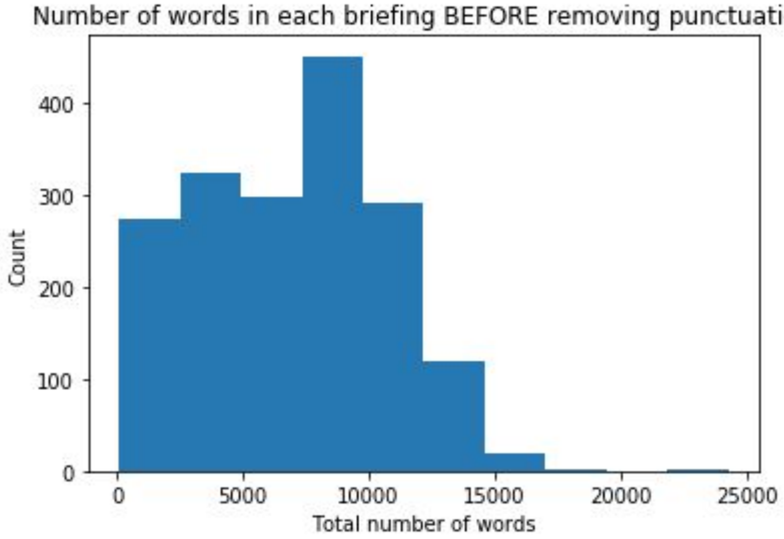
- Tokenization : Tokenized every words in contents

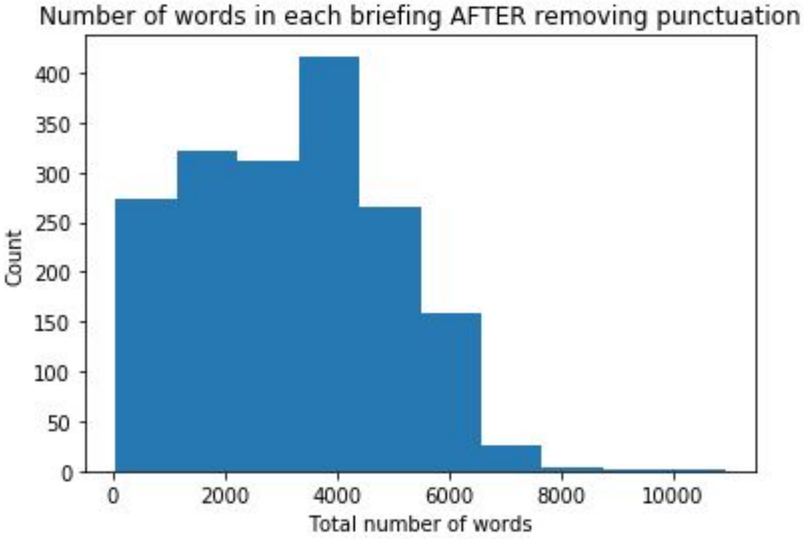
- Removing morphological affixes: “stemmer.stem(token)” was used to remove any unnecessary morphological affixes for better analysis

- Construct word vs ID mapping table to map same tokens to the same IDs for managing efficiently

Data Storytelling

Let’s check the number of words before removing punctuation. The mean is 7016.93 and the minimum and maximum are 101 and 24271. The frequency bar graphs BEFORE and AFTER removing punctuation are like below.

