Final Report-LDA Model of Trump Tweet

Jingshu Song, Qianhui Cen, Daniel Vitale, Yiming Dong

Introduction

As we all know, Trump's tweets are a popular topic in our daily life. This project seeks to examine topics of interest in U.S. President, Donald Trump's tweets. We choose dataset about Trump's tweets from May 2009 to July 2019 from Harvard Dataverse Website. By acquiring a dataset of all his tweets within this time window, we build monthly topic models using LDA model and TF-IDF model to figure out the change in areas of interest of the President of the United States over time.

Based on the analysis on the models, we are able to derive a set of topics from the corpus of Trump Tweets, which shed light on his life and historical public exposure while making sense in the context of changes known to have occurred in Trump's life over time, leading up to, and during his Presidency.

Data-Preprocessing

Firstly, we create a data frame after we import 10-year-long dataset as Trump`s Tweets. There are a lot of variables inside the whole input data. Like 'created_at' , 'favorite_count', 'Retweet_count', 'in_reply_to_screen_name', 'retweeted','text', 'truncated'. Our goal is to analyze the topic therefore our model is LDA.

1. Removed stop words, punctuation, and hyperlinks as well as Donald Trump's name
2. Lemmatized text
3. Then used corpora to create a dictionary and term matrix per month
4. We then produced topic models for each month

In the following figures, we can get a look at the changes among the 10 years by monthly basis. It can be seen that in the recent 2019 how the number of tweets fluctuated.

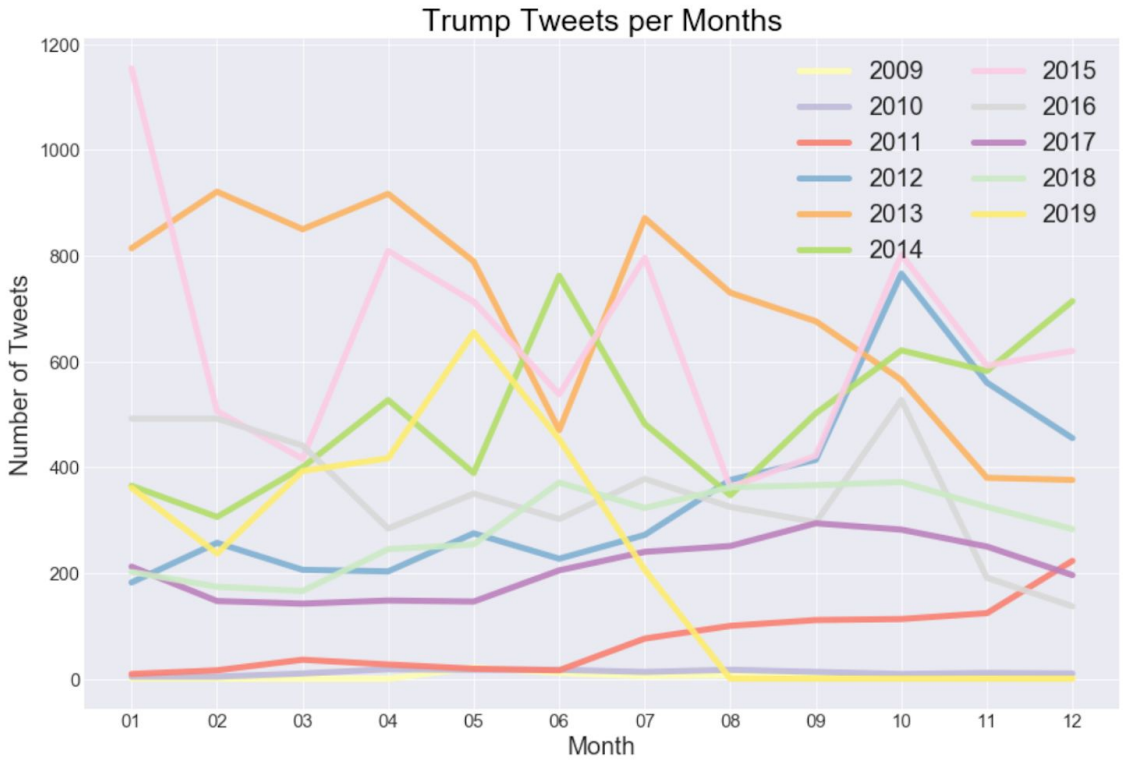

Figure 1

In the following figures, we can get a look at the changes among the 10 years. It can be seen that 2013 Trump released the most tweets compared with the amount of tweets in other 9 years. And Trump tweets most in January 2015.

