

INSY 5378-002: Data Science: A Programming Approach

Spring 2017

Group Project 1: Social Media Analytics

Members:

|  |  |  |
| --- | --- | --- |
| **Name** | **Major** | **Email** |
| Anish Grandhi | Information Systems | anish.grandhi@mavs.uta.edu |
| Meghna Goswami | Information Systems | meghna.goswami@mavs.uta.edu |
| Vinothkumaar Udhayakumar | Industrial Engineering | vinothkumaar.udhayakumar@mavs.uta.edu |

Table of Contents

[I. Data Collection 1](#_Toc478763313)

[II. Sentiment Analysis 1](#_Toc478763314)

[III. Word Cloud 3](#_Toc478763315)

[IV. Topic Modeling 3](#_Toc478763316)

[V. Data Collection on location based tweets 5](#_Toc478763317)

[VI. Sentiment Analysis of location based tweets 5](#_Toc478763318)

[VII. Word Clouds of all the location based tweets 10](#_Toc478763319)

[VIII. Topic Modeling on location based tweets 12](#_Toc478763320)

[IX. Insights 19](#_Toc478763321)

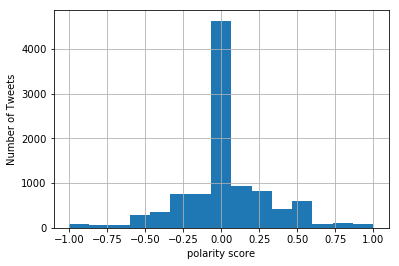
[X. References 19](#_Toc478763322)

# Data Collection

Twython module was used to collect 10K tweets from twitter using the keyword “trump”. The code can be found in CollectTweets.py file.

# Sentiment Analysis

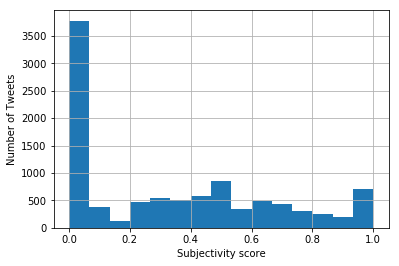
Using Textblob, polarity and subjectivity scores were computed on the 10K tweet corpus. The histograms below were drawn using Matplotlib where X axis is the score and Y axis is the count.



Figure

Figure 1 shows the polarity scores histogram. The scores show that around 4700 tweets have a neutral score (zero) while the rest are either positive or negative with a majority being near the polarity score of zero.

The **average polarity score** computed is 0.0235.



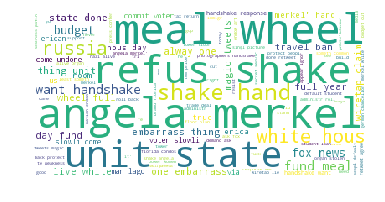
Figure

Figure 2 shows the subjectivity score histogram. Of the 10K tweets, approximately 3700 tweets have a subjectivity score of zero. The subjectivity of zero means that the tweets are objective in nature i.e., they state facts rather than opinion. The rest of the tweets have varying levels of subjectivity scores indicating that they are opinions to some degree. The subjectivity score of 1 shows that they are purely subjective in nature.

The **average subjectivity score** computed is 0.331

# Word Cloud

Each tweet was then cleaned by removing stop words using nltk package and then stemmed using the Porter stemmer. The words of the tweets were then fed into the Word Cloud module which generated the word cloud as shown –



Figure

# Topic Modeling

After removing stop-words and stemming, Non-negative Matrix Factorization (NMF) from Scikit-Learn and Latent Dirichlet Allocation (LDA) from GENSIM were used to conduct topic analysis. After the varying the number of topics, the following are the models of NMF and LDA:

NMF Model –

[u'merkel', u'wheel', u'claim', u'meet', u'fund', u'come', u'year', u'russia', u'lie', u'fox']

[u'wiretap', u'state', u'claim', u'meet', u'russia', u'lie', u'day', u'erica', u'look', u'thing']

[u'budget', u'state', u'wheel', u'administr', u'meet', u'year', u'russia', u'make', u'erica', u'tower']

[u'handshak', u'wheel', u'administr', u'come', u'fox', u'tower', u'unit', u'thing', u'report', u'germani']

[u'want', u'year', u'lie', u'russia', u'ask', u'day', u'make', u'erica', u'look', u'live']

[u'hand', u'administr', u'claim', u'meet', u'fund', u'year', u'lie', u'russia', u'fox', u'ask']

[u'shake', u'fund', u'come', u'day', u'ask', u'fox', u'live', u'alway', u'germani', u'report']

[u'angela', u'state', u'administr', u'meet', u'year', u'ask', u'make', u'erica', u'look', u'live']

[u'cut', u'administr', u'meet', u'year', u'ask', u'tower', u'erica', u'look', u'protect', u'10']

[u'like', u'wheel', u'claim', u'fund', u'fox', u'day', u'unit', u'tower', u'thing', u'report']

[u'news', u'state', u'come', u'russia', u'day', u'ask', u'alway', u'germani', u'protect', u'ban']

[u'meal', u'state', u'come', u'russia', u'make', u'live', u'thing', u'ban', u'man', u'erican']

[u'hous', u'claim', u'meet', u'fund', u'russia', u'lie', u'ask', u'live', u'look', u'thing']

[u'embarrass', u'state', u'wheel', u'come', u'day', u'unit', u'alway', u'report', u'germani', u'ban']

[u'white', u'fund', u'come', u'fox', u'day', u'make', u'unit', u'tower', u'report', u'german']

[u'refus', u'peopl', u'meet', u'russia', u'lie', u'ask', u'make', u'erica', u'live', u'look']

