Project 1: Social Media Analytics

**Group No.16**

**Kartik Doulat Chavan**

**Vidyalakshmi Krishnakumar**

**Roma Pradeep Patil**

Contents

[1 Introduction: 3](#_Toc478765130)

[2 Data Collection: 3](#_Toc478765131)

[3 Sentiment Analysis & Word Cloud: 3](#_Toc478765132)

[3.1 Global Sentiment Analysis: 3](#_Toc478765133)

[3.2 Global Word Cloud: 3](#_Toc478765134)

[4 Sentiment Analysis of US States: 4](#_Toc478765135)

[4.1.1 Polarity: 4](#_Toc478765136)

[4.1.2 Subjectivity: 5](#_Toc478765137)

[4.2 Word Cloud of US States: 6](#_Toc478765138)

[4.2.1 Blue States 6](#_Toc478765139)

[4.2.2 Red States 6](#_Toc478765140)

[5 Topic Modeling 8](#_Toc478765141)

[6 Insights 9](#_Toc478765142)

[7 References: 9](#_Toc478765143)

# Introduction:

In this project, we have used the Twitter Streaming API to collect the tweets about US President Donald Trump. This guide us towards the first step of having a Twitter account for us to create an application. This application provides us with the keys and tokens which are the required credentials to collect tweets for further analysis. We used string library to remove punctuation and digits. Along with the additional list created in the program, nltk provided us with the packages to remove stopwords. Then we performed sentiment analysis using TextBlob to measure the polarity and subjectivity of the collected tweets. We plotted the histograms of polarity and subjectivity using matplotlib. This analysis was first performed for the 10,000 collected tweets which was then followed by analyzing differently for six pars of different states across the USA to understand variation in the opinion. Further we created word cloud using WordCloud and nltk libraries. Further we conducted topic modeling on the tweets using Non-Negative Matrix Factorization from Scikit\_learn and Latent Dirichlet Allocation from Gensim.

# Data Collection:

The tweets were collected using the Twython module. Twython is the premier Python library providing an easy (and up-to-date) way to access Twitter data. Actively maintained and featuring support for Python 2.6+ and Python 3. Twitter lists being one of the many features of Twython module helps us collect the tweets which are stored in the json file. To narrow down the tweet collection, we used the track =’trump’ where trump is the keyword which helped us collect all the tweets related to President Donald Trump. We combined several states for tweet collection. The collection of location wise tweets is according to Red and Blue states.

Following is the list of Blue States:

1. California and Washington
2. Massachusetts and New Jersey
3. New York and Virginia

Following is the list of Red States:

