# 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

News company : who want to know which area Obama considered important and mentioned often in his briefings such as new classification.

Social media like facebook, youtube, Instagram and ect: Topic analysis can be used to analyze which topic people most interested at that time.

### Data

Obama's briefings were found in the whitehouse 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

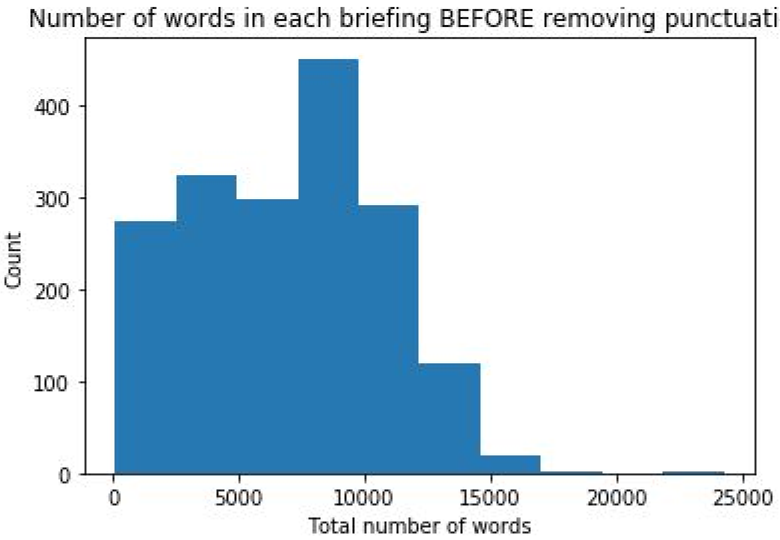
- Punctuation : punctuation such as "a", "!" , "?", ets were removed at each contents of briefings using "string.punctuation" provided from Python. After removing punctuation, 54.3% was reserved.
- Unicode implementation: To have words encoded using Unicode encoding, I encoded text type to utf-8 and remove any numbers and also change every words to lower cases.

## Data Preparation

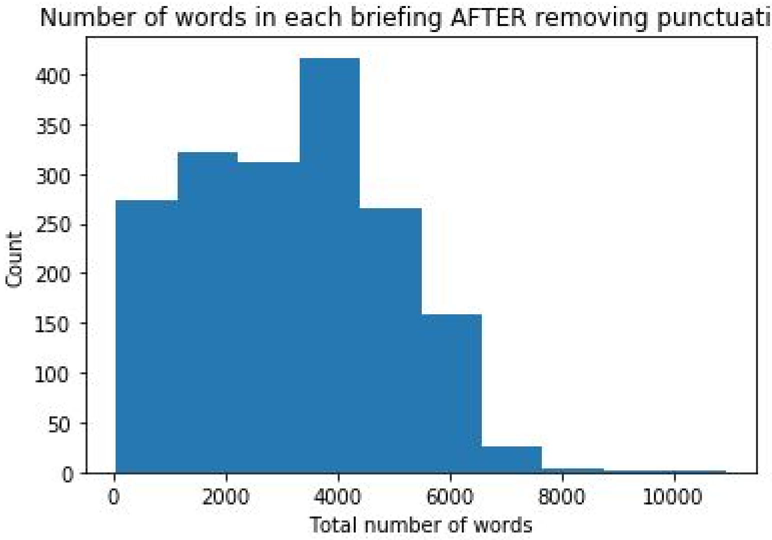
- Tokenization : Tokenized every words in contents
- Removing morphological affixes: "stemmer.stem(token)" was used to remove any unnecessary morphological affixes for better analysis
- Construct word vs ID mapping table to map same tokens to the same IDs for managing efficiently

## 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 plot is like below.



Let’s check the number of words after removing punctuation. The mean is 3207.11 and the minimum and maximum are 62 and 10929.



From two bar plots above, we can see that graphs look alike. After removing punctuation, about 50 % was removed.