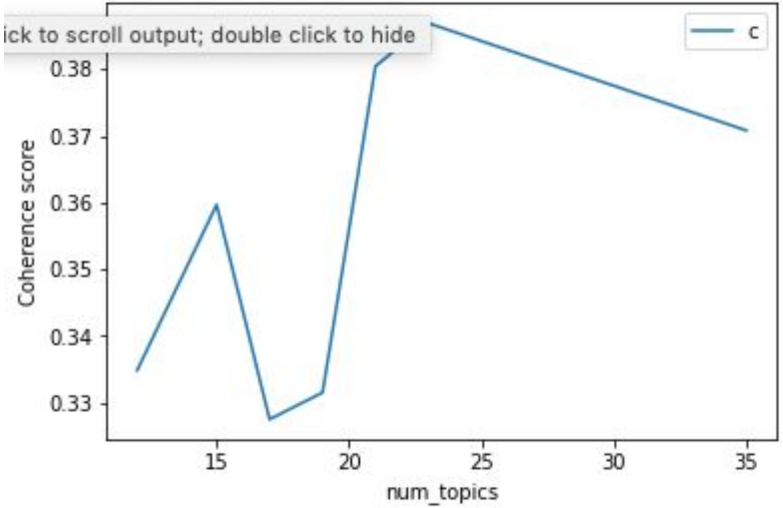




Machine Learning - LDA

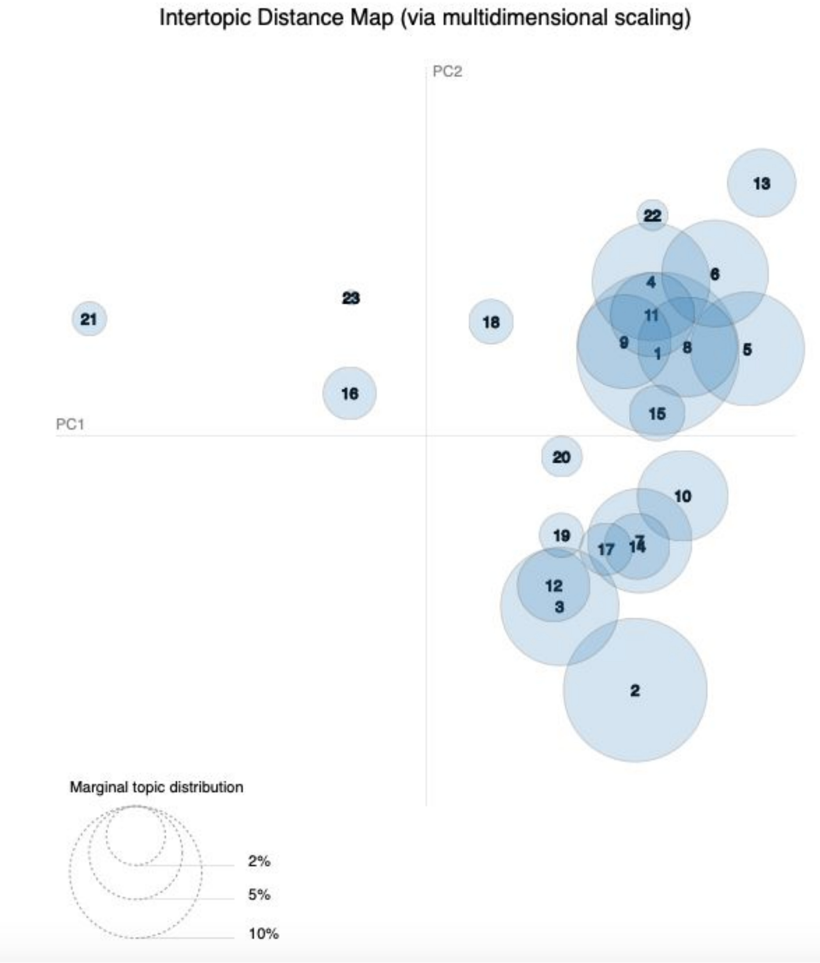
This is one the most popular topic modeling algorithms today. It is a generative model in that it assumes each document is a mixture of topics and in turn, each topic is a mixture of words. So LDA groups words into topics. Then it compares to two topics and determine which topic is close.

The number of topics which are from 12 to 35 were applied to the model to find the best model. From the graph below, the model with 23 numbers of topics are the best model and the coherence score is 0.39.