LDA Model –

[u'0.054\*"russia" ', u' 0.029\*"wiretap" ', u' 0.027\*"meal" ', u' 0.026\*"wheel" ', u' 0.024\*"hous" ', u' 0.024\*"someth" ', u' 0.023\*"white" ', u' 0.022\*"fund" ', u' 0.022\*"singl" ', u' 0.021\*"live"']

[u'0.062\*"3" ', u' 0.061\*"magic" ', u' 0.028\*"tower" ', u' 0.016\*"media" ', u' 0.016\*"realli" ', u' 0.014\*"eye" ', u' 0.013\*"l" ', u' 0.013\*"see" ', u' 0.010\*"snub" ', u' 0.009\*"refuge"']

[u'0.059\*"room" ', u' 0.027\*"fe" ', u' 0.024\*"build" ', u' 0.013\*"plan" ', u' 0.013\*"ili" ', u' 0.012\*"vote" ', u' 0.012\*"face" ', u' 0.012\*"best" ', u' 0.011\*"shit" ', u' 0.011\*"way"']

[u'0.035\*"erican" ', u' 0.020\*"maga" ', u' 0.016\*"point" ', u' 0.015\*"read" ', u' 0.013\*"great" ', u' 0.012\*"he" ', u' 0.011\*"chang" ', u' 0.010\*"stop" ', u' 0.009\*"busi" ', u' 0.008\*"scienc"']

[u'0.043\*"erica" ', u' 0.023\*"help" ', u' 0.021\*"tax" ', u' 0.020\*"maralago" ', u' 0.020\*"va" ', u' 0.019\*"today" ', u' 0.013\*"time" ', u' 0.013\*"talk" ', u' 0.013\*"return" ', u' 0.013\*"id"']

[u'0.052\*"merkel" ', u' 0.043\*"hand" ', u' 0.043\*"shake" ', u' 0.040\*"angela" ', u' 0.036\*"refus" ', u' 0.020\*"ignor" ', u' 0.017\*"press" ', u' 0.017\*"german" ', u' 0.014\*"chancellor" ', u' 0.014\*"power"']

[u'0.026\*"meet" ', u' 0.017\*"rt" ', u' 0.017\*"outrag" ', u' 0.016\*"first" ', u' 0.014\*"resist" ', u' 0.014\*"doesnt" ', u' 0.014\*"hotel" ', u' 0.012\*"pay" ', u' 0.011\*"kick" ', u' 0.010\*"agenc"']

[u'0.029\*"ban" ', u' 0.025\*"travel" ', u' 0.021\*"like" ', u' 0.017\*"putin" ', u' 0.015\*"weak" ', u' 0.015\*"germani" ', u' 0.014\*"te" ', u' 0.014\*"make" ', u' 0.013\*"well" ', u' 0.013\*"twitter"']

[u'0.062\*"true" ', u' 0.018\*"florida" ', u' 0.017\*"iliar" ', u' 0.017\*"brand" ', u' 0.016\*"spi" ', u' 0.016\*"condo" ', u' 0.016\*"word" ', u' 0.015\*"didnt" ', u' 0.014\*"shock" ', u' 0.014\*"come"']

[u'0.028\*"aign" ', u' 0.021\*"respond" ', u' 0.020\*"dem" ', u' 0.017\*"want" ', u' 0.015\*"fire" ', u' 0.014\*"new" ', u' 0.014\*"pull" ', u' 0.012\*"secretari" ', u' 0.012\*"hous" ', u' 0.012\*"play"']

[u'0.034\*"report" ', u' 0.020\*"r" ', u' 0.018\*"real" ', u' 0.016\*"agre" ', u' 0.016\*"work" ', u' 0.015\*"us" ', u' 0.014\*"cy" ', u' 0.013\*"e" ', u' 0.012\*"journal" ', u' 0.011\*"week"']

[u'0.036\*"embarrass" ', u' 0.030\*"ever" ', u' 0.030\*"thing" ', u' 0.030\*"state" ', u' 0.028\*"one" ', u' 0.027\*"unit" ', u' 0.027\*"alway" ', u' 0.026\*"done" ', u' 0.014\*"n" ', u' 0.013\*"world"']

[u'0.041\*"handshak" ', u' 0.028\*"get" ', u' 0.026\*"respons" ', u' 0.025\*"mer" ', u' 0.024\*"photograph" ', u' 0.024\*"stolen" ', u' 0.022\*"want" ', u' 0.022\*"organ" ', u' 0.021\*"aliv" ', u' 0.020\*"nazi"']

[u'0.039\*"budget" ', u' 0.035\*"news" ', u' 0.032\*"fox" ', u' 0.031\*"pictur" ', u' 0.031\*"cut" ', u' 0.029\*"lie" ', u' 0.022\*"ask" ', u' 0.020\*"demand" ', u' 0.019\*"er" ', u' 0.015\*"nigerian"']

[u'0.018\*"good" ', u' 0.017\*"tion" ', u' 0.016\*"leader" ', u' 0.014\*"wont" ', u' 0.013\*"alli" ', u' 0.013\*"v" ', u' 0.013\*"kill" ', u' 0.012\*"bullshit" ', u' 0.010\*"outlandish" ', u' 0.009\*"critic"']

[u'0.027\*"wall" ', u' 0.017\*"h" ', u' 0.016\*"support" ', u' 0.014\*"pathet" ', u' 0.013\*"small" ', u' 0.012\*"w" ', u' 0.012\*"htt" ', u' 0.012\*"behaviour" ', u' 0.011\*"apolog" ', u' 0.010\*"judg"']

[u'0.040\*"administr" ', u' 0.038\*"voter" ', u' 0.032\*"commit" ', u' 0.031\*"undon" ', u' 0.031\*"slowli" ', u' 0.029\*"come" ', u' 0.029\*"protect" ', u' 0.027\*"student" ', u' 0.026\*"back" ', u' 0.025\*"peopl"']

# Data Collection on location based tweets

Twython module was used to collect tweets from twitter using the keyword “trump” from **New Jersey, California, Central USA** (which consists the states of Texas, Colorado, Nebraska, Kansas, Oklahoma, Iowa, Missouri, Illinois, Mississippi). and **North West USA** (which consists the states of Washington, Oregon and Idaho). The code can be found in CollectTweets.py file. The latitude and longitude in the code was changed according to the region for which the tweets were being collected.

