**Group12: Niranjan Naik ;Dhaval Metre; Anurag Patil**

1. **Collecting and Storing Tweets**

Collected 10k tweets from twitter filtering out tweets related to “ trump”. We used (collectTenKTweets.py) file to achieve it. In this file we used the twitter API called “tweepy” to collect the tweets and used MongoDB to store them. We used (extractDataByLocation.py) to extract the “text” part of the tweet to respective text files based on locations as well as 10k tweets. References used: refer 1

1. **For Sentiment Analysis**

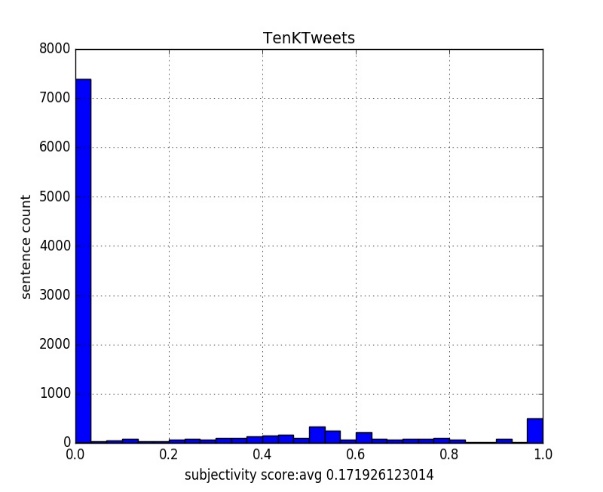
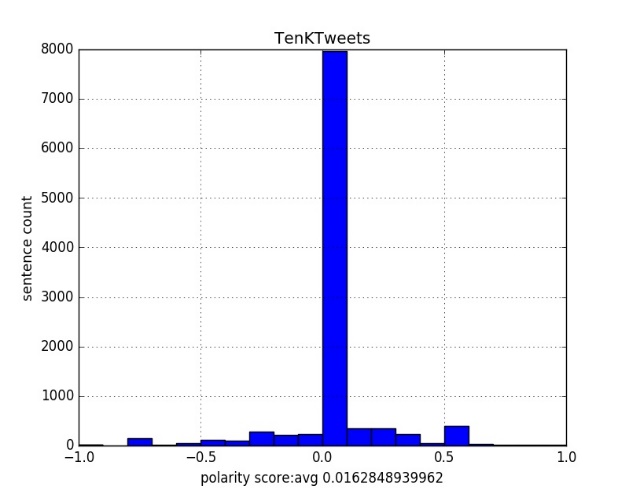
In this section, we used Text Blob to get polarity and subjectivity score for the tweets.

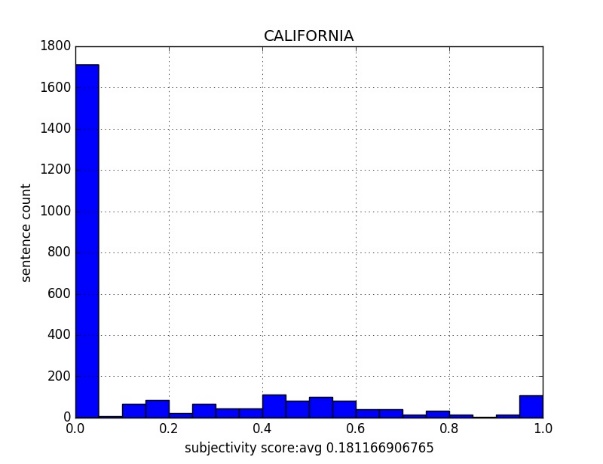
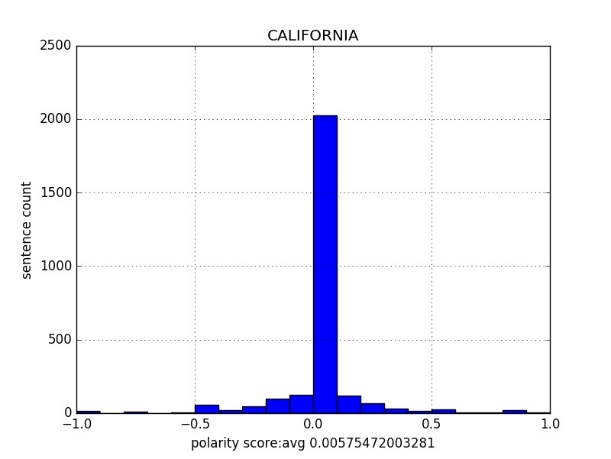
And used matplotlib.pyplot to plot the histogram. The results of the sentiment analysis for 10K tweets as well as location wise tweets are as follows:

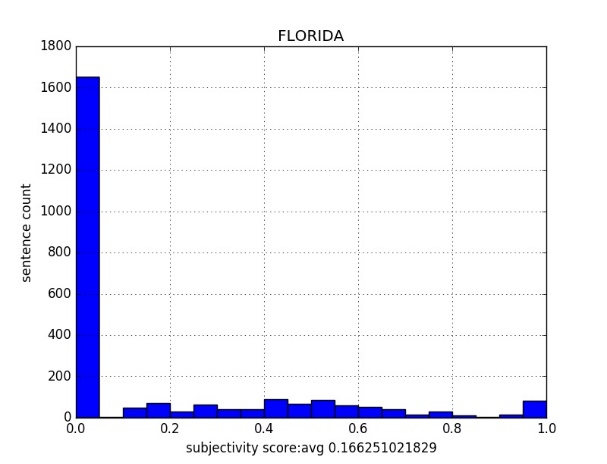
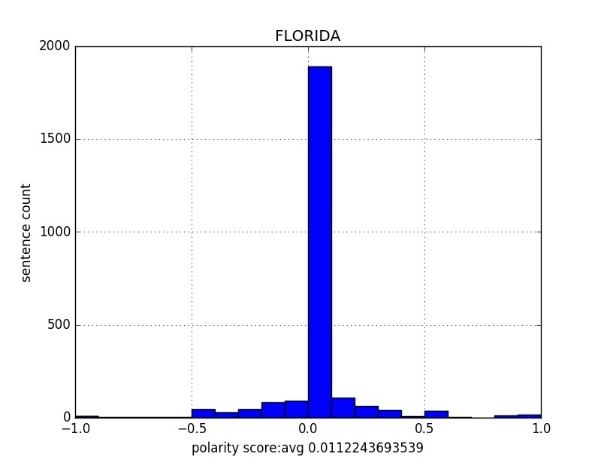
References used: refer 2

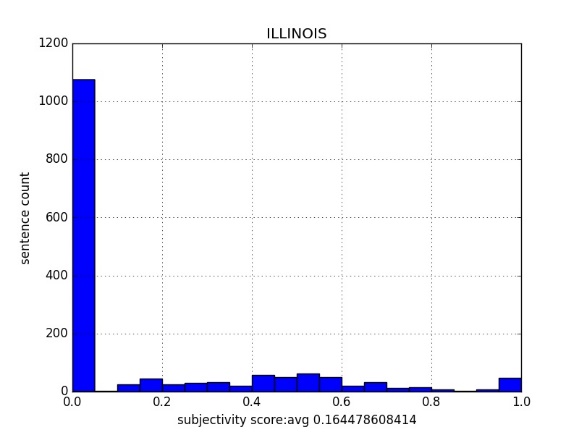
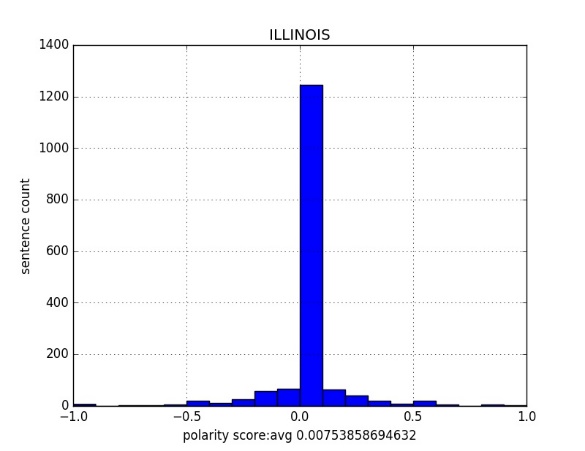
|  |  |  |
| --- | --- | --- |
| **Location** | **Average Polarity** | **Average Subjectivity** |
| New York(NY) | 0.0151670791815 | 0.170331685919 |
| Texas(TX) | 0.0048271161148 | 0.165944307105 |
| California(CA) | 0.00575472003281 | 0.181166906765 |
| Florida(FL) | 0.0112243693539 | 0.166251021829 |
| Illinois(IL) | 0.00753858694632 | 0.164478608414 |
| Oregon(OR) | 0.00683843955449 | 0.178420103647 |
| 10K Tweets | 0.0162848939962 | 0.171926123014 |