# Machine Learning

First, mapping table on "descriptions" vs ID was created to efficiently apply values to models. Then,dictionary was converted into a bag-of-words. The result, corpus , is a list of vectors equal to the number of documents. As a result, we have frequency table of each words

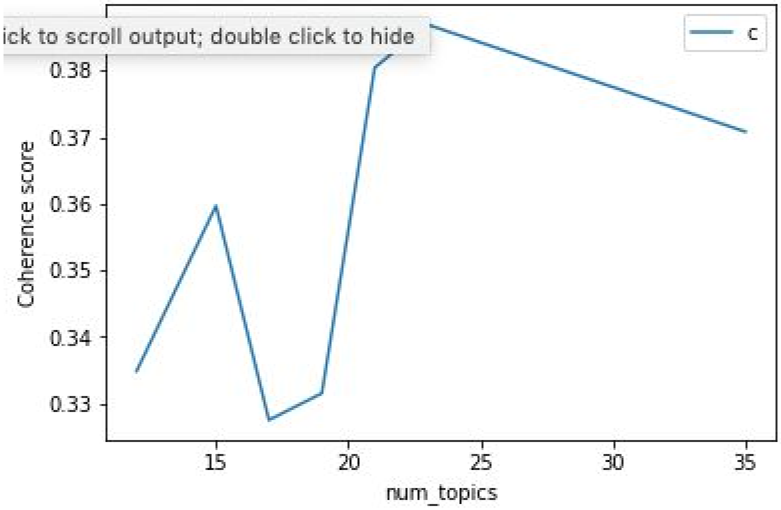
To run any mathematical model on text corpus, it is converted into a matrix representation.

## LDA (Latent Dirichlet Allocation)

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.

LDA require a num_topics parameter (set to 200 by default) to limit number of topics in the model. So several numbers of topics were tested on the model to pick the best one. Coherence score was used to measure the model. Coherence socre was used to measure the model. Topic coherence measures the human interpretability of a topic model. Topic coherence evaluates topic models how models are interpretable using wikipedia or wordnet. Thus this can be used to compare different topic models.

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.