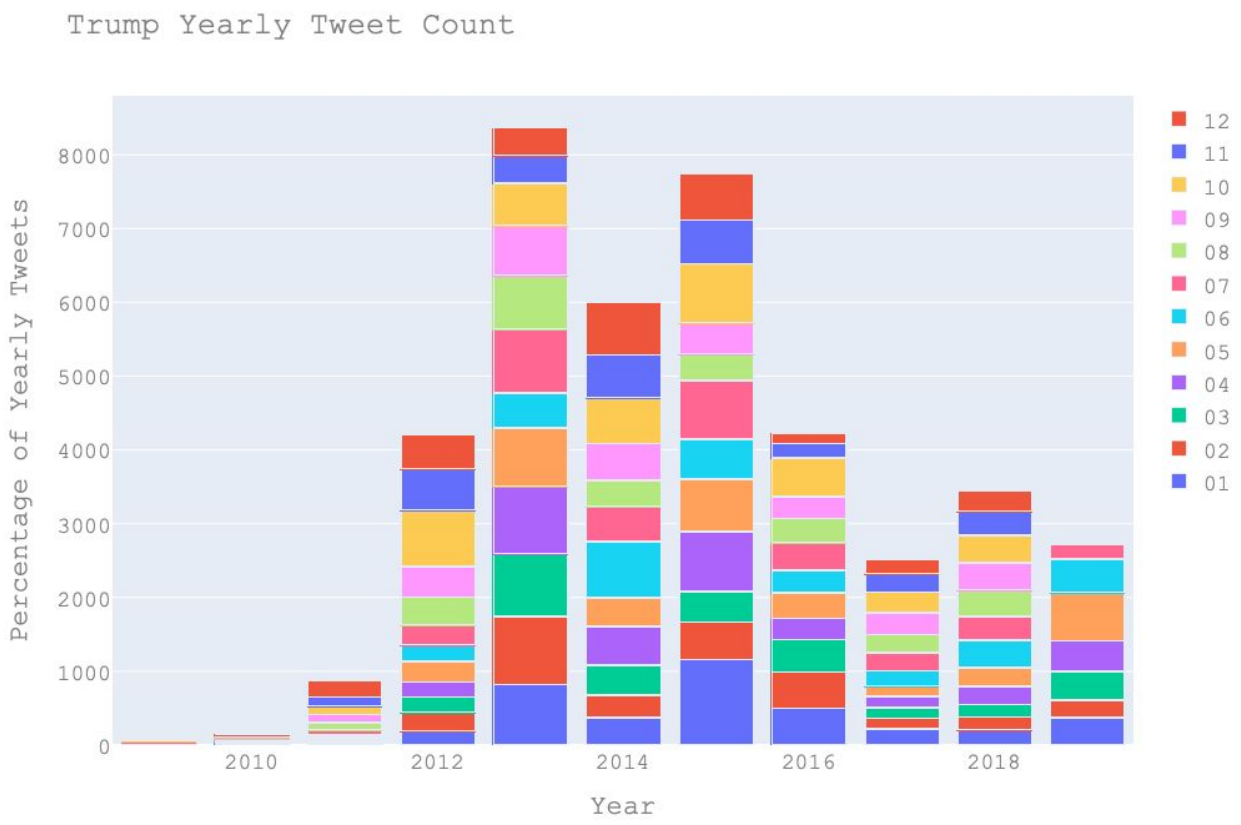

Figure 2

LDA model

After we preprocessed our data, we worked on the LDA model. The basic knowledge of LDA will be introduced first. The Latent Dirichlet Allocation (LDA) is a kind of way to identify the topics.In natural language processing, LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is a form of unsupervised learning that views documents as bags of words.