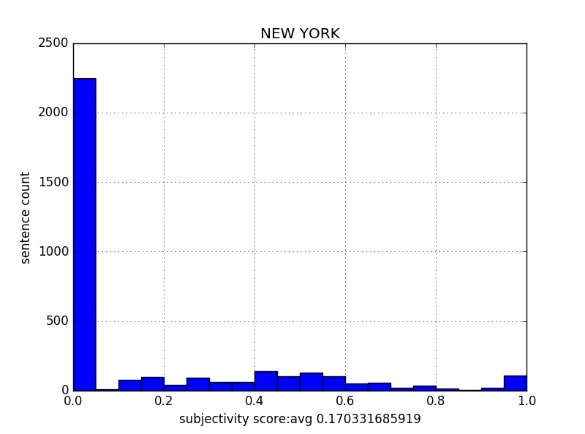
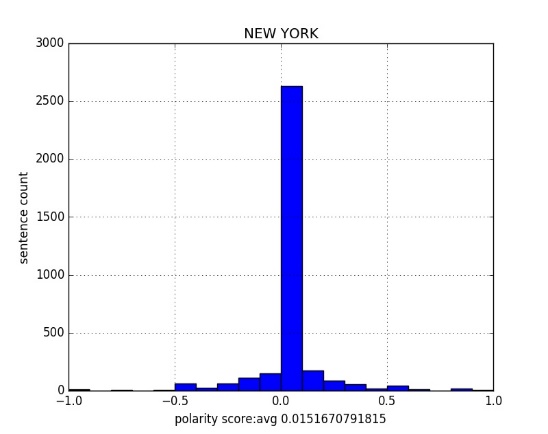
Refer to respective Jpeg Files for the histogram images.

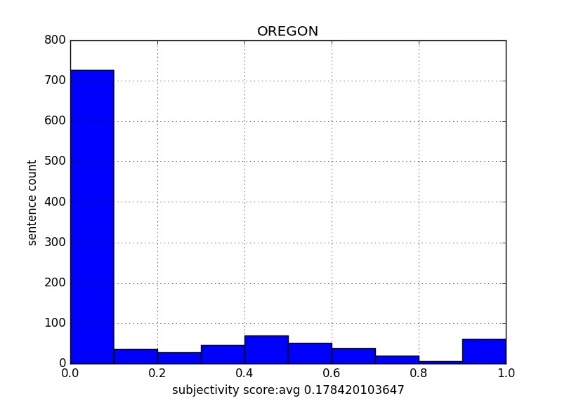
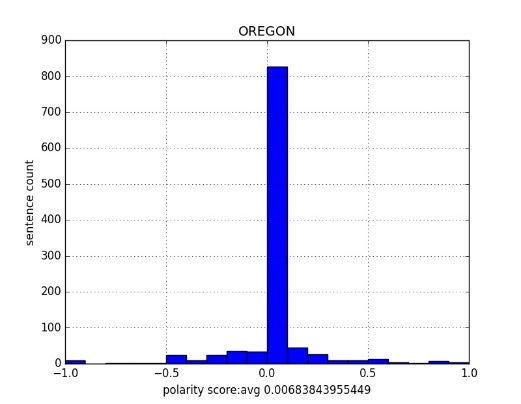


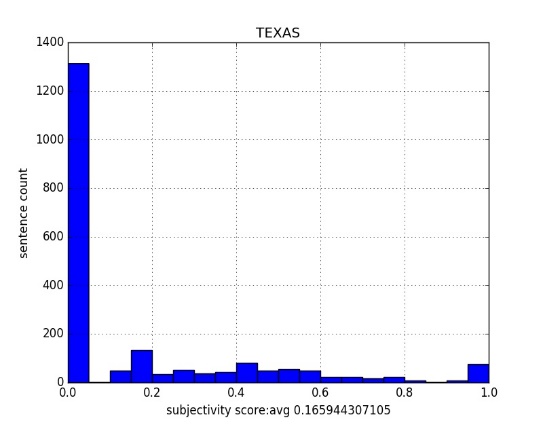
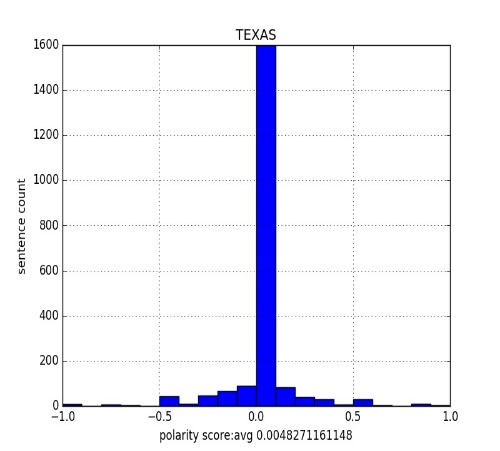