Machine Learning - LDA

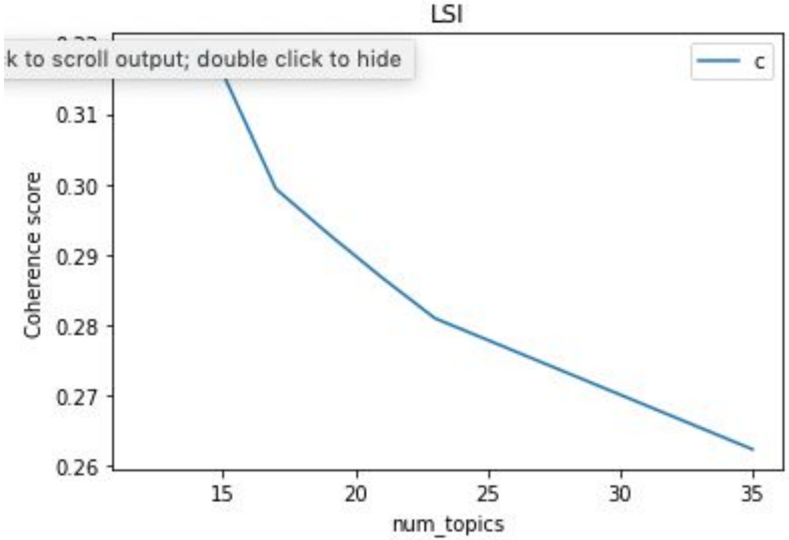
The dominants of each topic can be visualized like below. Topics with the bigger circle mean the more importance relative to the data. From the graph below, we can see that there are two clusters. Clusters on a graph mean the similarity between topics.



Machine Learning - LSI

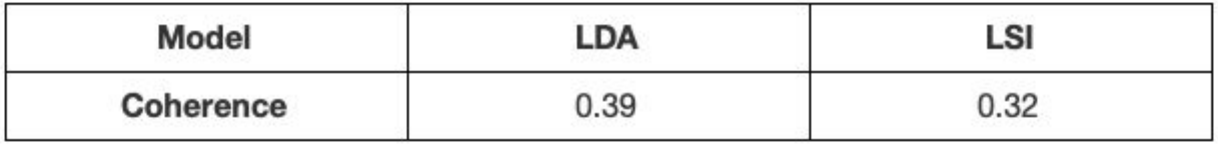
This is a useful topic modeling algorithm in that it can rank topics by itself. Thus it outputs topics in a ranked order. LSI examines the words and looks for how they are related to each other. For example, whether they are similar words, e.g. car, automobile and auto , they are kind of something else, e.g. car and vehicle, or they are part of a larger concept, e.g. engine and car. In practice, LSI is much faster to train than LDA, but has lower accuracy.

The model also needs to decide the number topics to apply to the model. From the graph, the number of topics of 13 has the optimal result. The coherence score is 0.32.



Conclusion

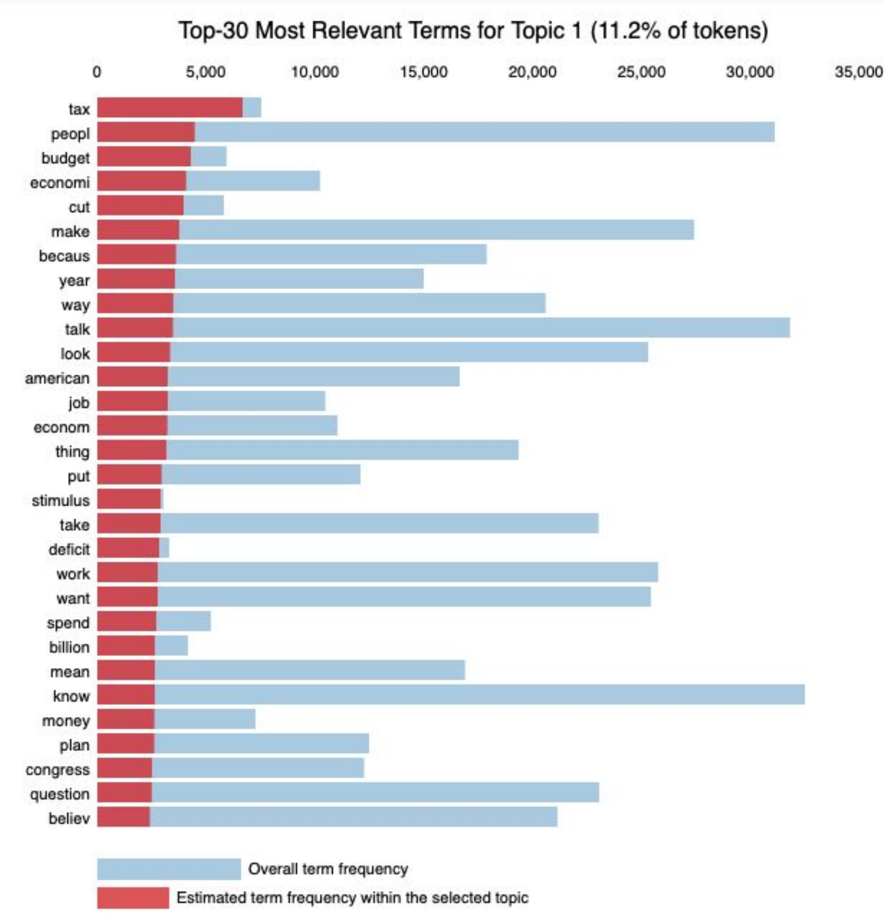
Topic coherence in essence measures the human interpretability of a topic model. It uses wordnet, wikipedia or google to measure how list of words were explained. The coherence for LDA is higher than LSI. From the result of LDA, the topics Obama mentioned often were diplomacy, health care, Congress and economy.



