Ramabhadra Raju Tirumalaraju

March 30, 2016

Data science

Text analytics mini project

**INSY-5378-001- Data Science – Text Analytics Mini Project**

# **Techniques used for the project**

## **Reading the tweets**

We used Search API provided by twitter using Twython. To avoid the rate limit error we used “app auth” to connect to twitter rest API. We were able to get 450 requests x 100 tweets (45000 tweets) in a single 15 minute window.

## **Getting tweets based on location**

For collecting tweets based on location we have used the geolocalization feature in the search api

twython\_object.cursor(twython\_object.search, q = query, count=100, Geolocalization = “37.781157,-122.398720,1mi”)

It takes a location and within the specified radius the tweets will be extracted. For our purpose we have taken the midpoint (center) location of all 50 states, then based on the size (area) of the state we determined the radius to maximize the coverage area. We have collected a small sample of 200 tweets for each candidate in every state. So at a go we are collecting 4 candidates \* 50 states \* 200 tweets = 40000 tweets.

## **Processing the tweet text**

The following preprocessing steps were undertaken:

Convert to lowercase

Remove the URL’s

Removed usernames as they are not useful in sentiment analysis

Removed additional white spaces

Removed the stop words.

## **Performance**

We weren’t able to do a detailed performance testing but the entire code takes around 5-10 minutes for extracting the tweets and calculating the sentiments. We have used a list vs set comparison while calculating the sentiment score which is faster than list vs list. Also the calculation for ANEW module is fast as we are storing all the 10000 processed tweets at calculating the score in a single go.

## **Visualization using Tableau**

The results were visualized using tableau. We used the tableau dashboard to associate the various worksheets we made the Tableau workbook. This was published using Tableau Public server. We have mentioned the Tableau links for the various results that we got.

1. Use the Twitter API (Twython, Twitter, Tweepy, or any other Python-Twitter module) to download about 10,000 tweets for each of the presidential candidates.

Please find the attachments in the mail.

1. Calculate sentiments for each of the candidates using: 1) AFFIN11.txt (see Blackboard); 2) the Hu & Liu lexicon of positive and negative words; 3) ANEW module (check the following site to figure out how to use the module: <http://www.csc.ncsu.edu/faculty/healey/maa-16/text/>). Who has the most positive comments? The most negative?

**Solution:**

Sentiments are calculated based on each of Affin,Hu & Liu lexicon and Anew module. Following are the results

1. **AFFIN11.txt:**

|  |  |  |  |
| --- | --- | --- | --- |
| **candidate** | **Positive affinity** | **negative affinity** | **overall affinity** |
| **trump** | 14228 | -11150 | 3078 |
| **hillary** | 12725 | -12405 | 320 |
| **sanders** | 14810 | -6800 | 8010 |
| **cruz** | 5848 | -12099 | -6251 |

Based on the results obtained Sanders got the most positive sentiment and Ted cruz got the most negative sentiment.

The results are visualized for comparison using tableau, this can be seen from the following link [View affin results](https://public.tableau.com/profile/publish/AffinityScores/Sheet1#!/publish-confirm)

1. **Hu & Liu lexicon:**

|  |  |  |  |
| --- | --- | --- | --- |
| **candidate** | **positive lexicon** | **negative lexicon** | **overall lexicon** |
| **trump** | 15299 | 5745 | 9554 |
| **hillary** | 7353 | 8486 | -1133 |
| **sanders** | 7409 | 4700 | 2709 |
| **cruz** | 8214 | 7479 | 735 |

Based on Hu & lexicon trump has the most positive sentiment and Hilary has the most negative sentiment.

The results can be viewed in Tableau [View Lexicon results](https://public.tableau.com/profile/publish/LexiconScores_0/LexiconScores#!/publish-confirm)

1. **Anew Module:**

|  |  |  |
| --- | --- | --- |
| candidate | a new arousal | a new valence |
| trump | 245063.42 | 228541.98 |
| hillary | 179321.41 | 187624.92 |
| sanders | 205381.05 | 210645.04 |
| cruz | 206177.92 | 200811.02 |

Based on the Anew results trump has the highest arousel and valence and hilary has the lowest arousal and valence.

The results are compared using tableau [View Anew results](https://public.tableau.com/profile/publish/AnewScores_0/Anewscores#!/publish-confirm)

1. Repeat (b) by any states (e.g., TX, AZ, CA, etc.). For each candidate, display the top five states by: 1) positive sentiments and 2) negative sentiments

Below are the results for the top five states with positive and negative sentiments

The results can be viewed in tableau [View State wise Results](https://public.tableau.com/profile/publish/Statewise_3/Sheet4#!/publish-confirm)

**AFFIN Results:**

**Trump:**

Top five Positive states:Wyoming, Wisconsin, West Virginia, Washington Virginia

Top five Negative States: Indiana ,Illinois ,Idaho ,Hawaii and Iowa

**Cruz:**

Top five Positive states: Alabama ,Alaska ,Arizona , Arkansas,California

Top five Negative States: Wyoming ,Wisconsinv ,West Virginia , Washington Virginia

**Hillary:**

Top five Positive states: Wyoming ,Wisconsinv ,West Virginia ,Washington

Virginia

Top five Negative States: Alabama,Alaska ,Arizona ,Arkansas California.