1. Pennsylvania and Ohio
2. Texas and Louisiana
3. Utah and Arizona

# Sentiment Analysis & Word Cloud:

## Global Sentiment Analysis:

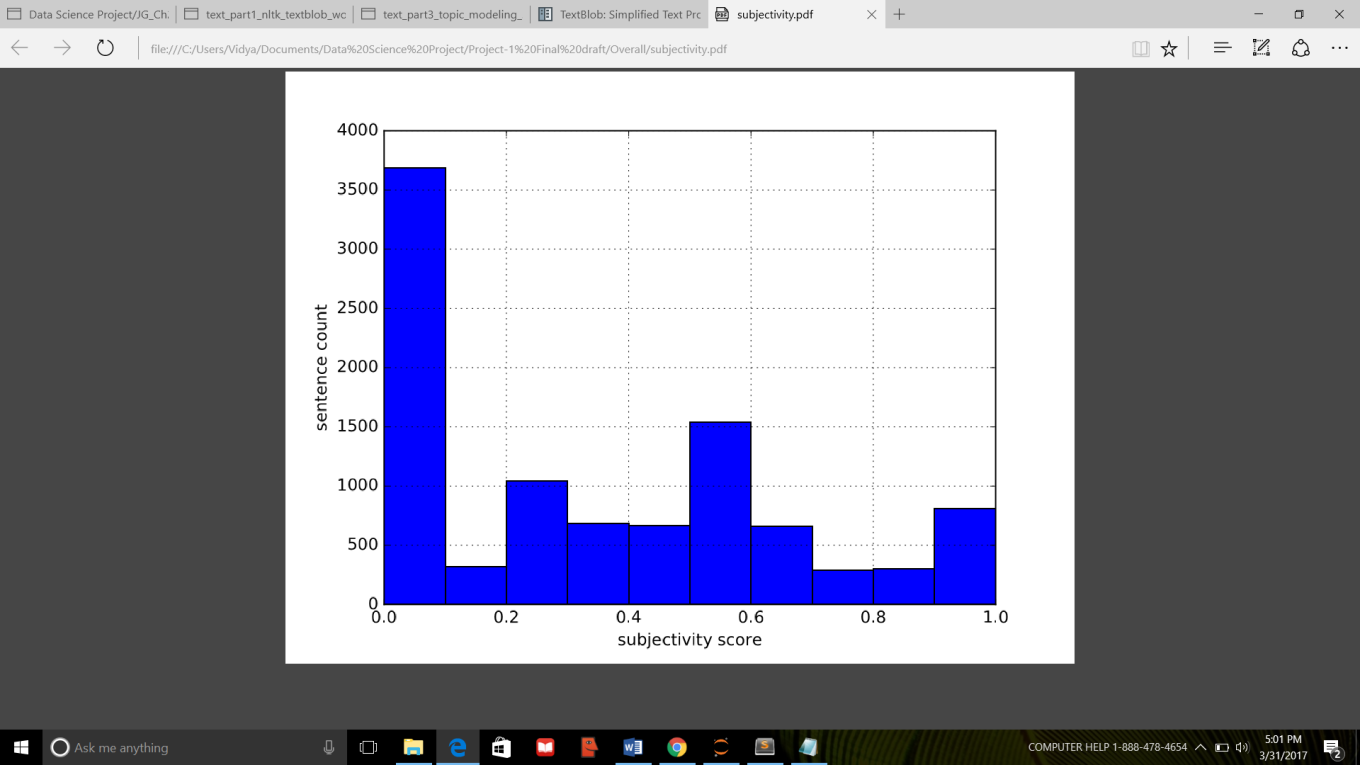
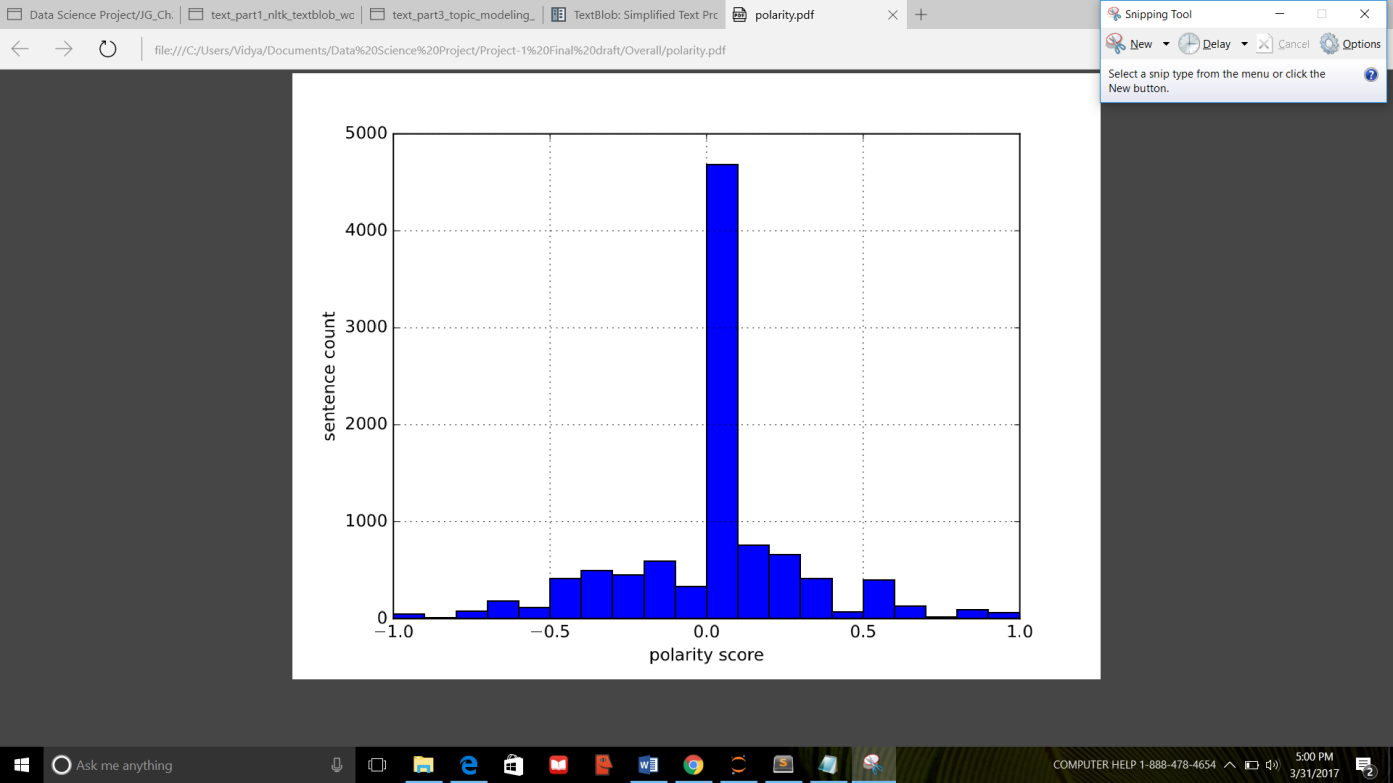
Since we are using TextBlob for sentiment analyzing, not all the words are ranked with the polarity and subjectivity. Hence, we see a high frequency at polarity =0 and subjectivity =0.