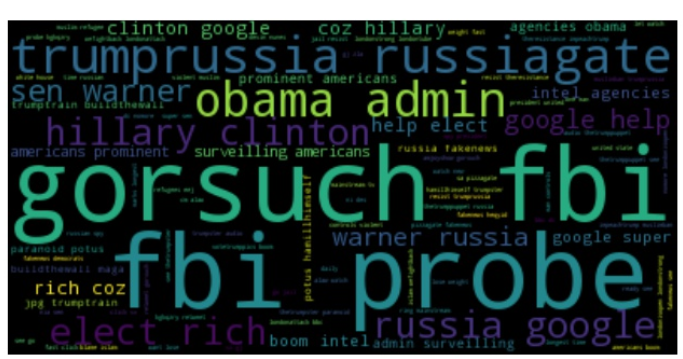


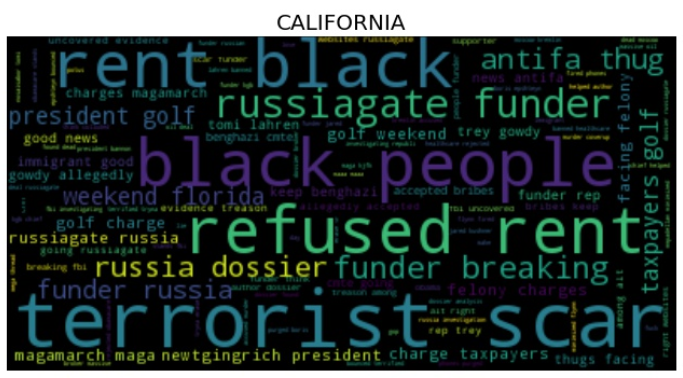


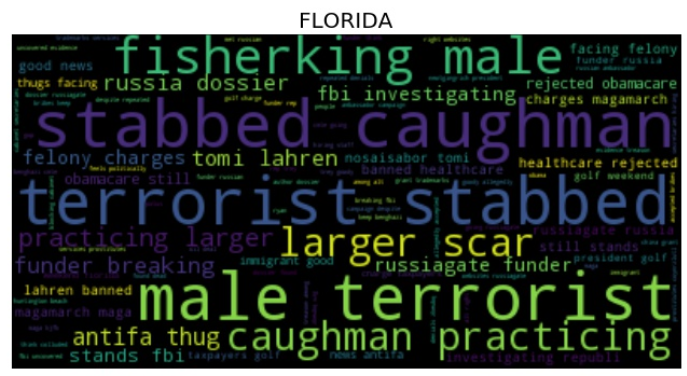
1. **Word Cloud**

We got the input from the text files generated as output of **A** and removed the punctuations, stop words and other digits. Then provided the resulting string as input to word cloud generate() method. References used: refer 3

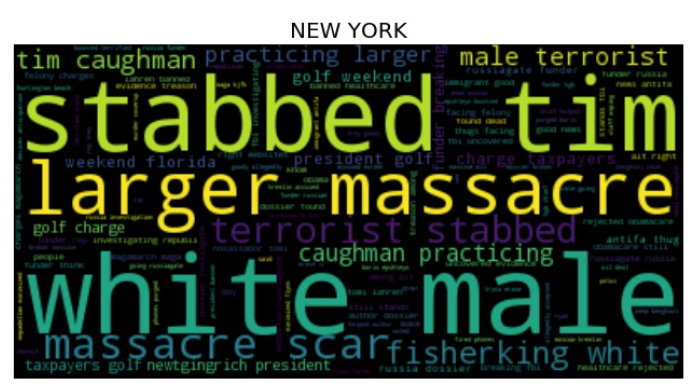
TenKTweets



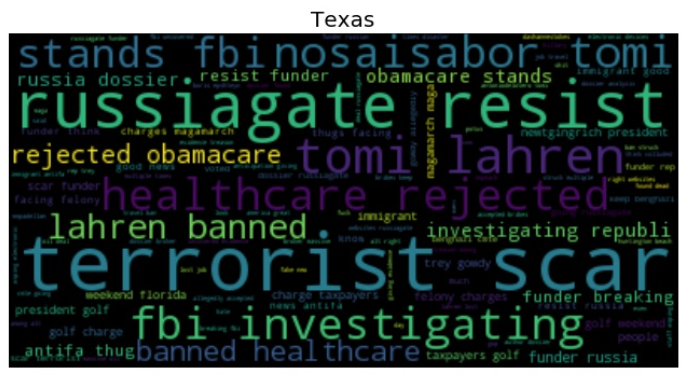












1. **NMF and LDA**

For NMF,

We used sklearn package. We got the input from the text files generated as output of **A** and removed the punctuations, stop words and other digits. Then provided the resulting list of strings containing words in each tweet as input to vectorizer to get document matrix. This matrix is then sorted based on the highest frequency of words. For the results please refer **Appendix 1.** References used: refer 4

