Priya Sanodia

Binoy Dutt

Social Media Analytics

Report

Contents

[1 Data Collection: 2](#_Toc478751976)

[2 Sentiment Analysis: 3](#_Toc478751977)

[3 Word-Cloud: 4](#_Toc478751978)

[4 Topic Modelling: 5](#_Toc478751979)

[5 State Wise Tweet Collection and Analysis 6](#_Toc478751980)

[5.1 Geographic Variation Table: 6](#_Toc478751981)

[5.2 New York 7](#_Toc478751982)

[5.2.1 Overview 7](#_Toc478751983)

[5.2.2 Sentiment Analysis 7](#_Toc478751984)

[5.2.3 Word Cloud 8](#_Toc478751985)

[5.2.4 Topic Modelling 8](#_Toc478751986)

[5.3 Georgia 10](#_Toc478751987)

[5.3.1 Overview 10](#_Toc478751988)

[5.3.2 Sentiment Analysis 10](#_Toc478751989)

[5.3.3 Word Cloud 11](#_Toc478751990)

[5.3.4 Topic Modelling 11](#_Toc478751991)

[5.4 California 13](#_Toc478751992)

[5.4.1 Overview 13](#_Toc478751993)

[5.4.2 Sentiment Analysis 13](#_Toc478751994)

[5.4.3 Word Cloud 14](#_Toc478751995)

[5.4.4 Topic Modelling 14](#_Toc478751996)

[5.5 Texas 15](#_Toc478751997)

[5.5.1 Overview 15](#_Toc478751998)

[5.5.2 Sentiment Analysis 15](#_Toc478751999)

[5.5.3 Word Cloud 16](#_Toc478752000)

[5.5.4 Topic Modelling 16](#_Toc478752001)

[5.6 Illinois 17](#_Toc478752002)

[5.6.1 Overview: 17](#_Toc478752003)

[5.6.2 Sentiment Analysis 17](#_Toc478752004)

[5.6.3 Word Cloud 18](#_Toc478752005)

[5.6.4 Topic Modelling 18](#_Toc478752006)

[6 Insights 19](#_Toc478752007)

# Data Collection:

Twitter developer application was created and the ‘Consumer Key’, ‘Consumer Secret’, ‘Access Token’, ’Access Token Secret’ credentials were generated and stored in the form of a dictionary in a json file. The tweets were collected using **Twython** using keyword tracking.

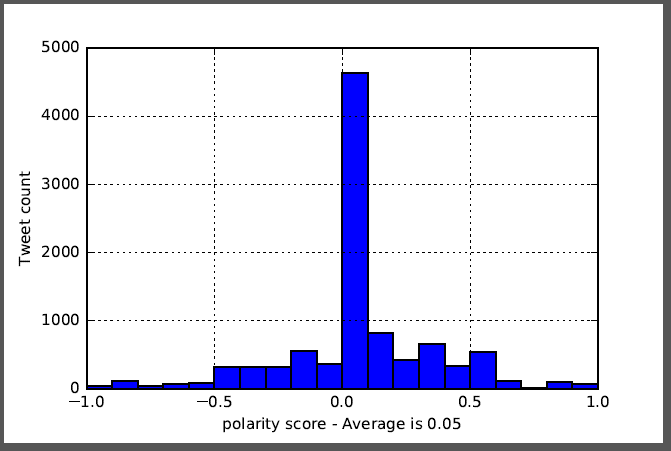
Few issues that were faced and were rectified as follows:

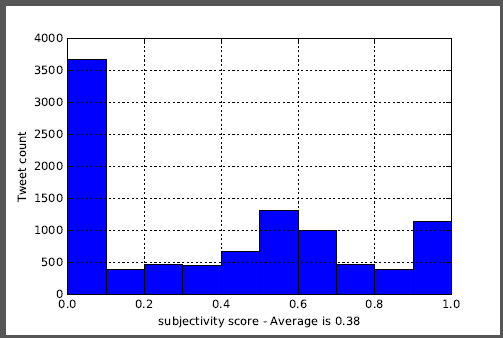
* To avoid the Unicode error we have changed the print condition in success method from [print 'received tweet #', len(tweets), data['text'][:100]] to [print 'received tweet #', len(tweets)] as the data received was not always in UTF-8 format.
* To avoid the Chunked Encoding Error due to connection failure we have added the try and except condition under a while loop.
* To handle other errors such as internet connectivity and key board interruption or we added a generic except statement which would make sure that the collected tweets are stored in a json file and the tweets are not lost.