# Sentiment Analysis of location based tweets

Using Textblob, polarity and subjectivity scores were computed on the region wise tweet corpuses. The histograms below were drawn using Matplotlib where X axis is the score and Y axis is the count. The **polarity scores** are -

**New Jersey**:

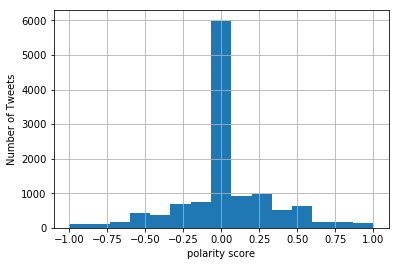
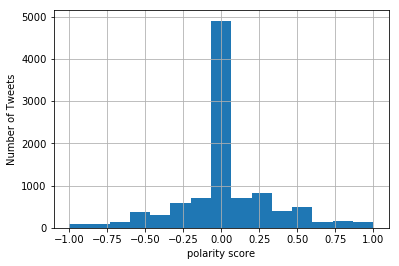


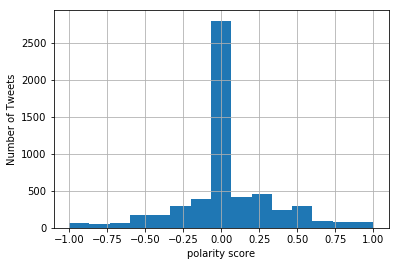
Figure 4

**California**:



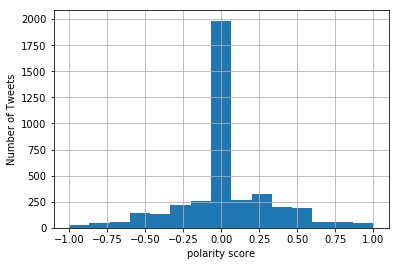
Figure

**Central**:



Figure

**North West**:



Figure

The **subjectivity scores** are as follows -

**New Jersey**:

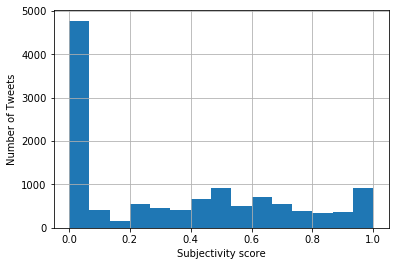
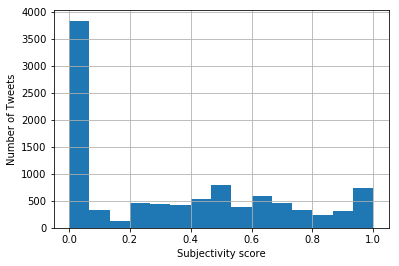


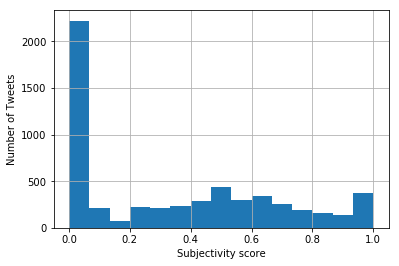
Figure 8

**California**:



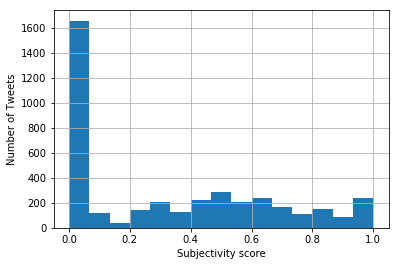
Figure

**Central**



Figure

**North West**



Figure

Averages of Polarity and Subjectivity of geographic regions –

|  |  |  |
| --- | --- | --- |
| Region | Average Polarity Score | Average Subjectivity score |
| New Jersey | 0.022 | 0.34 |
| California | 0.025 | 0.34 |
| Central USA | 0.03 | 0.33 |
| North West USA | 0.022 | 0.32 |

# Word Clouds of all the location based tweets

Each tweet was then cleaned by removing stop words using nltk package and then stemmed using the Porter stemmer. The words of the tweets were then fed into the Word Cloud module which generated the word cloud as shown –

New Jersey:

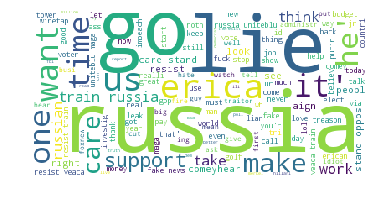


Figure 12

California:

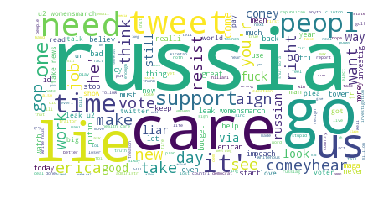


Figure 13

Central:

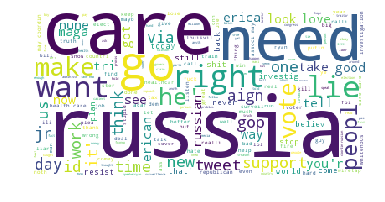


Figure 14

North West:

Before filtering out news channels since they do not make sense:



Figure 15

After filtering out news channels:



Figure 16

# Topic Modeling on location based tweets

On varying number of topics to get the optimal number, the following NMF and LDA models have been obtained on location based tweets:

**New Jersey**:

Topic Modeling NMF

[u'russia', u'work', u'win', u'think', u'stand', u'erican', u'elect', u'oppos', u'cut', u'fashion']

[u'care', u'work', u'fake', u'win', u'stand', u'countri', u'erican', u'elect', u'veaca', u'comey']

[u'resist', u'countri', u'world', u'tower', u'stop', u'id', u'look', u'come', u'putin', u'love']

[u'lie', u'wiretap', u'win', u'think', u'elect', u'oppos', u'cut', u'new', u'right', u'comey']

[u'like', u'think', u'world', u'new', u'vote', u'right', u'comey', u'look', u'gop', u'impeach']

[u'erica', u'wiretap', u'make', u'think', u'countri', u'elect', u'world', u'veaca', u'cut', u'comey']

[u'want', u'wiretap', u'win', u'think', u'world', u'comey', u'right', u'job', u'cut', u'budget']

[u'maga', u'countri', u'comeyhear', u'erican', u'elect', u'world', u'tower', u'tax', u'stop', u'id']

[u'train', u'win', u'comeyhear', u'new', u'veaca', u'vote', u'job', u'impeach', u'let', u'doesnt']

[u'russian', u'fake', u'win', u'comeyhear', u'veaca', u'right', u'job', u'gop', u'impeach', u'look']

[u'news', u'fake', u'countri', u'comeyhear', u'elect', u'new', u'vote', u'job', u'tower', u'stop']

[u'support', u'think', u'countri', u'erican', u'right', u'tower', u'stop', u'id', u'look', u'golf']

[u'time', u'wiretap', u'work', u'win', u'countri', u'elect', u'cut', u'comey', u'tax', u'budget']

[u'peopl', u'wiretap', u'investig', u'win', u'make', u'think', u'erican', u'comeyhear', u'elect', u'oppos']

Topic Modeling LDA

[u'0.035\*"maga" ', u' 0.034\*"resist" ', u' 0.032\*"train" ', u' 0.029\*"lie" ', u' 0.025\*"support" ', u' 0.023\*"wiretap" ', u' 0.019\*"uniteblu" ', u' 0.019\*"russia" ', u' 0.012\*"veaca" ', u' 0.011\*"fault"']

[u'0.020\*"comey" ', u' 0.015\*"hear" ', u' 0.015\*"leak" ', u' 0.014\*"collus" ', u' 0.011\*"health" ', u' 0.011\*"illeg" ', u' 0.010\*"give" ', u' 0.009\*"becom" ', u' 0.009\*"hate" ', u' 0.009\*"gop"']

[u'0.024\*"right" ', u' 0.015\*"elect" ', u' 0.013\*"suck" ', u' 0.013\*"get" ', u' 0.011\*"life" ', u' 0.010\*"yeah" ', u' 0.010\*"told" ', u' 0.008\*"anyth" ', u' 0.007\*"meal" ', u' 0.007\*"realli"']

[u'0.067\*"erica" ', u' 0.019\*"win" ', u' 0.015\*"fire" ', u' 0.015\*"ok" ', u' 0.013\*"new" ', u' 0.012\*"everyth" ', u' 0.011\*"f" ', u' 0.011\*"that" ', u' 0.011\*"russiag" ', u' 0.010\*"pleas"']

[u'0.049\*"e" ', u' 0.021\*"believ" ', u' 0.017\*"tower" ', u' 0.014\*"fbi" ', u' 0.014\*"seat" ', u' 0.014\*"jr" ', u' 0.012\*"continu" ', u' 0.010\*"way" ', u' 0.009\*"ball" ', u' 0.009\*"bullshit"']

[u'0.014\*"immigr" ', u' 0.013\*"peopl" ', u' 0.013\*"power" ', u' 0.012\*"women" ', u' 0.010\*"big" ', u' 0.009\*"wnyc" ', u' 0.009\*"must" ', u' 0.009\*"mar" ', u' 0.009\*"comment" ', u' 0.009\*"congress"']

[u'0.017\*"ralli" ', u' 0.012\*"cy" ', u' 0.012\*"myideaofhellwouldb" ', u' 0.011\*"low" ', u' 0.010\*"truth" ', u' 0.010\*"great" ', u' 0.010\*"pussi" ', u' 0.010\*"order" ', u' 0.009\*"joke" ', u' 0.009\*"live"']

[u'0.017\*"plan" ', u' 0.014\*"budget" ', u' 0.014\*"involv" ', u' 0.011\*"left" ', u' 0.010\*"fool" ', u' 0.010\*"n" ', u' 0.009\*"hirem" ', u' 0.008\*"voter" ', u' 0.008\*"report" ', u' 0.008\*"scandal"']

[u'0.023\*"fake" ', u' 0.021\*"investig" ', u' 0.020\*"news" ', u' 0.013\*"still" ', u' 0.013\*"2" ', u' 0.011\*"tie" ', u' 0.010\*"insur" ', u' 0.009\*"russia" ', u' 0.009\*"traitor" ', u' 0.009\*"putin"']

[u'0.024\*"erican" ', u' 0.015\*"dont" ', u' 0.011\*"ever" ', u' 0.010\*"u" ', u' 0.010\*"liar" ', u' 0.009\*"golf" ', u' 0.008\*"work" ', u' 0.008\*"time" ', u' 0.008\*"would" ', u' 0.008\*"think"']