For LDA,

We got the input from the text files generated as output of **A.** and removed the punctuations, stop words and other digits. Then created a corpora dictionary and stored the tweets for comparison. Then used TfidfModel and LdaModel models respectively to generate the LDA model where each word was compared over number of passes to decide the significance in the model..

We changed the combination of number of topics and no of instances to get different results and the results which gave us the better understanding and related results were selected. Also the model with low perplexity was considered over the other. For results of every location please refer to **Appendix 2**. References used: refer 5

1. **Location wise Tweet collection:**

Similarly, we collected tweets for different locations using different “locations” as filter, stored the tweets into respective collections in MongoDB as Documents and extracted “text” part with

extractDataByLocation.py. For Collecting tweets by location, we combined few states in to one group representing one location. Details are as follows:

|  |  |  |
| --- | --- | --- |
| **Location** | **States** | **Co-ordinates** |
| New York(NY) | New York, Massachusetts, New Jersey, Pennsylvania | -78.49,37.86, -61.08,47.87 |
| Texas(TX) | Texas, Oklahoma, Louisiana, Arkansas | -104.74,27.88, -93.54,37.68 |
| California(CA) | California, Nevada | -122.61,33.21, -112.76,41.57 |
| Florida(FL) | Florida, Georgia, North Carolina, South Carolina | -90.97,24.29, -73.56,36.1 |
| Illinois(IL) | Illinois, Detroit, Toronto, Cincinnati, Michigan | -91.41,38.1, -77.43,46.71 |
| Oregon(OR) | Oregon, Washington, Idaho | -125.86,43.45, -116.02,50.63 |

References used: refer 1

Files Used: tweetByLocation-FL.py, tweetByLocation-CA.py, tweetByLocation-IL.py,   
 tweetByLocation-NY.py, tweetByLocation-OR.py, tweetByLocation-TX.py.

**F.**

**a.** From the sentiment analysis we were able to show the polarity of tweets as positive or negative. We came to the conclusion that the location New York has most positive sentiments on the basis of the number of tweets received and the average polarity score.

b. The state of California has some unique tweets based on the latest insights of views by FOX News on scams and Trump relations.

c. The State of Oregon has most number of negative tweets based on the polarity score. Being a blue state there are more anti Trump tweets based on the Topic of Russia

d. State Florida and New York have a very similar sentiment based on the killing of Tim Caughman based on resent event that took place in United States which was based on racial matters.

e. As we went through collecting tweets we found that the data collected without location had tweets based on the decision that Trump has taken. Also, the state of Texas had many tweets related to Obamacare and pro trump tweets.