Model	LDA	LSI
Coherence	0.39	0.32

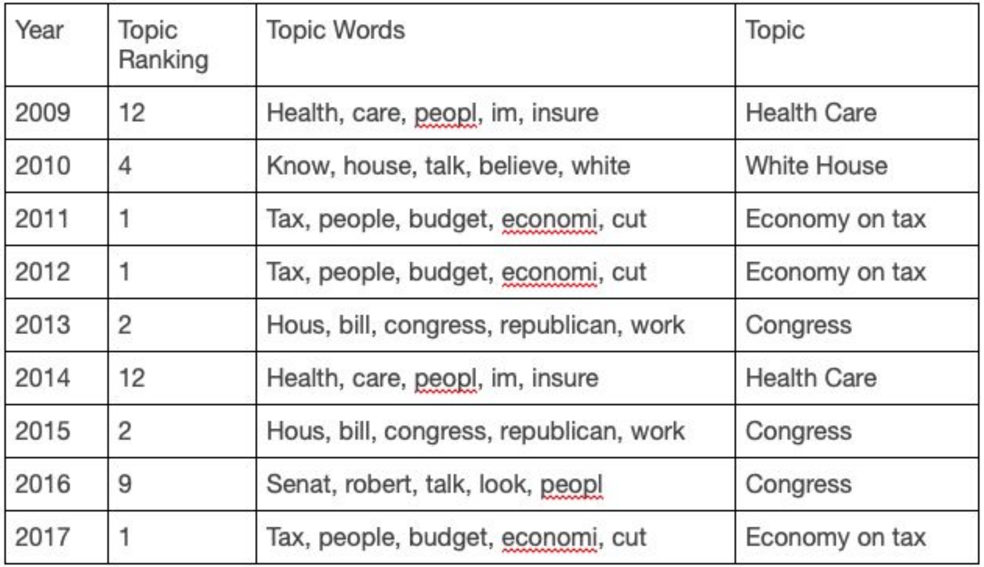
Conclusion

The top topic from LDA model can be visualized like below. The length of red bars for each words means the estimated term frequency within the topic. We can see that topic is related to economy, tax.



Conclusion

The table below shows how dominant topics of briefings change each year. We can see that Obama mentioned health care and white house in his early presidency years. Then, economy became the important topic at 2011 and 2012. After that, congress has been mostly a main issue from 2013 to 2016. During 2017 economy become important.



Year	Topic Ranking	Topic Words	Topic
2009	12	Health, care, peopl, im, insure	Health Care
2010	4	Know, house, talk, believe, white	White House
2011	1	Tax, people, budget, economi, cut	Economy on tax
2012	1	Tax, people, budget, economi, cut	Economy on tax
2013	2	Hous, bill, congress, republican, work	Congress
2014	12	Health, care, peopl, im, insure	Health Care
2015	2	Hous, bill, congress, republican, work	Congress
2016	9	Senat, robert, talk, look, peopl	Congress
2017	1	Tax, people, budget, economi, cut	Economy on tax