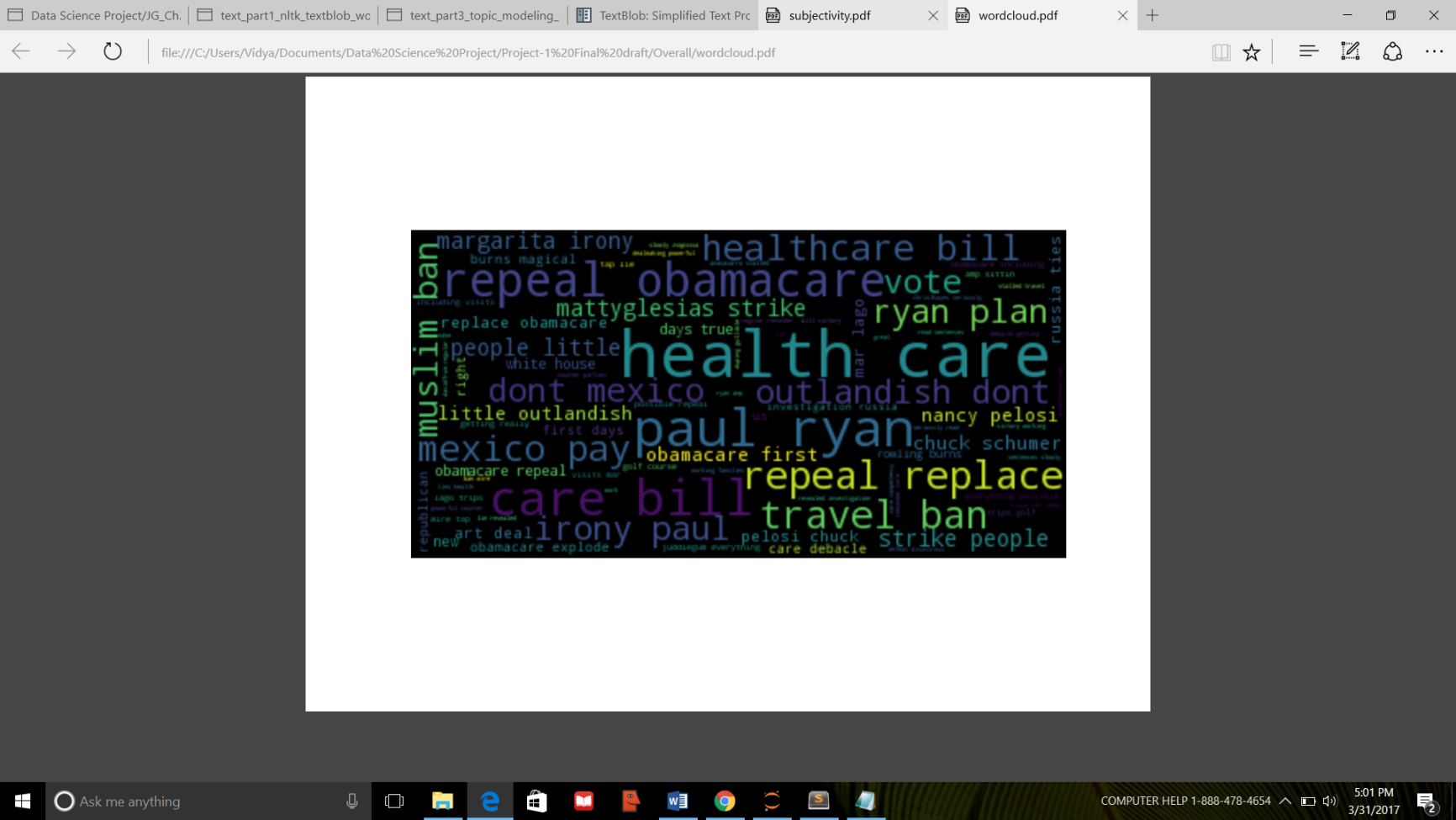
Global Polarity: According to the graph shown below, we observe scattered distribution of the polarity, with the average polarity equals to 0.00889909921553.