For this project, we want to know what topic that Trump's tweets talked about each month from 2009 to 2019, so we created dictionaries of his tweets for each month and we removed some words like "Donald" and "trump" because these words have a high frequency in the tweets and

they can't help us about topic identification. Besides that, we also removed stop words and punctuation, and then lemmatized text. After that,

We used corpora to create a dictionary and term matrix per month, and produced topic models for each month. In the following picture, you can see the result of the LDA model for each month.

```
Out[7]: {'2009-05': [(0, '0.049*"read" + 0.049*"top" + 0.049*"ten"'),
          (1, '0.053*"keep" + 0.053*"like" + 0.053*"think"'),
          (2, '0.054*"read" + 0.054*"great" + 0.054*"new"'),
          (3, '0.067*"discussing" + 0.067*"new" + 0.067*"listen"'),
          (4, '0.047*"always" + 0.047*"fired" + 0.047*"second"'),
          (5, '0.039*"apprentice" + 0.039*"celebrity" + 0.039*"champion"'),
          (6, '0.056*"like" + 0.056*"afraid" + 0.029*"champion"'),
          (7, '0.049*"my" + 0.049*"never" + 0.049*"them"'),
          (8, '0.032*"think" + 0.032*"like" + 0.032*"champion"'),
          (9, '0.039*"direct" + 0.039*"could" + 0.039*"self"')],
         '2009-06': [(0, '0.013*"big" + 0.013*"rating" + 0.013*"“expand"'),
          (1, '0.013*"big" + 0.013*"rating" + 0.013*"“expand"'),
          (2, '0.059*"rating" + 0.059*"check" + 0.059*"big"'),
          (3, '0.043*"re" + 0.043*"beckham" + 0.043*"chronicle"'),
          (4, '0.013*"big" + 0.013*"rating" + 0.013*"“expand"'),
          (5, '0.044*"today" + 0.044*"record" + 0.044*"frank"'),
          (6, '0.086*"birthday" + 0.086*"wish" + 0.086*"thanks"'),
          (7, '0.044*"loss" + 0.044*"it’s" + 0.044*"michael"'),
          (8, '0.062*"champion" + 0.062*"last" + 0.062*"like"'),
```

Figure 3

We selected the number of the topic as 10. From the result, we can notice that in May 2009, Trump's tweets mentioned read, top and ten words most, while in June 2009, trump's tweets mentioned big, rating and expand. As you can see, words like" read", "discuss", "think" were came up with a lot in May and in June words like "expand", "rate", "big" and "Michael" were mentioned many times, so we can conclude that Trump prefer to expand US and improved the rate on some aspects.

After that, we selected the recent month (2019-07) and used PyLDAvis to present the result of LDA model.