# Sentiment Analysis:

* The collected 10000 tweets are successfully now stored in json file. Using **TextBlob** the subjectivity and polarity scores were calculated for each tweet of these tweets and stored in the form of a list.
* Histogram with 20 bins for polarity and 10 bins for subjectivity with average scores as labels was generated using **matplotlib.pyplot** module. Post sentiment analysis all the tweet text was stored in a json file for further use.

Below is the graph for same:

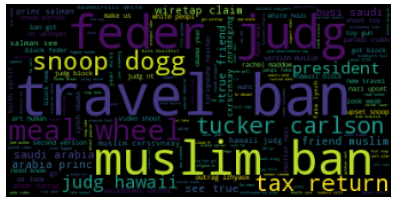




# Word-Cloud:

* As a first step the Punctuations and Digits were removed using String Class as they would hinder the stopwords removal and stemming. As all the tweet characters were not in UTF-8 format there was a mismatch error which was rectified by encoding.
* Each word from the tweet was then stemmed to get the root stem words to avoid words having the same roots but different forms being bifurcated.
* A set of ignore words and accept words were created based on the following 2 criteria:
  1. Few words that were not required to be stemmed as they were noun or irrelevant or articles
  2. Words that were showing up in the word cloud with high frequency without any relevant meaning.
* Spelling check in each word of tweet was attempted but encountered with few errors like Trump got spell-checked to Plump, tweets got converted to sweets and some other grammatical errors hence scrapped the idea of a spell check.
* Post stemming all the stop words were again removed as to eliminate few words that were getting generated due to stemming and have no relevant motive
* After all the pre-processing the filtered tweets were stored in a json file in the form of a list for future use.

Following Word-Cloud was generated using the 10,000 generic trump tweets



# Topic Modelling:

* As part of topic modelling, topics were generated using the number of topic as 4,5,6,7,10,15,20,25,30.
* With the number of topics as 5 and the word count as 6, the result seemed much more readable relatively whereas with the increase in number of topics the list of topics become more varied with many topics not making much sense.
* In general, NFM proved to give much more insightful topics compared to LDA topics. Few topics generated by LDA were difficult to understand.
* All the generated topics were stripped and stored in a json file in the form of a list for future use.

The list of topics can be found below:

**Topic List LDA General**

* true see busi friend princ arabia
* nashvil ralli meal chant lock news
* take claim full read wiretap stop
* ban feder block judg travel hawaii
* outrag lzhyaox akhdr obama fake people

**Topic List NFM General**

* ban judg block feder travel muslim,
* carlson tucker wiretap claim fr min
* video snoop shoot toy gun tariqnashe
* znrddxgzng cxrscvnxay salman arabia princ friend
* outrag lzhyaox akhdr terpenequeen kjhhl dsz

# State Wise Tweet Collection and Analysis

## Geographic Variation Table:

**Selection of States:**

* States were selected based on the Democratic or Republican support sourcing from <https://en.wikipedia.org/wiki/Red_states_and_blue_state>.
* Since the [2000 United States presidential election](https://en.wikipedia.org/wiki/United_States_presidential_election,_2000), red states and blue states have referred to [states](https://en.wikipedia.org/wiki/U.S._state) of the [United States](https://en.wikipedia.org/wiki/United_States) whose residents predominantly vote for either the [Republican Party](https://en.wikipedia.org/wiki/Republican_Party_(United_States)) (red) or [Democratic Party](https://en.wikipedia.org/wiki/Democratic_Party_(United_States)) (blue) presidential candidates. Since then, the use of the term has been expanded to differentiate between states being perceived as [liberal](https://en.wikipedia.org/wiki/Modern_liberalism_in_the_United_States) and those perceived as [conservative](https://en.wikipedia.org/wiki/Conservatism_in_the_United_States) - Wikipedia