[u'0.083\*"russia" ', u' 0.029\*"comeyhear" ', u' 0.020\*"id" ', u' 0.019\*"lose" ', u' 0.014\*"impeach" ', u' 0.011\*"start" ', u' 0.010\*"stop" ', u' 0.010\*"evil" ', u' 0.008\*"citizen" ', u' 0.008\*"last"']

[u'0.026\*"inc" ', u' 0.018\*"ili" ', u' 0.016\*"bad" ', u' 0.013\*"approv" ', u' 0.012\*"love" ', u' 0.011\*"lol" ', u' 0.011\*"rate" ', u' 0.009\*"1jul" ', u' 0.007\*"law" ', u' 0.007\*"alway"']

[u'0.037\*"aign" ', u' 0.013\*"today" ', u' 0.013\*"sinc" ', u' 0.011\*"suppos" ', u' 0.010\*"hope" ', u' 0.010\*"youv" ', u' 0.008\*"lawfirm" ', u' 0.008\*"keep" ', u' 0.008\*"repub" ', u' 0.008\*"slowli"']

[u'0.056\*"care" ', u' 0.026\*"stand" ', u' 0.023\*"oppos" ', u' 0.019\*"go" ', u' 0.016\*"loudobb" ', u' 0.015\*"treason" ', u' 0.015\*"thing" ', u' 0.012\*"done" ', u' 0.011\*"hous" ', u' 0.010\*"white"']

[u'0.017\*"tax" ', u' 0.015\*"first" ', u' 0.013\*"co" ', u' 0.011\*"administr" ', u' 0.011\*"resign" ', u' 0.011\*"russian" ', u' 0.011\*"never" ', u' 0.010\*"day" ', u' 0.010\*"ing" ', u' 0.009\*"asshol"']

**California**:

Topic Modeling with NMF

[u'russia', u'vote', u'day', u'russian', u'right', u'wiretap', u'fbi', u'gop', u'elect', u'come']

[u'care', u'vote', u'resist', u'want', u'job', u'impeach', u'think', u'good', u'fake', u'new']

[u'lie', u'vote', u'resist', u'impeach', u'right', u'think', u'work', u'elect', u'gop', u'realli']

[u'like', u'vote', u'russian', u'job', u'fbi', u'wiretap', u'think', u'work', u'elect', u'fake']

[u'peopl', u'day', u'russian', u'job', u'right', u'impeach', u'wiretap', u'fbi', u'gop', u'good']

[u'leak', u'day', u'russian', u'right', u'wiretap', u'think', u'gop', u'elect', u'come', u'good']

[u'time', u'russian', u'wiretap', u'work', u'fbi', u'gop', u'good', u'erican', u'great', u'liar']

[u'dont', u'want', u'resist', u'right', u'good', u'fake', u'new', u'erican', u'voter', u'fuck']

[u'aign', u'day', u'russian', u'job', u'wiretap', u'work', u'fbi', u'come', u'good', u'liar']

[u'comeyhear', u'russian', u'job', u'right', u'wiretap', u'fbi', u'gop', u'elect', u'fake', u'realli']

[u'erica', u'vote', u'news', u'day', u'right', u'fbi', u'think', u'work', u'elect', u'come']

[u'womensmarch', u'russian', u'job', u'wiretap', u'fbi', u'work', u'gop', u'come', u'good', u'fake']

Topic Modeling LDA

[u'0.016\*"wiretap" ', u' 0.013\*"id" ', u' 0.011\*"tie" ', u' 0.011\*"claim" ', u' 0.011\*"n" ', u' 0.011\*"approv" ', u' 0.010\*"man" ', u' 0.009\*"rate" ', u' 0.009\*"fuck" ', u' 0.009\*"go"']

[u'0.031\*"care" ', u' 0.013\*"right" ', u' 0.012\*"news" ', u' 0.012\*"fake" ', u' 0.011\*"time" ', u' 0.009\*"good" ', u' 0.009\*"your" ', u' 0.009\*"whitehous" ', u' 0.008\*"ralli" ', u' 0.007\*"hear"']

[u'0.028\*"investig" ', u' 0.015\*"fbi" ', u' 0.011\*"offic" ', u' 0.011\*"loser" ', u' 0.009\*"fail" ', u' 0.009\*"call" ', u' 0.008\*"give" ', u' 0.008\*"golf" ', u' 0.008\*"first" ', u' 0.008\*"west"']

[u'0.031\*"leak" ', u' 0.025\*"womensmarch" ', u' 0.017\*"u2" ', u' 0.014\*"win" ', u' 0.014\*"congress" ', u' 0.012\*"get" ', u' 0.012\*"ever" ', u' 0.011\*"ye" ', u' 0.010\*"shut" ', u' 0.010\*"elect"']

[u'0.020\*"cy" ', u' 0.017\*"fire" ', u' 0.008\*"show" ', u' 0.008\*"evid" ', u' 0.008\*"l" ', u' 0.008\*"vote" ', u' 0.008\*"clinton" ', u' 0.008\*"stupid" ', u' 0.007\*"et" ', u' 0.007\*"mean"']

[u'0.021\*"lie" ', u' 0.017\*"tower" ', u' 0.011\*"resist" ', u' 0.010\*"great" ', u' 0.010\*"white" ', u' 0.010\*"hous" ', u' 0.010\*"wrong" ', u' 0.010\*"bill" ', u' 0.009\*"republican" ', u' 0.008\*"move"']

[u'0.017\*"erican" ', u' 0.014\*"voter" ', u' 0.010\*"bad" ', u' 0.010\*"cnn" ', u' 0.009\*"us" ', u' 0.008\*"listen" ', u' 0.007\*"via" ', u' 0.007\*"ass" ', u' 0.006\*"chang" ', u' 0.006\*"fall"']