**Appendix-1**

NMFC run on **10K tweets**

Topic 0: google nia warner hillary sen

Topic 1: nomore londonstrong wefightback londonisopen londonattack

Topic 2: aloha londontube blame islam fact

Topic 3: americans intel agencies prominent surveilling

Topic 4: bio check electronics ridiculous traders

Topic 5: maga immigrant waynebogda tcot waters

Topic 6: playing araviapa putie truckers fakepresident

Topic 7: daily donald immigrant ivanka president

Topic 8: longest marks rhysam spy united

Topic 9: brexit putin article bro amp

Topic 10: fbi coordinated associates suggests germanrlopez

Topic 11: unmasking woodward scheme bob facing

Topic 12: potus gop thetrumppuppet political whitehouse

NMFC Run On **California** Location

Topic 0: resist repadamschiff impeach theresistance nunesmustresign

Topic 1: male larger stabbed practicing caughman

Topic 2: russia investigation funder finished regime

Topic 3: investigating fbi words day presidents

Topic 4: stands obamacare repeal fail aca

Topic 5: rejected healthcare pulled xavierdleau wfvomsggl

Topic 6: broker massive oil deal dossier

Topic 7: fired flynn subpoenas flipped prepping

Topic 8: bounced epshteyn boris moderated panel

Topic 9: purged phones let drain swamp

Topic 10: benghazi accepted cmte allegedly bribes

Topic 11: guilty people innocent teapainusa things

Topic 12: finally grow foxnews beginning scam

NMFC Run On **Florida** Location

Topic 0: russiagate funder repadamschiff act amjoy

Topic 1: investigating fbi words day collusion

Topic 2: stands obamacare repeal fail failed

Topic 3: rejected healthcare proposal denied like

Topic 4: practicing fisherking massacre larger caughman

Topic 5: russia bbc repadamschiff backs claims

Topic 6: nosaisabor theblaze banned tomi lahren

Topic 7: uncovered websites alt treason right

Topic 8: oil broker massive deal dossier

Topic 9: funder dossier spy analysis mega

Topic 10: antifa thugs immigrant maga felony

Topic 11: fired flynn jared smuggled amb

Topic 12: resist repadamschiff russia gorkamustgo girl

NMFC Run On **Illinois** Location

Topic 0: russiagate funder amjoy gorka joyannreid

Topic 1: investigating fbi words investigation day

Topic 2: stands obamacare repeal fail healthcarereform

Topic 3: rejected healthcare failure job shut

Topic 4: tim caughman larger stabbed fisherking

Topic 5: dossier spy funder linked car

Topic 6: jared amb smuggled meeting towers

Topic 7: theblaze banned nosaisabor lahren tomi

Topic 8: resist repadamschiff nunesmustresign dashannestokes gop

Topic 9: golf taxpayers president charge newtgingrich

Topic 10: yea hell feel couple seen

Topic 11: oil broker deal massive dossier

Topic 12: bounced epshteyn boris believes head

NMFC Run On **New** **York** Location

Topic 0: theresistance resist russia russiagate bbc

Topic 1: caughman stabbed massacre practicing male

Topic 2: investigating fbi words day appears

Topic 3: stands obamacare repeal fail worth

Topic 4: rejected healthcare way sa atdavidhoffman

Topic 5: oil broker deal massive dossier

Topic 6: fired flynn smuggled amb towers

Topic 7: bounced epshteyn boris judge jeanine

Topic 8: purged phones asset liability experience

Topic 9: golf newtgingrich weekend charge florida

Topic 10: russiagate funder miss joyannreid putin

Topic 11: moscow kgb author helped coverup

Topic 12: websites uncovered treason alt evidence

NMFC Run On **Oregon** Location

Topic 0: russiagate putin brilliant strategy funder

Topic 1: claims backs source bbc dossier

Topic 2: websites uncovered alt treason right

Topic 3: male fisherking larger practicing massacre

Topic 4: oil broker massive deal dossier

Topic 5: investigating fbi words day presidents

Topic 6: stands obamacare fail repeal man

Topic 7: rejected healthcare fail liberals gorkamustgo

Topic 8: finally scam beginning artist wake

Topic 9: fired flynn flipped subpoenas wipi

Topic 10: epshteyn bounced boris judgejeanine sundaymorning

Topic 11: purged phones decodes grantstern article

Topic 12: golf weekend charge president florida

NMFC Run On **Texas** Location

Topic 0: massacre stabbed larger caughman fisherking

Topic 1: theresistance russia bbc muslimban miss

Topic 2: investigating fbi investigation presidents like

Topic 3: stands obamacare repeal fail wfvomsggl

Topic 4: rejected healthcare pulled wfvomsggl xavierdleau

Topic 5: scar got blame regime like

Topic 6: theblaze nosaisabor banned lahren tomi

Topic 7: massive broker oil deal dossier

Topic 8: russiagate funder amjoy joyannreid scandal

Topic 9: allegedly trey gowdy bribes benghazi

Topic 10: guilty people come therickydavila leaders

Topic 11: fired flynn smuggled amb meeting

Topic 12: bounced epshteyn boris judge sundaymorning

**Appendix 2**

**TenKTweets**

Number of topics=3

Number of passes=50

[(0, u'0.192\*"trump" + 0.015\*"immigrant" + 0.011\*"jalloyd" + 0.011\*"wa" + 0.009\*"trumprt" + 0.008\*"impeachtrump" + 0.007\*"hillari" + 0.007\*"nia" + 0.006\*"dashannestokes" + 0.006\*"donaldtrump"'), (1, u'0.032\*"t" + 0.029\*"co" + 0.029\*"https" + 0.025\*"maga" + 0.023\*"trumprussia" + 0.018\*"resist" + 0.015\*"new" + 0.014\*"flynn" + 0.012\*"thi" + 0.009\*"hi"'), (2, u'0.019\*"russia" + 0.015\*"amp" + 0.014\*"latest" + 0.012\*"thank" + 0.012\*"googl" + 0.009\*"theresist" + 0.008\*"use" + 0.008\*"russiag" + 0.007\*"get" + 0.007\*"help"')]