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **State** | **South West** | | **North East** | | **National Popular vote** |
|  | **Longitude** | **Latitude** | **Longitude** | **Latitude** |
| **NY** | -74.156292 | 40.5543273 | 44.9864273 | -73.4465251 | **Clinton** |
| **Cal** | -123.8906282 | 30.76638363 | -117.688944 | 34.0122346 | **Clinton** |
| **Illinois** | -90.21595 | 38.166165 | -87.312178 | 42.351516 | **Clinton** |
| **Texas** | -98.230012 | 26.203407 | -95.769713 | 34.93343 | **Trump** |
| **Georgia** | -84.280733 | 30.438256 | -82.39401 | 34.852618 | **Trump** |

* State wise tweets were collected using the above-mentioned latitude and longitudes using the locations keyword and filter method under Twython.
* To validate the tweets were collected by converting the keyword and collected tweet text to lower case and then validated if the keyword exists in the tweet text.
* For each state, apart from the general stopwords few other stopwords were identified manually based on their occurrence in the WordCloud.
* Major challenge faced during tweet collection was the rate of tweet collection as geo tagging is not available in re-tweets and many tweets without geo tags having the required keyword must have been not picked up. However, patience attributed in successful collection of 10,000 tweets for each of the selected states.

## New York

### Overview

*As of April 2016,*[*Democrats*](https://en.wikipedia.org/wiki/Democratic_Party_(United_States))*represented a plurality of voters in New York State, constituting over twice as many*[*registered voters*](https://en.wikipedia.org/wiki/Voter_registration)*as any other political party affiliation*[*or lack thereof*](https://en.wikipedia.org/wiki/Independent_politician)*.*[*[182]*](https://en.wikipedia.org/wiki/New_York#cite_note-NYSRegisteredVoterAffiliation-182)*Since the second half of the 20th century, New York has generally supported candidates belonging to the Democratic Party in national elections. Democratic presidential candidate*[*Barack Obama*](https://en.wikipedia.org/wiki/Barack_Obama)*won New York State by over 25 percentage points in both 2012 and 2008. https://en.wikipedia.org/wiki/New\_York#Politics*

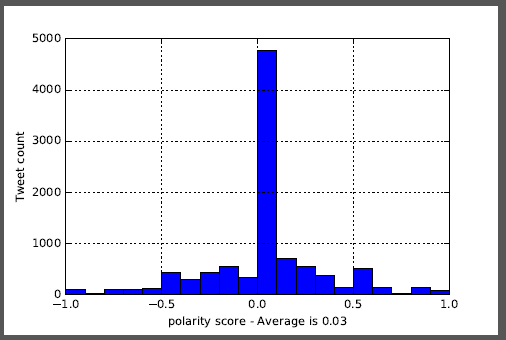
**Identified ignore words and accept words are mentioned below:**

ignore\_words = ['RT', 'rt','http', 'https', 'trump','donald', 'realdonaldtrump', 'thi', 'amp','hi','whi','co','ha','potus','potu','wa','president']

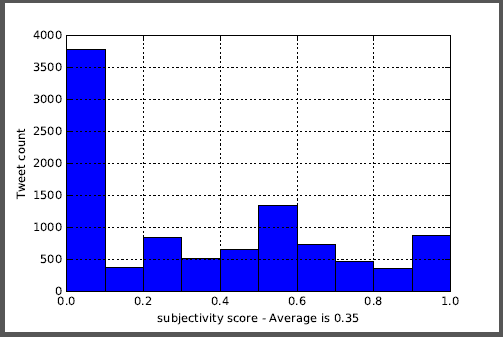
accept\_words = ['trumpcare', 'obamacare','notmypresident']