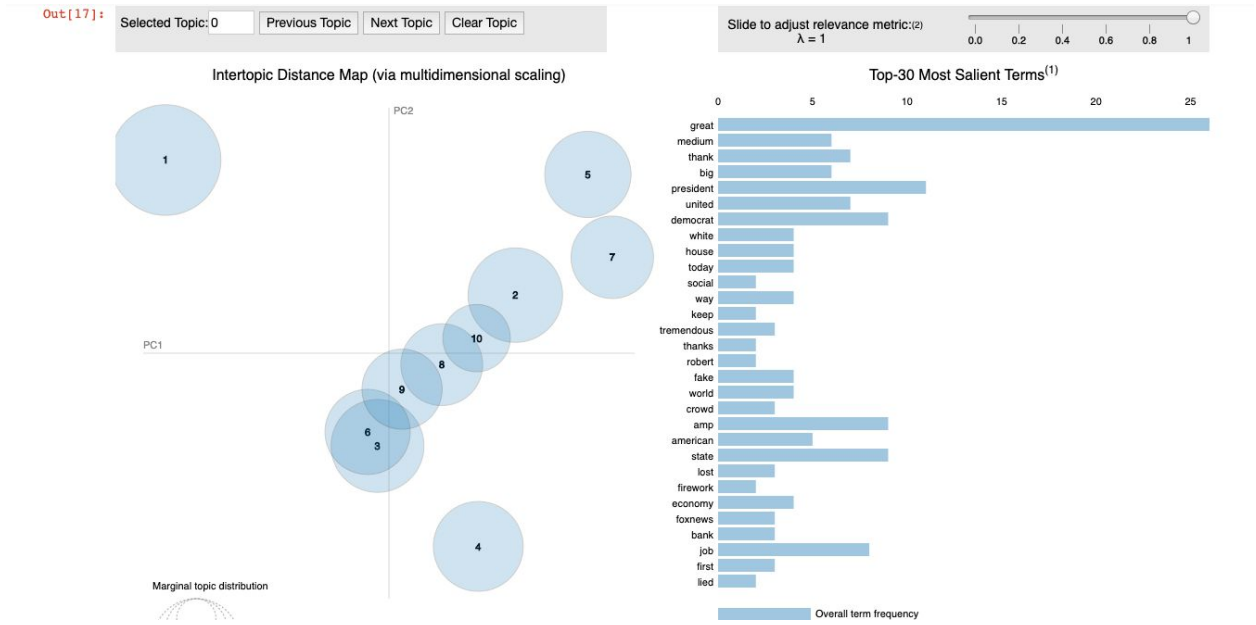

Figure 4

As you can see, on the left, the topics are plotted as circles, whose centers are defined by the computed distance between topics. In this project, we selected the number of topic as 10. The prevalence of each topic is indicated by the circle's area. On the right, two juxtaposed bars showing the topic-specific frequency of each term (in red) and the corpus-wide frequency (in blueish gray). When no topic is selected, the right panel displays the top 30 most salient terms for the dataset. In July, Trump's tweets mentioned these words with high frequency, for example, great, medium, thank, big and so on. When we select the first topic, you can see the result in figure 5.

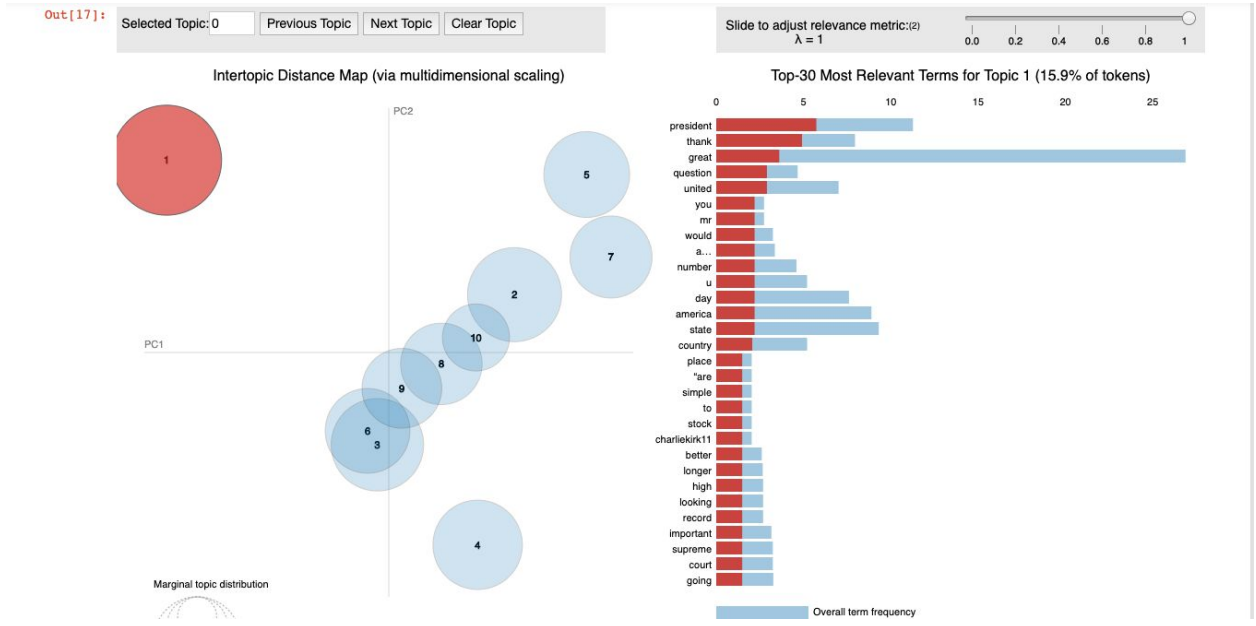

Figure 5

The red words are the top 30 most relevant terms were mentioned in the topic 1, for example, president, thank and great were the top 3 terms in the topic 1. When we set the relevant as 1, it shows the frequency that these words are camp up with in all topics. It can be noticed that the word “great” is mentioned most in all topics and the second one is “president”.