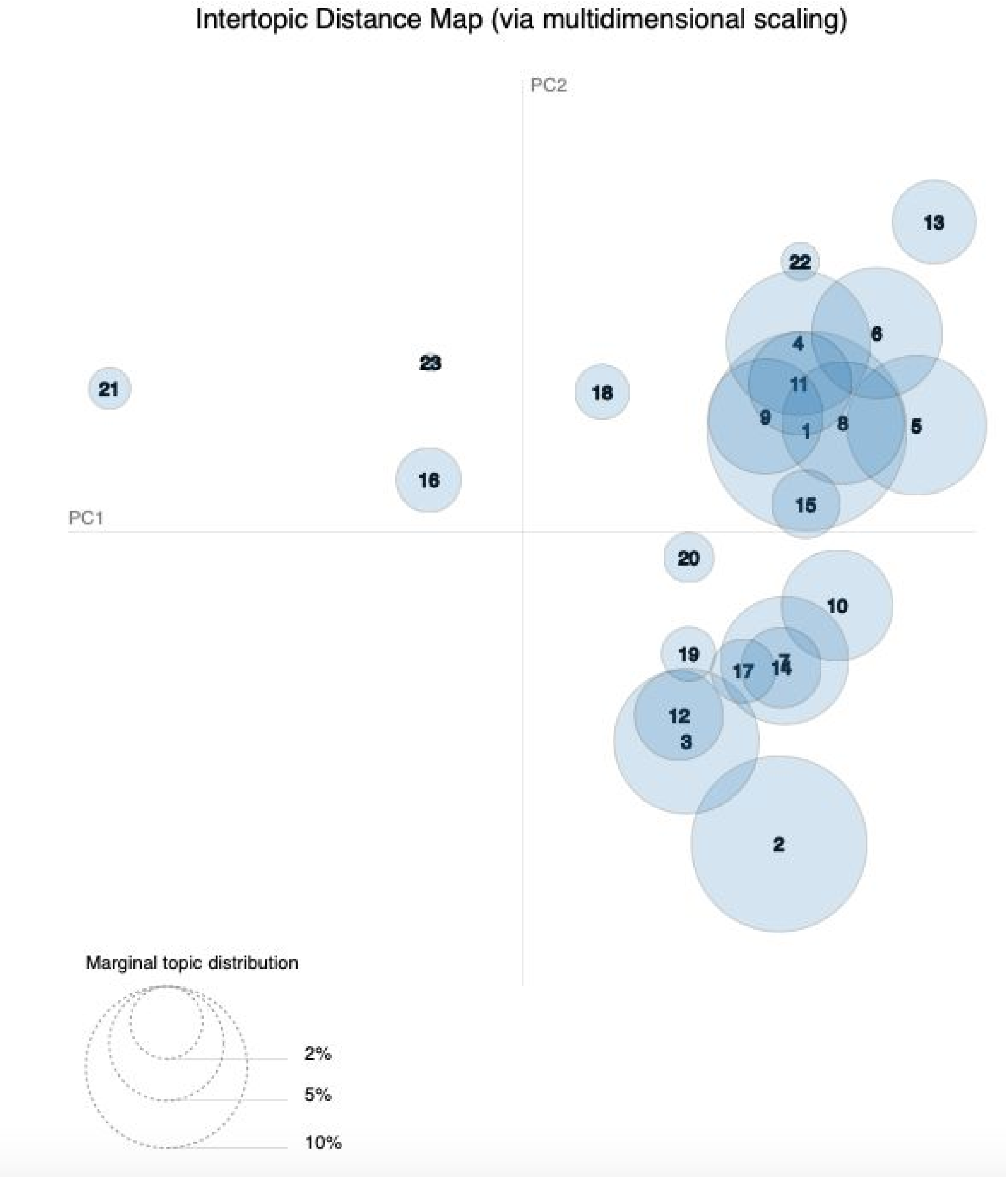


The result of LDA with 23 topics is like below. Possible subject for each topic was based on searches at google with words in each topic.

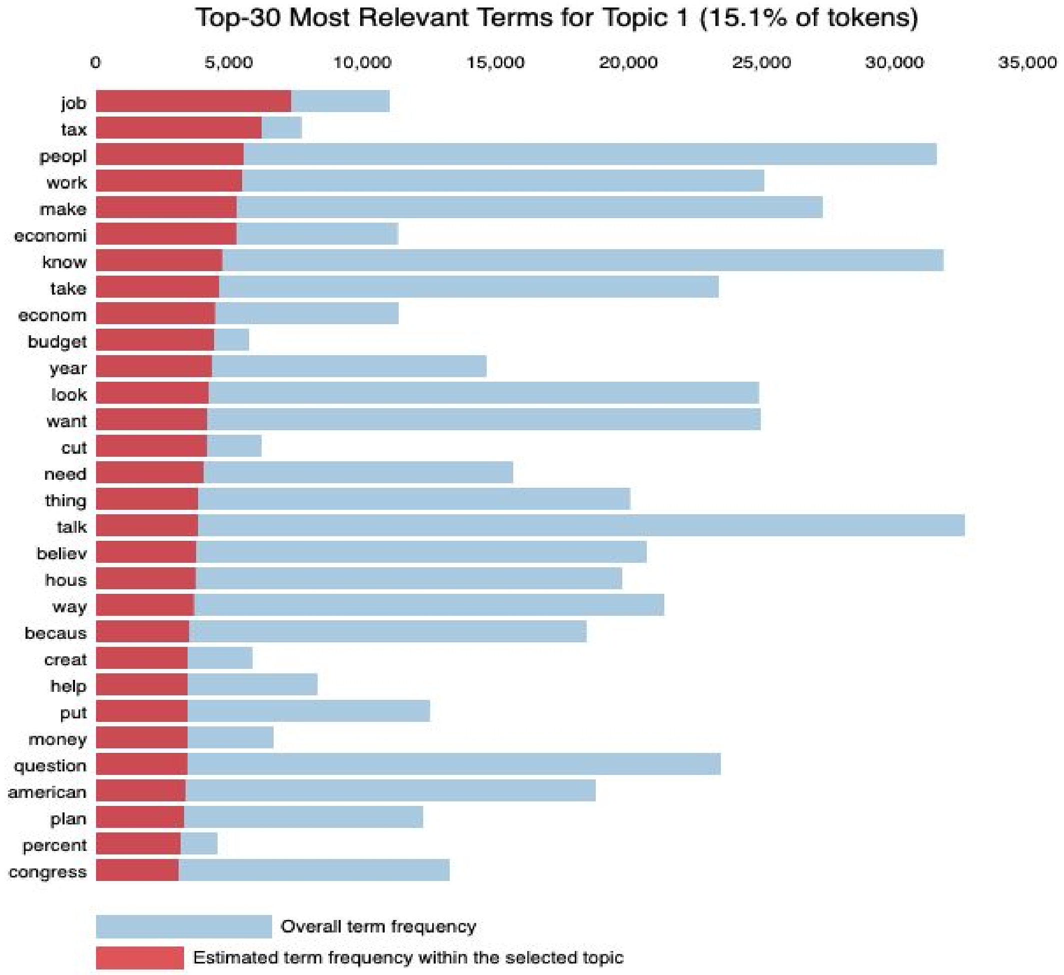
The dominants of topics are like below. From the table below, we can see that top 1 and 2 topics are economy. Top 3 topic is about diplomacy on Iraq.