Global Subjectivity: According to the graph shown below, we observe scattered distribution of subjectivity, with the average subjectivity equals to 0.334338079988.

## Global Word Cloud:

The word cloud here shows that most of the tweets contain the words or phrases like health care, repeal Obamacare, healthcare bill, Muslim ban, travel ban etc.





# Sentiment Analysis of US States:

Since we are using TextBlob for sentiment analyzing, not all the words are ranked with the polarity and subjectivity. Hence, we see a high frequency at polarity =0 and subjectivity =0.

### Polarity:

#### Blue States

1) California and Washington: According to the graph shown below, we observe scattered distribution of the polarity, with the average polarity equals to 0.0165527719691.

2) Massachusetts and New Jersey: According to the graph shown below, we observe scattered distribution of the polarity, with the average polarity equals to 0.0123465942787.

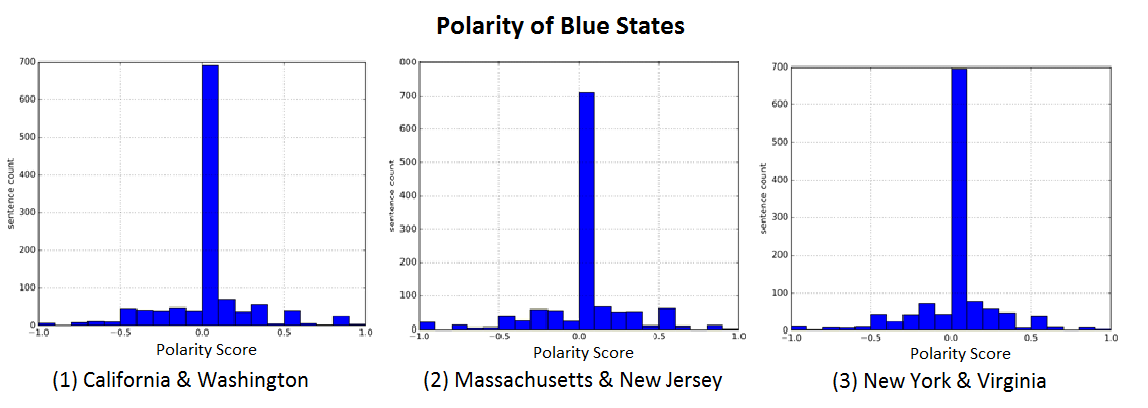
3) New York and Virginia: According to the graph shown below, we observe scattered distribution of the polarity, with the average polarity equals to 0.00967384529822.

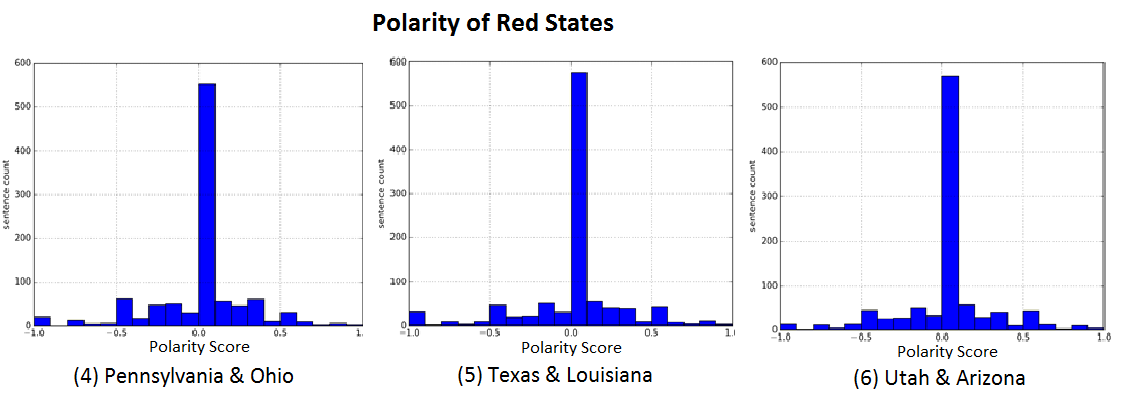
#### Red States

4) Pennsylvania and Ohio: According to the graph shown below, we observe scattered distribution of the polarity, with the average polarity equals to -0.0038193397532.

5) Texas and Louisiana: According to the graph shown below, we observe scattered distribution of the polarity, with the average polarity equals to -0.00162209522504.

6) Utah and Arizona: According to the graph shown below, we observe scattered distribution of the polarity, with the average polarity equals to 0.0079255389055.





### Subjectivity:

#### Blue States

1) California and Washington: According to the graph shown below, we observe scattered distribution of the subjectivity, with the average subjectivity equals to 0.307584722951.

2) Massachusetts and New Jersey: According to the graph shown below, we observe scattered distribution of the subjectivity, with the average subjectivity equals to 0.316684263695.