INFO : -8.618 per-word bound, 393.0 perplexity estimate based on a held-out corpus of 225160 documents with 287858 words

**California**

topics= 3

passes= 100

[(0, '0.058\*"funder" + 0.046\*"russiag" + 0.028\*"donald" + 0.016\*"know" + 0.012\*"russian" + 0.011\*"golf" + 0.011\*"black" + 0.010\*"rent" + 0.010\*"keep" + 0.010\*"think"'), (1, '0.068\*"trumprussia" + 0.022\*"t" + 0.020\*"thi" + 0.020\*"russia" + 0.020\*"amp" + 0.017\*"got" + 0.015\*"white" + 0.014\*"dossier" + 0.014\*"mmpadellan" + 0.012\*"peopl"'), (2, '0.034\*"resist" + 0.017\*"presid" + 0.015\*"terrorist" + 0.012\*"massacre" + 0.012\*"tim" + 0.012\*"male" + 0.012\*"larger" + 0.012\*"caughman" + 0.012\*"practic" + 0.012\*"scar…"')]

INFO : -7.551 per-word bound, 187.5 perplexity estimate based on a held-out corpus of 33539 documents with 35389 words

**Florida**

Number of topics=5

Number of passes=100

[(0, u'0.095\*"trump" + 0.037\*"resist" + 0.030\*"thi" + 0.024\*"hi" + 0.023\*"ha" + 0.018\*"maga" + 0.016\*"got" + 0.015\*"immigrant" + 0.014\*"golf" + 0.013\*"white"'), (1, u'0.071\*"trumprussia" + 0.064\*"funder" + 0.017\*"get" + 0.014\*"support" + 0.011\*"one" + 0.011\*"peopl" + 0.011\*"don" + 0.010\*"tim" + 0.010\*"male" + 0.010\*"stab"'), (2, u'0.034\*"russia" + 0.024\*"presid" + 0.018\*"mmpadellan" + 0.014\*"he" + 0.014\*"russian" + 0.013\*"rt" + 0.013\*"becaus" + 0.013\*"realdonaldtrump" + 0.011\*"still" + 0.009\*"healthcar"'), (3, u'0.058\*"t" + 0.046\*"https" + 0.044\*"russiag" + 0.035\*"co" + 0.021\*"\u2026rt" + 0.019\*"dossier" + 0.017\*"think" + 0.014\*"like" + 0.013\*"investig" + 0.012\*"terrorist"'), (4, u'0.026\*"wa" + 0.025\*"amp" + 0.024\*"go" + 0.018\*"fbi" + 0.015\*"u" + 0.014\*"it" + 0.013\*"want" + 0.012\*"theresist" + 0.011\*"make" + 0.010\*"time"')]

INFO : -7.532 per-word bound, 185.1 perplexity estimate based on a held-out corpus of 36449 documents with 44120 words

**Illinois**

Number of topics=3

Number of passes=500

[(0, u'0.060\*"trump" + 0.052\*"trumprussia" + 0.032\*"russiag" + 0.022\*"russia" + 0.011\*"dossier" + 0.011\*"investig" + 0.010\*"white" + 0.010\*"u" + 0.010\*"immigrant" + 0.010\*"go"'), (1, u'0.042\*"funder" + 0.036\*"t" + 0.027\*"https" + 0.019\*"co" + 0.016\*"golf" + 0.013\*"\u2026rt" + 0.012\*"amp" + 0.012\*"rt" + 0.012\*"think" + 0.012\*"maga"'), (2, u'0.025\*"resist" + 0.018\*"wa" + 0.017\*"thi" + 0.017\*"hi" + 0.013\*"presid" + 0.012\*"ha" + 0.009\*"news" + 0.009\*"terrorist" + 0.009\*"support" + 0.008\*"male"')]

INFO : -7.383 per-word bound, 166.9 perplexity estimate based on a held-out corpus of 23283 documents with 27729 words