### Sentiment Analysis

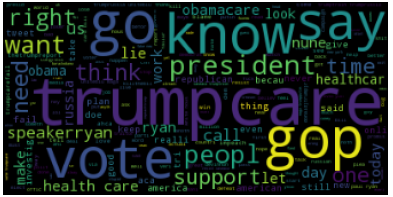
* **Polarity Scores Histogram**



* **Subjectivity Scores Histogram**



### Word Cloud



### Topic Modelling

**NFM Generated Topics**

1. trumpcare fail vote obamacare gop lose
2. trumprussia trumptrain uniteblu resist maga saveaca
3. like know peopl obamacare say time
4. win loser tire lose speakerryan trumpcarefail
5. care vote health ryan hous gop

**Interpretation of Topics:**

* Topics 1, 4, 5 talks about the trumpcare health bill proposed by President Trump and the fact that Paul Ryan cancelled the House Health Bill vote.
* Topic 2 talks about the Trump- Russia Links, MAGA and Save the Affordable Care Act.

## Georgia

### Overview

*In the 21st century, many conservative Democrats, including former U.S. Senator and governor*[*Zell Miller*](https://en.wikipedia.org/wiki/Zell_Miller)*, decided to support Republicans. https://en.wikipedia.org/wiki/Georgia\_(U.S.\_state)#Politics*

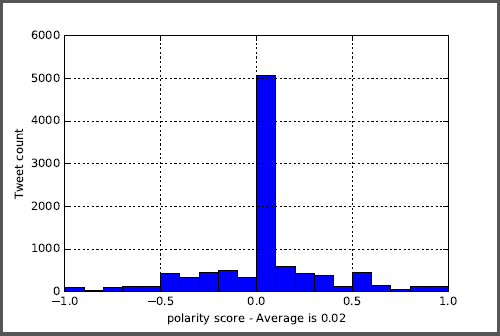
**Identified ignore words and accept words are mentioned below:**

ignore\_words = ['RT', 'rt','http', 'https', 'trump','donald', 'realdonaldtrump', 'thi', 'amp','hi','whi','co','ha','potus','potu','wa', 'president', 'know','america','american']

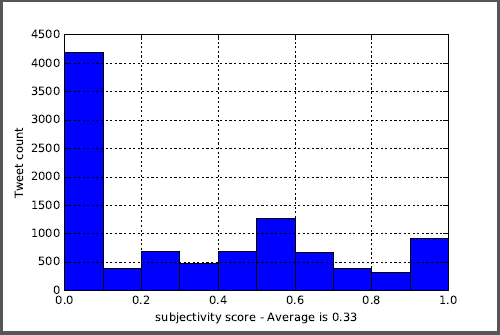
accept\_words = ['trumpcare', 'obamacare','notmypresident']

### Sentiment Analysis

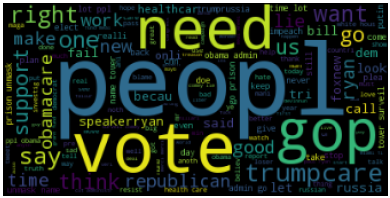
* **Polarity Scores Histogram**



* **Subjectivity Score Histogram**



### Word Cloud



### Topic Modelling

**NFM Generated Topic**

1. time admin unmask surveil prison tower
2. job hire miami hotel fl hospit
3. cut administr fund meal grant wheel
4. care peopl like vote gop need
5. trumpcare rep wld insur save bc

**Interpretation of Topics:**

* Topic 1 talks about wire tapping of Trump before the elections news.
* Many job posting for Trump Hotels were posted which have been picked up as a topic 2.
* Topic 3 talks about Trump Proposal cut on grants for meal on wheel.
* Topic 4 talks about Paul Ryan cancelled the House Health Bill vote.
* Topic 5 is due to the news that Trump supporting states will have to pay more for insurance