[u'0.017\*"acar" ', u' 0.013\*"carefail" ', u' 0.013\*"administr" ', u' 0.011\*"budget" ', u' 0.010\*"jr" ', u' 0.009\*"god" ', u' 0.008\*"seanspic" ', u' 0.008\*"replac" ', u' 0.007\*"histori" ', u' 0.006\*"thank"']

[u'0.086\*"russia" ', u' 0.021\*"comeyhear" ', u' 0.013\*"ili" ', u' 0.010\*"tax" ', u' 0.009\*"collus" ', u' 0.009\*"maga" ', u' 0.008\*"cut" ', u' 0.007\*"lose" ', u' 0.007\*"ask" ', u' 0.006\*"doesnt"']

[u'0.018\*"support" ', u' 0.011\*"better" ', u' 0.009\*"es" ', u' 0.009\*"accus" ', u' 0.008\*"stori" ', u' 0.008\*"ryan" ', u' 0.007\*"comey" ', u' 0.007\*"also" ', u' 0.006\*"fix" ', u' 0.006\*"caredef"']

[u'0.030\*"impeach" ', u' 0.015\*"russiag" ', u' 0.013\*"traitor" ', u' 0.010\*"theresist" ', u' 0.010\*"wall" ', u' 0.010\*"ster" ', u' 0.009\*"must" ', u' 0.007\*"worst" ', u' 0.007\*"arent" ', u' 0.007\*"border"']

[u'0.029\*"e" ', u' 0.013\*"today" ', u' 0.012\*"truth" ', u' 0.010\*"putin" ', u' 0.009\*"help" ', u' 0.008\*"impeach45" ', u' 0.007\*"ban" ', u' 0.007\*"alreadi" ', u' 0.006\*"point" ', u' 0.006\*"list"']

[u'0.031\*"erica" ', u' 0.016\*"liar" ', u' 0.008\*"w" ', u' 0.007\*"train" ', u' 0.006\*"use" ', u' 0.006\*"friend" ', u' 0.006\*"congratul" ', u' 0.006\*"crimin" ', u' 0.006\*"cost" ', u' 0.005\*"import"']

[u'0.029\*"aign" ', u' 0.015\*"treason" ', u' 0.011\*"ing" ', u' 0.010\*"day" ', u' 0.010\*"deal" ', u' 0.010\*"peopl" ', u' 0.009\*"attack" ', u' 0.008\*"love" ', u' 0.007\*"democrat" ', u' 0.007\*"ur"']

**Central USA**

Topic Modeling NMF

[u'care', u'support', u'aign', u'id', u'work', u'surveil', u'wiretap', u'erican', u'jr', u'ralli']

[u'russia', u'id', u'surveil', u'jr', u'ralli', u'think', u'look', u'day', u'pleas', u'world']

[u'like', u'id', u'surveil', u'ralli', u'think', u'look', u'pleas', u'hous', u'good', u'world']

[u'dont', u'id', u'surveil', u'maga', u'ralli', u'think', u'day', u'hous', u'good', u'nune']

[u'investig', u'support', u'time', u'aign', u'work', u'maga', u'erican', u'wiretap', u'health', u'nune']

[u'peopl', u'support', u'time', u'work', u'erican', u'ralli', u'health', u'think', u'look', u'day']

[u'want', u'support', u'aign', u'time', u'work', u'maga', u'wiretap', u'erican', u'health', u'jr']

[u'make', u'time', u'id', u'surveil', u'ralli', u'health', u'think', u'hous', u'love', u'day']

[u'russian', u'support', u'aign', u'time', u'maga', u'work', u'erican', u'wiretap', u'jr', u'health']

[u'gop', u'support', u'time', u'work', u'maga', u'erican', u'nune', u'day', u'love', u'tri']

[u'erica', u'id', u'surveil', u'ralli', u'think', u'look', u'pleas', u'world', u'louisvil', u'love']

Topic Modeling LDA

[u'0.013\*"traitor" ', u' 0.011\*"help" ', u' 0.010\*"even" ', u' 0.010\*"happen" ', u' 0.009\*"2020" ', u' 0.009\*"truck" ', u' 0.009\*"work" ', u' 0.009\*"yet" ', u' 0.008\*"told" ', u' 0.008\*"wrong"']

[u'0.019\*"administr" ', u' 0.011\*"today" ', u' 0.010\*"comey" ', u' 0.010\*"devinnun" ', u' 0.010\*"treason" ', u' 0.009\*"es" ', u' 0.009\*"shit" ', u' 0.009\*"agenda" ', u' 0.008\*"watch" ', u' 0.008\*"compromis"']

[u'0.017\*"win" ', u' 0.011\*"im" ', u' 0.009\*"start" ', u' 0.009\*"cnn" ', u' 0.009\*"ass" ', u' 0.009\*"line" ', u' 0.008\*"bad" ', u' 0.008\*"via" ', u' 0.008\*"stay" ', u' 0.008\*"fall"']

[u'0.019\*"lie" ', u' 0.015\*"ster" ', u' 0.010\*"et" ', u' 0.010\*"stop" ', u' 0.009\*"u" ', u' 0.008\*"back" ', u' 0.008\*"russiag" ', u' 0.007\*"cover" ', u' 0.007\*"knew" ', u' 0.007\*"best"']

[u'0.031\*"erica" ', u' 0.018\*"investig" ', u' 0.012\*"fbi" ', u' 0.012\*"pass" ', u' 0.011\*"hell" ', u' 0.010\*"russian" ', u' 0.010\*"account" ', u' 0.009\*"replac" ', u' 0.009\*"cnnpolit" ', u' 0.008\*"didnt"']