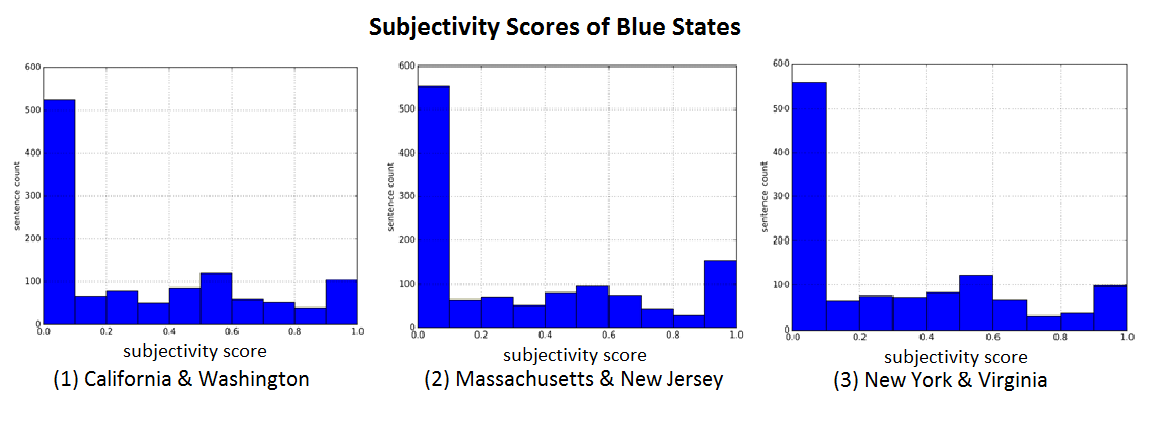
3) New York and Virginia: According to the graph shown below, we observe scattered distribution of the subjectivity, with the average subjectivity equals to 0.285811110911.

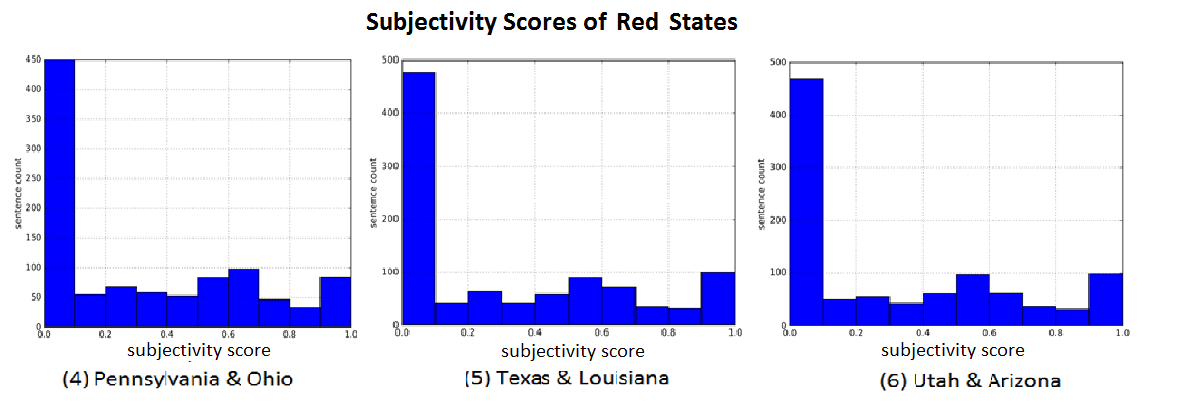
#### Red States

4) Pennsylvania and Ohio: According to the graph shown below, we observe scattered distribution of the subjectivity, with the average subjectivity equals to 0.31114620257.

5) Texas and Louisiana: According to the graph shown below, we observe scattered distribution of the subjectivity, with the average subjectivity equals to 0.301285402189.

6) Utah and Arizona: According to the graph shown below, we observe scattered distribution of the subjectivity, with the average subjectivity equals to 0.301195853105.





## Word Cloud of US States:

### Blue States

1) California and Washington: The word cloud below shows that most of the tweets contain the words or phrases like Russia, Obama, people , supporter etc

2) Massachusetts and New Jersey: The word cloud below shows that most of the tweets contain the words or phrases like Obama, Ivanka, white house, administration etc

3) New York and Virginia: The word cloud below shows that most of the tweets contain the words or phrases like white house, Ivanka , Russian, approval etc.

### Red States

4) Pennsylvania and Ohio: The word cloud below shows that most of the tweets contain the words or phrases like climate change, Russia, people , new etc.

5) Texas and Louisiana: The word cloud below shows that most of the tweets contain the words or phrases like Obama, Russia, people , administration etc.