## California

### Overview

*The state's trend towards the*[*Democratic Party*](https://en.wikipedia.org/wiki/Democratic_Party_(United_States))*and away from the*[*Republican Party*](https://en.wikipedia.org/wiki/Republican_Party_(United_States))*can be seen in state elections. From 1899 to 1939, California had Republican governors. Since 1990, California has generally elected Democratic candidates to federal, state and local offices, including current Governor*[*Jerry Brown*](https://en.wikipedia.org/wiki/Jerry_Brown)*; however, the state has elected Republican Governors, though many of its Republican Governors, such as*[*Arnold Schwarzenegger*](https://en.wikipedia.org/wiki/Arnold_Schwarzenegger)*, tend to be considered moderate Republicans and more*[*centrist*](https://en.wikipedia.org/wiki/Centrism)*than the national party. https://en.wikipedia.org/wiki/California#Government\_and\_politics*

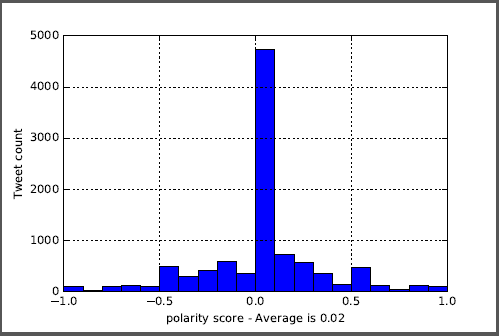
**Identified ignore words and accept words are mentioned below:**

ignore\_words = ['RT', 'rt','http', 'https', 'trump','donald', 'realdonaldtrump', 'thi', 'amp','hi','whi','co','ha','potus','potu','wa','president']

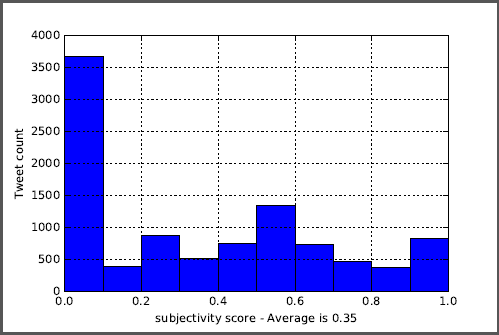
accept\_words = ['trumpcare', 'obamacare','notmypresident']

### Sentiment Analysis

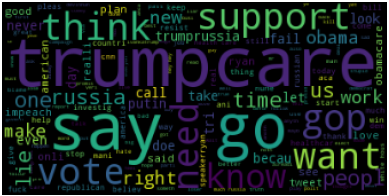
* **Polarity Score Histogram**



* **Subjectivity Score Histogram**



### Word Cloud



### Topic Modelling

1. hey impeachtrump pay russia resist tie
2. trumpcare vote obamacare gop fail killthebil
3. like say care lie support time
4. investig nune fbi video cooper zapa
5. peopl american need know job hear

**Interpretation of Topics:**

* Topics 1 and 3 talks about Trump ties with Russia and FBI investigation on the same.
* Topic 2 talks about the GOP loose and Trump Care fails
* Topic 5 talks about the jobs to the Americans

## Texas

### Overview

*Although Texas was essentially a one-party state during this time and the Democratic primary was viewed as the real election, the Democratic Party had conservative and liberal factions, which became more pronounced after the*[*New Deal*](https://en.wikipedia.org/wiki/New_Deal)*.*[*[137]*](https://en.wikipedia.org/wiki/Texas#cite_note-PoliticalPartiesTSHA-138)*Additionally, several factions of the party briefly split during the 1930s and 40s.*[*[137]*](https://en.wikipedia.org/wiki/Texas#cite_note-PoliticalPartiesTSHA-138)