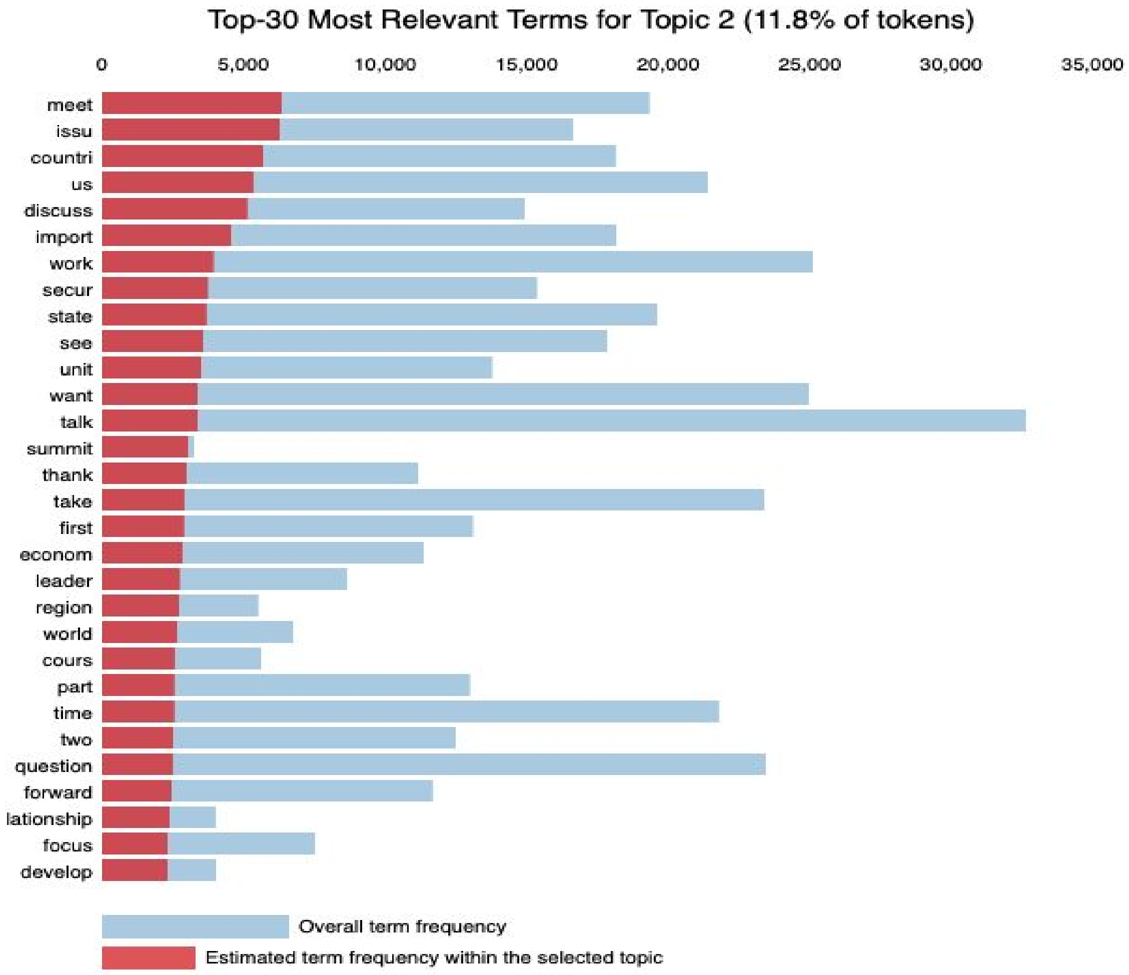
| | | **Words in each topic** |
|---|---|---|
| 1 | 15.1% | Job, tax, peopl, work, make, economi,know,take econom,budget, |
| 2 | 11.8% | Meet,issu,countri,us,discuss,import, work,secur,state,see |
| 3 | 8% | Iraq, militari, unit, forc, secur, libya, iraqi, us, govern, take, peopl |
| 4 | 7.9% | Know, talk, look, peopl, campaig, time, hous, laughter, week, make |
| 5 | 7.4% | Republican, american, congress, way, talk, hous, believe, negoti,thing, peopl |
| 6 | 6.5% | Health, care, peopl, im, talk, know, look, hous, want, make mani |
| 7 | 6.2% | Afghanistan, secur, troop, pakistan, us, afghan, administr, govern, forc |
| 8 | 5.7% | Know, robert, obvious,believ, im, talk, hous, yes, question, senat, look, work, |
| 9 | 5% | Administr, law, mexico, immigr, border, peopl, talk, state, make, use, somth |
| 10 | 4.7% | Iran, nuclear, iranian, intern, sanction, weapon, make, know, take, talk, progr |
| 11 | 4% | Inform, secretari, administr, know, investig, question, made, depart, general |
| 12 | 3% | Oil, energi, secretari, look, us, people, take, like know, want, question, see |
| 13 | 2.6% | Senat, court, forclosur, process,republican, look, nomine,judge,nomin, pee |
| 14 | 2.4% | Burton, israel, prime, know, minist, talk, isra, meet, palestinian, peac, speech |
| 15 | 1.8% | Make, jone, people, look, time, talk, know, take, point, obvious, decis, state, |
| 16 | 1.6% | Dr, medic, respons, test, leste,flu, health, coordin, viewpoint, peopl, helen, k |
| 17 | 1.5% | Russin, russian, medvedev, ukrain, moscow, savannah, georgia, vice, missil, |
| 18 | 1.1% | Financi, bank, quto, administr, car, recoveri, compani, industri, money, taxp |
| 19 | 1.1% | Reinvest, peopl, earmark, qaddafi, becaus, know, govern, us, work, haiti, st |