6) Utah and Arizona: The word cloud below shows that most of the tweets contain the words or phrases like Obama, Russian , white house, job etc.



# Topic Modeling

Topic Modeling was performed considering two models namely, Non-Negative Matrix Factorization, and Latent Dirichlet Allocation.

Through the process of analyzing NMF, we obtained the topics which contained the words providing less clarity of the topics being discussed, as there was less correlation between the words.

On the other hand, LDA portrayed high correlation between words, resulting in better understanding of the subject.

We observed 3 factors that could help determine the best LDA’s topic model:

1.Time and number of Passes – The number of passes if kept between 100 and 150 gave a topic with an improved correlation between the words within an acceptable amount of time. Hence, we chose the count to be 120.

2.Number of topics- Considering the range between 13-17, which gave us a better understanding about the topic because of the increased number of correlated words, we kept the count as 15.

3.Density- We observed that the density of words in the topic is related to the above mentioned two factors. The combination of number of passes and number of topics resulted in better topics which guided us towards choosing the best topic model.

Hence, the best topic model for this project considered by us is LDA with number of topics equal to 15 and number of passes equal to 120.

Top 15 topics according to our LDA model are-

Topic #1 (0, u'0.012\*"care" + 0.010\*"health" + 0.010\*"getting" + 0.009\*"travel" + 0.009\*"debacle" + 0.009\*"ban" + 0.008\*"lie" + 0.008\*"revealed" + 0.008\*"tap" + 0.008\*"stalled"')

Topic #2 (1, u'0.023\*"days" + 0.020\*"replace" + 0.018\*"repeal" + 0.014\*"true" + 0.012\*"obamacare" + 0.010\*"sopandeb" + 0.009\*"would" + 0.008\*"within" + 0.008\*"first" + 0.007\*"kdik"')

Topic #3 (2, u'0.009\*"spectacular" + 0.009\*"better" + 0.008\*"ryan" + 0.008\*"treason" + 0.008\*"tedlieu" + 0.008\*"focus" + 0.008\*"thehill" + 0.007\*"issu" + 0.007\*"nytimes" + 0.007\*"go"')

Topic #4 (3, u'0.006\*"theresistance" + 0.005\*"ivanka" + 0.005\*"crowd" + 0.005\*"resist" + 0.004\*"bill" + 0.004\*"coming" + 0.004\*"thank" + 0.004\*"care" + 0.004\*"hey" + 0.004\*"breaking"')

Topic #5 (4, u'0.010\*"deal" + 0.009\*"art" + 0.007\*"ryan" + 0.007\*"obamacare" + 0.006\*"imploding" + 0.005\*"mmflint" + 0.005\*"much" + 0.005\*"reporter" + 0.005\*"gave" + 0.005\*"whether"')

Topic #6 (5, u'0.012\*"schumer" + 0.011\*"chuck" + 0.010\*"pelosi" + 0.010\*"nancy" + 0.010\*"reminder" + 0.009\*"losers" + 0.008\*"davidfrum" + 0.008\*"parties" + 0.008\*"competency" + 0.008\*"powerful"')

Topic #7 (6, u'0.013\*"muslim" + 0.013\*"ban" + 0.010\*"daughter" + 0.008\*"tiffany" + 0.008\*"gjsw" + 0.008\*"jwpbw" + 0.007\*"recap" + 0.007\*"battle" + 0.007\*"yeah" + 0.006\*"wiretapping"')

Topic #8 (7, u'0.014\*"burns" + 0.014\*"rowling" + 0.014\*"magical" + 0.009\*"realantionship" + 0.009\*"jwdfzvbqu" + 0.006\*"zb" + 0.006\*"nnxnthe" + 0.006\*"followme" + 0.005\*"na" + 0.005\*"gon"')