*The state's conservative white voters began to support Republican presidential candidates by the mid-20th century. After this period, they supported Republicans for local and state offices as well, and most whites became Republican Party members.*[*[138]*](https://en.wikipedia.org/wiki/Texas#cite_note-SouthWon-139)*The party also attracted some minorities, but many have continued to vote for Democratic candidates. The shift to the Republican Party is much-attributed to the fact that the Democratic Party became increasingly*[*liberal*](https://en.wikipedia.org/wiki/Liberalism_in_the_United_States)*during the 20th century, and thus increasingly out-of-touch with the average Texas voter.*[*[139]*](https://en.wikipedia.org/wiki/Texas#cite_note-HistoryTXvoters-140)*As Texas was always a*[*conservative*](https://en.wikipedia.org/wiki/Conservatism_in_the_United_States)*state, voters switched to the GOP, which now more closely reflected their beliefs. https://en.wikipedia.org/wiki/Texas#Politics*

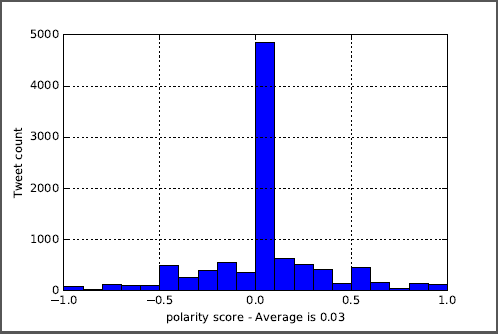
**Identified ignore words and accept words are mentioned below:**

ignore\_words = ['RT', 'rt','http', 'https', 'trump','donald', 'realdonaldtrump', 'thi', 'amp','hi','whi','co','ha','potus','potu','wa', 'president','go','need','know','say']

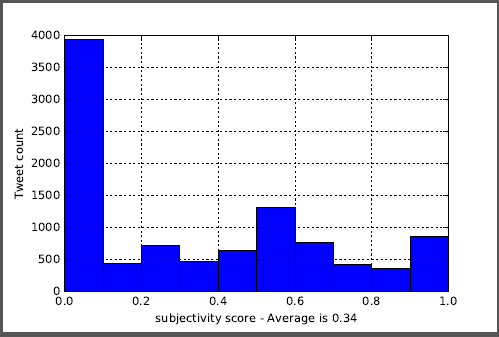
accept\_words = ['trumpcare', 'obamacare','notmypresident']

### Sentiment Analysis

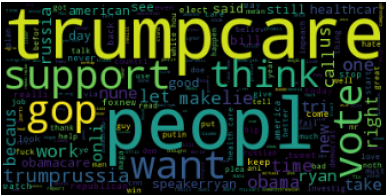
* **Polarity Scores Histogram**



* **Subjectivity Sores Histogram**



### Word Cloud



### Topic Modelling

1. trumpcare vote pleas notrumpcar gop lose
2. trumprussia theresist right russiag investig fakehealthcar
3. play song anymor griev father golf
4. peopl care obamacare make want repeal
5. like look feel realli make tweet

**Interpretation of Topics:**

* Topic 1 talks about trumpcare fails and GOP loose.
* Topic 2 talks about the Trump Russia ties which people are dissatisfied about
* Topic 3 Due to trump care bill a trump supporter who used to sing Trump songs have now stopped due to the new health care bill.

## Illinois

### Overview:

*Illinois is a*[*Democratic stronghold*](https://en.wikipedia.org/wiki/Red_states_and_blue_states)*and considered one of the most Democratic states in the nation. Historically, Illinois was a political*[*swing state*](https://en.wikipedia.org/wiki/Swing_state)*, with near-parity existing between the*[*Republican*](https://en.wikipedia.org/wiki/Republican_Party_(United_States))*and the*[*Democratic*](https://en.wikipedia.org/wiki/Democratic_Party_(United_States))*parties. However, in recent elections, the Democratic Party has gained ground and Illinois has come to be seen as a solid*[*blue state*](https://en.wikipedia.org/wiki/Red_states_and_blue_states)*in presidential contests.* - *https://en.wikipedia.org/wiki/Illinois#Politics*