The dominants of each topic can be visualized like below.



The top 2 topics are like below.

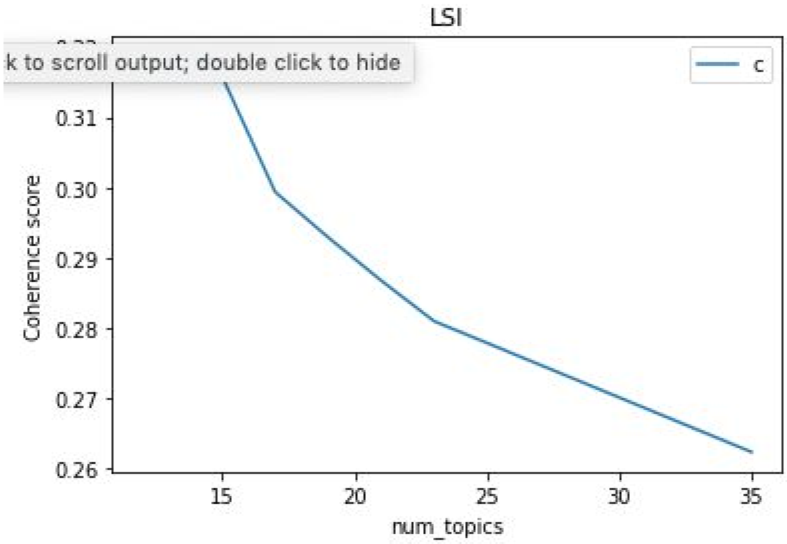




## LSI (Latent Semantic Indexing)

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.



Let’s find possible subject like LDA. The problem here is that words are repeated across each topic. For example, Iran is mentioned at topic 5,6,7,8 and 11 while iran is mentioned only once at LDA. Also if you look at topic 8, words related to health care like health, care are at the same group with words related to diplomacy like iran, cuban. When we search words in each topic on Google, possible subject didn’t come up as list of words are not related to each other.  Therefore, LSI is not suitable as much as LDA.

| | **Words in each topic** | **Possible subject** |
|---|---|---|
| 1 | state, make, peopl, know, unit, talk, work, take, us, becaus | |
| 2 | unit, state, secur, tax, countri, job, look, militari, cut, isil, | |
| 3 | republican, hous, congress, senat, us, meet, issu, discuss, american, democrat | |
| 4 | tax, health, congress, cut, peopl, know, care, look, economi, deficit, | |
| 5 | iran, agreement, militari, forc, nuclear, republican, american, senat, meet, iraq, | |
| 6 | iran, peopl, hous, nuclear, countri, agreement, white, tax, year, american | |
| 7 | care, insur, afford, health, senat, iran, peopl, forc, becaus, isil, | |
| 8 | iran, agreement, issu, nuclear, health, care, action, isil, cuban, budget, | |
| 9 | senat, care, administr, republican, insur, health, meet, job, tax, afford | |

| | | |
|---|---|---|
| 10 | syria, russia, hous, cuban, russian, secur, assad, weapon, govern, cuba | |
| 11 | american, iran, russia, budget, secretari, unit, peopl, tax, immigr, nuclear | |
| 12 | tax, cuban, congress, govern, secur, peopl, cuba, budget, white, made | |

## 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 |