Topic #9 (8, u'0.011\*"week" + 0.010\*"like" + 0.008\*"repeal" + 0.007\*"obamacare" + 0.006\*"rmugwmbrnq" + 0.006\*"office" + 0.006\*"alifaith" + 0.005\*"obama" + 0.005\*"analysis" + 0.005\*"first"')

Topic #10 (9, u'0.010\*"democrats" + 0.008\*"gop" + 0.008\*"support" + 0.007\*"short" + 0.006\*"lack" + 0.006\*"bill" + 0.006\*"fell" + 0.005\*"fault" + 0.005\*"obamacare" + 0.005\*"due"')

Topic #11 (10, u'0.013\*"victory" + 0.011\*"defeat" + 0.010\*"families" + 0.010\*"disastrous" + 0.010\*"care" + 0.010\*"working" + 0.009\*"health" + 0.009\*"sensanders" + 0.009\*"stood" + 0.008\*"bill"')

Topic #12 (11, u'0.023\*"irony" + 0.023\*"margarita" + 0.021\*"paul" + 0.021\*"plan" + 0.016\*"ryan" + 0.012\*"slowly" + 0.009\*"amp" + 0.008\*"failed" + 0.006\*"two" + 0.006\*"broken"')

Topic #13 (12, u'0.014\*"read" + 0.013\*"chrislhayes" + 0.013\*"sentences" + 0.013\*"zsqpsssa" + 0.013\*"seriously" + 0.009\*"lago" + 0.009\*"mar" + 0.009\*"golf" + 0.009\*"juddlegum" + 0.009\*"course"')

Topic #14 (13, u'0.017\*"dont" + 0.016\*"little" + 0.016\*"mattyglesias" + 0.016\*"strike" + 0.016\*"outlandish" + 0.016\*"pay" + 0.016\*"mexico" + 0.014\*"people" + 0.013\*"get" + 0.005\*"dealmaker"')

Topic #15 (14, u'0.009\*"pass" + 0.009\*"house" + 0.007\*"blame" + 0.007\*"republicans" + 0.007\*"trying" + 0.006\*"votes" + 0.006\*"senate" + 0.005\*"bill" + 0.005\*"democrats" + 0.005\*"loss"')

# Insights

1. We have observed that the average polarity score of red states is negative and that of blue states is positive. This observation is surprising as we expect positive polarity score from red states.
2. States Pennsylvania-Ohio and Texas-Louisiana show negative polarity despite them being red states, may be reason is increase in taxes for the retired, his negative comments about women, or his relations with Russia.
3. Comparing with other states, tweets from Utah and Arizona have unique topic on Obama’s defense deputy admitting about spying on Trump.
4. Texas and Luisiana being the red states and New Jersey and Massachusetts being the blue states and also geographically distant have similar topics being discussed.

# References:

1. <http://stackoverflow.com/questions/510348/how-can-i-make-a-time-delay-in-python>
2. <http://stackoverflow.com/questions/20199126/reading-json-from-a-file>
3. <http://stackoverflow.com/questions/493819/python-join-why-is-it-string-joinlist-instead-of-list-joinstring>
4. [https://dev.twitter.com/streaming/overview/request-parameters#track](https://dev.twitter.com/streaming/overview/request-parameters%23track)
5. <http://boundingbox.klokantech.com/>
6. <https://pypi.python.org/pypi/twython>
7. <http://localhost:8888/notebooks/Data%20Science%20Project/jg_ch09_getting_data/04_api/2_test_twitter_search.ipynb>
8. <http://localhost:8888/notebooks/Data%20Science%20Project/jg_ch09_getting_data/04_api/stream.ipynb>
9. <http://localhost:8888/notebooks/Data%20Science%20Project/JG_Ch20_NLP_Text_Mining/0_notebooks/text_part1_nltk_textblob_wordcloud_section001.ipynb>
10. http://localhost:8888/notebooks/Data%20Science%20Project/JG\_Ch20\_NLP\_Text\_Mining/0\_notebooks/text\_part3\_topic\_modeling\_Section002.ipynb