[u'0.097\*"russia" ', u' 0.024\*"surveil" ', u' 0.015\*"may" ', u' 0.012\*"believ" ', u' 0.011\*"tie" ', u' 0.011\*"coordin" ', u' 0.010\*"tower" ', u' 0.008\*"cy" ', u' 0.008\*"like" ', u' 0.008\*"commun"']

[u'0.016\*"time" ', u' 0.012\*"er" ', u' 0.012\*"get" ', u' 0.011\*"health" ', u' 0.011\*"man" ', u' 0.010\*"ing" ', u' 0.009\*"report" ', u' 0.008\*"tax" ', u' 0.008\*"kill" ', u' 0.008\*"care"']

[u'0.029\*"id" ', u' 0.014\*"moveon" ', u' 0.010\*"polit" ', u' 0.010\*"clinton" ', u' 0.009\*"reason" ', u' 0.009\*"vetheaca" ', u' 0.009\*"joke" ', u' 0.008\*"ge" ', u' 0.008\*"what" ', u' 0.007\*"cut"']

[u'0.023\*"ralli" ', u' 0.017\*"great" ', u' 0.012\*"louisvil" ', u' 0.011\*"jr" ', u' 0.011\*"theresist" ', u' 0.010\*"fire" ', u' 0.010\*"doesnt" ', u' 0.010\*"run" ', u' 0.009\*"job" ', u' 0.009\*"order"']

[u'0.033\*"right" ', u' 0.030\*"ili" ', u' 0.028\*"maga" ', u' 0.021\*"resist" ', u' 0.021\*"associ" ', u' 0.018\*"support" ', u' 0.017\*"your" ', u' 0.014\*"lol" ', u' 0.011\*"promis" ', u' 0.011\*"news"']

[u'0.139\*"care" ', u' 0.016\*"train" ', u' 0.013\*"anyth" ', u' 0.013\*"impeach" ', u' 0.011\*"hous" ', u' 0.011\*"vote" ', u' 0.010\*"bill" ', u' 0.010\*"ryancar" ', u' 0.008\*"stupid" ', u' 0.008\*"fail"']

[u'0.038\*"aign" ', u' 0.009\*"soon" ', u' 0.008\*"put" ', u' 0.008\*"eful" ', u' 0.007\*"must" ', u' 0.006\*"wouldnt" ', u' 0.006\*"in" ', u' 0.006\*"around" ', u' 0.005\*"inki" ', u' 0.005\*"die"']

[u'0.021\*"wiretap" ', u' 0.016\*"n" ', u' 0.013\*"speakerryan" ', u' 0.013\*"claim" ', u' 0.010\*"speak" ', u' 0.009\*"2" ', u' 0.009\*"comeyhear" ', u' 0.008\*"thing" ', u' 0.008\*"1" ', u' 0.008\*"h"']

**North West USA**

Topic Modeling NMF

[u'bc', u'zero', u'feder', u'fast', u'faster', u'fat', u'father', u'fault', u'favor', u'favorit']

[u'kremlinrussia\_', u'astoria', u'plan', u'zero', u'feder', u'fast', u'faster', u'fat', u'father', u'fault']

[u'russia', u'work', u'healthcar', u'real', u'right', u'fuck', u'erica', u'dont', u'girl', u'health']

[u'putinrf\_eng', u'ryan', u'right', u'erica', u'talk', u'plan', u'health', u'aca', u'id', u'time']

[u'like', u'work', u'healthcar', u'real', u'fuck', u'erica', u'right', u'girl', u'health', u'im']

[u'loser', u'work', u'right', u'fuck', u'dont', u'talk', u'erica', u'girl', u'fail', u'aca']

[u'acar', u'astoria', u'fuck', u'plan', u'aca', u'ilwaco', u'fail', u'id', u'girl', u'help']

[u'vote', u'ryan', u'astoria', u'girl', u'earth', u'id', u'dea', u'internet', u'read', u'great']

[u'make', u'healthcar', u'real', u'astoria', u'erica', u'health', u'fail', u'ilwaco', u'win', u'erican']

[u'paulryanpeak', u'ryan', u'right', u'talk', u'plan', u'health', u'time', u'id', u'shit', u'deal']

[u'lie', u'real', u'astoria', u'erica', u'talk', u'ilwaco', u'girl', u'aca', u'im', u'id']

[u'want', u'work', u'healthcar', u'right', u'dont', u'fuck', u'fail', u'im', u'use', u'stop']

[u'cop', u'ryan', u'astoria', u'plan', u'aca', u'time', u'help', u'deal', u'dea', u'think']

u'gop', u'ryan', u'right', u'dont', u'talk', u'plan', u'ilwaco', u'health', u'id', u'deal']

Topic Modeling LDA

[u'0.015\*"trend" ', u' 0.014\*"keep" ', u' 0.013\*"fbi" ', u' 0.013\*"investig" ', u' 0.012\*"russia" ', u' 0.011\*"killthebil" ', u' 0.011\*"ed" ', u' 0.010\*"done" ', u' 0.009\*"worri" ', u' 0.009\*"ryan"']

[u'0.015\*"mond" ', u' 0.014\*"ili" ', u' 0.011\*"way" ', u' 0.010\*"today" ', u' 0.010\*"l" ', u' 0.009\*"let" ', u' 0.008\*"better" ', u' 0.008\*"everi" ', u' 0.008\*"wrong" ', u' 0.008\*"go"']

[u'0.028\*"loser" ', u' 0.014\*"fail" ', u' 0.012\*"everyth" ', u' 0.012\*"travel" ', u' 0.011\*"time" ', u' 0.010\*"kill" ', u' 0.009\*"deal" ', u' 0.008\*"believ" ', u' 0.008\*"art" ', u' 0.008\*"man"']