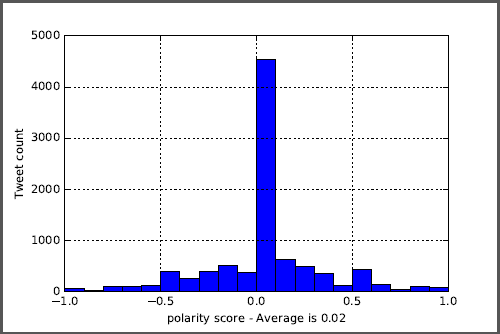
**Identified ignore words and accept words are mentioned below:**

ignore\_words = ['RT', 'rt','http', 'https', 'trump','donald', 'realdonaldtrump', 'thi', 'amp','hi','whi','co','ha','potus','potu','wa','president','know','go','say']

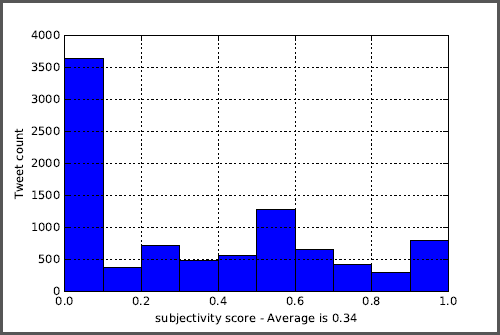
accept\_words = ['trumpcare', 'obamacare','notmypresident']

### Sentiment Analysis

* **Polarity Scores Histogram**



* **Subjectivity Scores Histogram**



### Word Cloud



### Topic Modelling

* 1. trumpcare vote care health gop obamacare
  2. need becaus russian wire tap spi
  3. peopl lie obama russia right time
  4. like look sound act doe putin
  5. job chicago hotel hire il work

**Interpretation of Topics:**

* Topics 2, 3, and 4 are based on the Trump Russia Ties and wire tapping.
* Topic 1 is based on the TrumpCare fail and GOP loose.
* Many job posting for Trump Hotels were posted which have been picked up as a topic 5.

# Insights

**General Tweets observations:**

* An interesting thing to notice is not within the states but comparing USA states with the rest of the word results – The world is talking more about the travel ban policy that President Trump had put forth and the stay on the same by Federal Judges twice whereas the important topics discussed within USA are more related to US policies specifically the recent Trumpcare which makes sense as any nation is more concerned about the policies being implemented within the country compared to foreign policies.
* The worldwide tweets talk more about the Travel Ban on 6 Muslim nations, Carlson Tucker and President Trump’s claim for wiretapping, Snoop Dogg videos which displays President Trump being wearing a clown attire, Comments made by Arabic prince in support of Trump’s travel Ban and referring him as a friend and the outrage of the Queen.
* General observation is that the tweets are more centred towards world politics and foreign policies.
* Most of the state are not discussing any of the topics mentioned above.

**State- wise Tweets observation:**

* States NY, California, Illinois, and Texas talks about Trump- Russia links and criticism over the same. Out of all Texas is the only Trump Supporting State where as Georgia being a Trump supporter refrains from majorly commenting on the current trending topic of Trump- Russia Ties.
* Interestingly Georgia and Illinois has substantial number of tweets related to job hiring in Trump Hotel which came out in the top topics. Considering 10,000 tweets for each state the number of job postings in them must be very high to be picked up in the top topics.
* All the states have been talking about the recent TrumpCare bill proposed by President Trump and Speaker Paul Ryan cancelling the House Health Bill Vote irrespective of historical support/oppose to Trump.
* Georgia being a trump supporting state talks about the news that Trump supporting states will have to pay more insurance which the other states are not discussing as much.
* Texas has tweeted enough to have a top topic on a trump supporter who used to sing Trump songs have now stopped due to the new health care bill.
* California known for major Information Technological presence also tweeted about bring jobs back to the Americans. This was a key topic during Trump’s Presidential elections campaign and even though California is not a trump supporting state it still have been to support this ideology.