TF-IDF and LDA

Lastly, we used our monthly corpus dictionary to first create a TF-IDF model, and then use the TF-IDF to run another LDA model per month.

As discussed in the original 2003 Blei paper, "Latent Dirichlet Allocation":

"LDA addresses the shortcomings of the TF-IDF model and leaves this approach behind. LSA is completely algebraic and generally (but not necessarily) uses a TF-IDF matrix, while LDA is a probabilistic model that tries to estimate probability distributions for topics in documents and words in topics. The weighting of TF-IDF is not necessary for this."

The normalization of terms by TF-IDF reduces the weight of terms which occur more frequently across the whole corpus- meaning the terms with the highest weights are those which occur least across all of the tweets for that month. Ultimately, the introduction of TF-IDF to the LDA model can produce a more sensitive topic model, tagging terms which may not be discovered without the normalization. Whether this is actually useful can be debatable, though it certainly returns a different set of topics than the regular LDA model. For example, the top 3 terms for the top 10 topics returned for July 2017 in the regular LDA model are as follows:

```
[(0, '0.014*"republican" + 0.011*"great" + 0.008*"american"'),
 (1, '0.020*"great" + 0.014*"thank" + 0.010*"senate"'),
 (2, '0.019*"great" + 0.011*"u" + 0.009*"year"'),
 (3, '0.015*"job" + 0.012*"amp" + 0.009*"vote"'),
 (4, '0.019*"working" + 0.016*"hard" + 0.010*"people"'),
 (5, '0.034*"amp" + 0.012*"investigation" + 0.009*"republican"'),
 (6, '0.023*"amp" + 0.010*"president" + 0.010*"honor"'),
 (7, '0.018*"amp" + 0.016*"obamacare" + 0.011*"repeal"'),
 (8, '0.011*"amp" + 0.009*"thank" + 0.009*"meeting"'),
 (9, '0.016*"amp" + 0.011*"republican" + 0.010*"u"')]
```

Figure 6

But when we go ahead and introduce the TF-IDF to the LDA model for the same time interval, we see somewhat similar results, but yet different terms which were not returned in the regular LDA model are returned:

```
[(0, '0.004*"great" + 0.004*"amp" + 0.003*"state"'),
 (1, '0.005*"thank" + 0.005*"bad" + 0.004*"first"'),
 (2, '0.011*"job" + 0.004*"evening" + 0.004*"maga🇺🇸"'),
 (3, '0.007*"u" + 0.005*"get" + 0.005*"amp"'),
 (4, '0.006*"speaking" + 0.006*"france" + 0.005*"morning"'),
 (5, '0.005*"you" + 0.004*"join" + 0.004*"member"'),
 (6, '0.008*"again" + 0.007*"make" + 0.007*"great"'),
 (7, '0.005*"year" + 0.005*"stock" + 0.004*"almost"'),
 (8, '0.004*"totally" + 0.004*"senatordole" + 0.004*"kind"'),
 (9, '0.006*"amp" + 0.005*"go" + 0.004*"vote"')]
```

Figure 7

As you can see, many of the same terms in the same topic clusters are returned, yet there are some terms such as “maga” and “France” which were not picked up in the first one. Utilizing both types of models over the same time period makes it possible to elucidate a potentially richer set of terms related to said period.

Conclusion

Ultimately, this project is a collection of topics which are mined from Donald Trump’s Twitter account which can serve as a quick reference for examining what the President of the United States was interested in across 10 years at any time period. We had hoped to use this LDA project to pursue direct lines of inquiry related to the change in specific topics over time. We had intended to use this as a baseline and then implement Dynamic Topic Modeling, as laid out in the Blei paper but we ran into some technical difficulties and decided to try to make the individual monthly topics models as thorough as possible instead. In short, we just ran out of time. Although this was not entirely what we had intended, it does serve as a quick reference to examine specific topics of interest at specific months as well as a good baseline from which a dynamic model could be compared.