**New York**

topics= 5

passes= 500

[(0, '0.093\*"trumprussia" + 0.048\*"resist" + 0.039\*"thi" + 0.024\*"presid" + 0.018\*"theresist" + 0.017\*"white" + 0.017\*"fbi" + 0.017\*"terrorist" + 0.015\*"peopl" + 0.014\*"keep"'), (1, '0.071\*"russiag" + 0.030\*"amp" + 0.022\*"got" + 0.020\*"like" + 0.020\*"get" + 0.015\*"investig" + 0.014\*"scar…" + 0.014\*"caughman" + 0.014\*"practic" + 0.013\*"rep"'), (2, '0.086\*"funder" + 0.027\*"mmpadellan" + 0.023\*"dossier" + 0.018\*"it" + 0.017\*"news" + 0.013\*"putin" + 0.012\*"make" + 0.012\*"still" + 0.011\*"charg" + 0.011\*"good"'), (3, '0.028\*"donald" + 0.026\*"russia" + 0.019\*"russian" + 0.017\*"he" + 0.015\*"one" + 0.013\*"don" + 0.012\*"say" + 0.011\*"would" + 0.011\*"all" + 0.011\*"y"'), (4, '0.037\*"t" + 0.024\*"golf" + 0.019\*"think" + 0.016\*"trump" + 0.012\*"weekend" + 0.012\*"gowdi" + 0.011\*"immigrant" + 0.011\*"fisherking" + 0.011\*"need" + 0.011\*"want"')]

INFO : -7.837 per-word bound, 228.6 perplexity estimate based on a held-out corpus of 46917 documents with 49532 words

**Oregon**

Number of topics=3

Number of passes=500

[(0, u'0.045\*"trumprussia" + 0.034\*"t" + 0.031\*"russiag" + 0.028\*"https" + 0.020\*"co" + 0.019\*"thi" + 0.014\*"ha" + 0.013\*"golf" + 0.013\*"wa" + 0.012\*"presid"'), (1, u'0.043\*"funder" + 0.026\*"resist" + 0.014\*"dossier" + 0.014\*"\u2026rt" + 0.011\*"go" + 0.011\*"mmpadellan" + 0.010\*"maga" + 0.010\*"realdonaldtrump" + 0.009\*"peopl" + 0.008\*"news"'), (2, u'0.064\*"trump" + 0.024\*"russia" + 0.017\*"hi" + 0.014\*"amp" + 0.013\*"rt" + 0.008\*"think" + 0.008\*"america" + 0.006\*"time" + 0.006\*"support" + 0.006\*"immigrant"')]

INFO : -7.413 per-word bound, 170.5 perplexity estimate based on a held-out corpus of 15872 documents with 18998 words

**Texas**

Number of topics=5

Number of passes=500

[(0, u'0.080\*"trumprussia" + 0.032\*"stab" + 0.032\*"practic" + 0.020\*"fbi" + 0.018\*"golf" + 0.017\*"mmpadellan" + 0.017\*"investig" + 0.013\*"still" + 0.013\*"u" + 0.013\*"it"'), (1, u'0.067\*"funder" + 0.053\*"russiag" + 0.031\*"male" + 0.027\*"amp" + 0.026\*"fisherking" + 0.023\*"presid" + 0.022\*"t" + 0.020\*"dossier" + 0.015\*"theresist" + 0.012\*"get"'), (2, u'0.042\*"t" + 0.041\*"https" + 0.039\*"resist" + 0.031\*"caughman" + 0.030\*"co" + 0.022\*"\u2026rt" + 0.020\*"go" + 0.018\*"ha" + 0.016\*"support" + 0.014\*"maga"'), (3, u'0.037\*"white" + 0.034\*"terrorist" + 0.033\*"russia" + 0.032\*"massacre" + 0.032\*"larger" + 0.032\*"scar\u2026" + 0.032\*"tim" + 0.013\*"peopl" + 0.011\*"time" + 0.011\*"obamacar"'), (4, u'0.089\*"trump" + 0.051\*"wa" + 0.039\*"got" + 0.030\*"thi" + 0.024\*"all" + 0.024\*"y" + 0.022\*"hi" + 0.017\*"rt" + 0.016\*"think" + 0.014\*"russian"')]

INFO : -7.348 per-word bound, 163.0 perplexity estimate based on a held-out corpus of 27550 documents with 32617 words

**References:**

1.

<http://stats.seandolinar.crom/collecting-twitter-data-storing-tweets-in-mongodb/>

<http://boundingbox.klokantech.com/>

<http://altons.github.io/python/2013/01/21/gentle-introduction-to-mongodb-using-pymongo/>

2.3.4.5

<https://www.dropbox.com/sh/cpkokzf1ko0p53t/AADzflCodDwlEMdliCvFRlL4a/JG_Ch20_Text_Mining/0_notebooks?dl=0>