[u'0.021\*"kremlinrussia\_" ', u' 0.020\*"bc" ', u' 0.016\*"dea" ', u' 0.016\*"astoria" ', u' 0.015\*"cop" ', u' 0.014\*"real" ', u' 0.012\*"nbc" ', u' 0.011\*"she" ', u' 0.010\*"abc" ', u' 0.009\*"check"']

[u'0.026\*"id" ', u' 0.016\*"bill" ', u' 0.011\*"see" ', u' 0.011\*"day" ', u' 0.010\*"video" ', u' 0.009\*"call" ', u' 0.009\*"pull" ', u' 0.008\*"healthcar" ', u' 0.008\*"offic" ', u' 0.007\*"militari"']

[u'0.035\*"erica" ', u' 0.034\*"ye" ', u' 0.017\*"neva" ', u' 0.009\*"paulryanpeak" ', u' 0.009\*"hero" ', u' 0.008\*"iran" ', u' 0.008\*"take" ', u' 0.008\*"putinrf\_eng" ', u' 0.008\*"best" ', u' 0.007\*"kremlinrussia\_"']

[u'0.011\*"25" ', u' 0.009\*"tax" ', u' 0.009\*"laugh" ', u' 0.009\*"march" ', u' 0.008\*"phone" ', u' 0.008\*"behind" ', u' 0.008\*"2017" ', u' 0.008\*"im" ', u' 0.008\*"busi" ', u' 0.008\*"weekli"']

[u'0.040\*"acar" ', u' 0.013\*"good" ', u' 0.013\*"mad" ', u' 0.013\*"look" ', u' 0.011\*"explod" ', u' 0.011\*"putinrf\_eng" ', u' 0.010\*"china" ', u' 0.009\*"love" ', u' 0.009\*"wow" ', u' 0.009\*"bc"']

[u'0.126\*"care" ', u' 0.023\*"carefail" ', u' 0.019\*"theresist" ', u' 0.017\*"resist" ', u' 0.016\*"russiag" ', u' 0.014\*"russia" ', u' 0.011\*"million" ', u' 0.011\*"health" ', u' 0.010\*"ing" ', u' 0.009\*"follow"']

[u'0.015\*"win" ', u' 0.013\*"lose" ', u' 0.011\*"sick" ', u' 0.010\*"clinton" ', u' 0.010\*"resign" ', u' 0.009\*"read" ', u' 0.009\*"scare" ', u' 0.008\*"togeth" ', u' 0.007\*"bigli" ', u' 0.007\*"never"']

[u'0.036\*"erican" ', u' 0.014\*"peopl" ', u' 0.013\*"ryan" ', u' 0.010\*"network" ', u' 0.009\*"address" ', u' 0.009\*"paul" ', u' 0.009\*"tire" ', u' 0.008\*"yet" ', u' 0.007\*"win" ', u' 0.007\*"ass"']

[u'0.058\*"russia" ', u' 0.012\*"girl" ', u' 0.011\*"failur" ', u' 0.011\*"cy" ', u' 0.009\*"putin" ', u' 0.009\*"next" ', u' 0.008\*"elect" ', u' 0.007\*"ever" ', u' 0.007\*"administr" ', u' 0.007\*"astoria"']

# Insights

1. All the polarity score histograms show that approximately 50%-60% of the tweets have a polarity score of zero while the rest 40-50% are distributed with positive or negative polarity scores.
2. From the subjectivity histograms, it shows that 30%-40% of the tweets have 0 subjectivity which means that they are objective in nature and the rest of the tweets are subjective to a degree.
3. The observation on topic models of the general corpus of 10K Trump tweets show that the topics tweeted about areare “Angela Merkel”. “Meal Wheel” and “Refus shake”. These are not surprising since President Trump has met with Angela Merkel and refused to shake hands with her. And the recent headlines that grabbed attention whether President was going to reduce the budget set aside for “Meals on Wheels” program and instead use it to build the Mexico Border Wall.
4. From the word clouds of location based tweets, it is evident that Russia is a high frequency occurring word across entire USA. North West region seems to be highly interested in News channels. After filtering out the news channels words, Russia again tops the list in the highest frequency words along with Kremlin. The name Kremlin means "fortress inside a city", and is often also used metonymically to refer to the government of the Russian Federation in a similar sense to how the White House is used to refer to the Executive Office of the President of the United States. Tweets from Central and California region seem to have a similar pattern since the word cloud has a high degree of similarity. Care, Need and Lie are high frequency words in California and Central region.
5. From the topic modelling, we can see that compared to other states, North-West states are more concerned about President Trump’s “health care”, “Paul Ryan” than the other states.
6. Both New Jersey and California have reacted more towards President Trump’s wiretap claims than the other states (Central states or North-West states) despite the geographic distance. It could probably because the both the states are more inclined against President Trump since the elections.
7. Similarly, New Jersey and California have mentioned about FBI Director Comey than compared to North-West states and states in Central USA.
8. The topic “Women’s March” is unique only in California.
9. One common topic across all the states is “Russia” which is probably because of the FBI’s investigation of Russia’s involvement in America’s election which started recently.

# References

1. How to use Pandas dataframe series for tweets - http://adilmoujahid.com/posts/2014/07/twitter-analytics/
2. Stop words removal from pandas dataframe series - http://stackoverflow.com/questions/33245567/stopword-removal-with-nltk-and-pandas
3. Stemming words in python pandas dataframe - http://stackoverflow.com/questions/37443138/python-stemming-with-pandas-dataframe
4. Remove Punctuation in python dataframe series - http://stackoverflow.com/questions/39782418/remove-punctuations-in-pandas