**Sander:**

Top five Positive states: Alabama ,Alaska ,Arizona ,Arkansas ,California

Top five Negative States: Vermont,Utah ,Texas, Virginia and Tennessee

**Lexicon Results:**

**Trump:**

Top five Positive states: North Carolina ,Ohio ,South Dakota ,Oklahoma and

South Carolina

Top five Negative States: Iowa, Kentucky, Kansas , Indiana and Mainev.

**Cruz:**

Top five Positive states: Illinois, Delaware ,Georgia ,Idaho and Hawaii

Top five Negative States: Washington ,West Virginia ,Wisconsinv ,Virginia and Wyoming

**Hillary:**

Top five Positive states: Alabama Alaska Arkansas Arizona

Georgia

Top five Negative States: Colorado, Minnesota, Iowa, Massachusetts

New Hampshire

**Sander:**

Top five Positive states:Nevada , Pennsylvania , Missouri, North Carolina

Oklahoma

Top five Negative States: Washington ,West Virginia , Wisconsinv , Virginia

Wyoming

**Anew :**

**Trump:**

Top five Arousal states: Minnesota Mississippi Michigan Massachusetts

Missouri

Top five Valence States: Minnesota Mississippi Michigan Massachusetts

Missouri

**Cruz:**

Top five Arousal states: Alabama, Alaska, California , Arkansas, Arizona

Top five Valence States: Arizona , Arkansas, California, Alaska

Alabama

**Hillary:**

Top five Positive states: Georgia , Florida , Alabama ,Alaska Arkansas

Top five Negative States: Georgia , Florida, Alabama , Alaska Arkansas

**Sander:**

Top five Positive states: Alabama, Illinois, Delaware , Alaska

Georgia

Top five Negative States: Alabama, Illinois , Delaware, Alaska , Georgia

1. Do a topic analysis on the entire set of tweets. You may choose 25 topics to extract from your topic modeling software – try NMF and LDA under Gensim.

I would encourage you to try out Tableau (I know I am at fault for not uploading the license – will do it tonight), seaborn/matplotlib in Python to visualize your results.

**Topic Modeling:**

Following are the top 25 topics from all the tweets:

Topic 0: pie sky argues economist fact economy make great feelthebern sanders tytlive feelthebe feelin fac wiprimary difference mainstreammedia hear bernie says

Topic 1: york new place cuddle used family str deserves learned hrc love wall bernie amp sanders debate hashing limb brockspin spew

Topic 2: amp subvert functions deceive sent trojan horse destroy soros saboteur believe gop hillary trump difference says candidate nominated necessity candidates

Topic 3: unfit sleazy elected wakeupamerica man corrupt wi president watch amp cruz vote sad prowar worldwide cnnmsnbc amazing says cias Libya

Topic 4: wiunion worth badgers fighting berniesanders hey think doesnt wisconsin vote hillary win words classified exactly kuddos gotten dangerous settle surging

Topic 5: fair includes signed right pledge story said cruz trump terrifying rightbernies biggest supporters notmeus best years cow lost grazing libyawhat

Topic 6: hillary door billclinton open person country quoteoftheday hillaryclinton imwithher spew brockspin targets ch id prefer hrc election clintoncash lewandowskigoing

Topic 7: continue threatened action hashtag legal let campaign cruzsexscandal cruz news repurposed unfairly disproven antihillary consistently lies credibility daddycruz cruzunderpressure heslut

Topic 8: voters mailer newlyregistered confusing nyc primary vote sanders trump official wld writein thank nominee president little authorities hurt confusion fixing

Topic 9: tonegate campaigns dismissal disaster condescending concern dems hillarys bernie fbi speech prowar cnnmsnbc worldwide amazing sad justice comey illegalemail director

Topic 10: wants women abortions punish wifecomparison cruzthe donald thing coat hangers trump jail abortion deathpenaltylotto illegal amp punished prosecute wonders punishment

Topic 11: tr refugees virtues touting busted muslim wiprimary camera cruz trump wisconsin friends ryan dumptheestablishment follow paul running dishonest hes dear

Topic 12: like teamdl michelleobama bad acted things president said cruz trump vote terrifying comes kills cringe prochoice esp issue republicans general

Topic 13: cruzsexscandal confirmed mistresses truth cruzcrew cruz daddycruz heslut neverlyingcheatingtedcruz cruzunderpressure ted votetrump hoted screwed rafael bc lied liars look national

Topic 14: ws ranting disgusted taking money send im hillary sanders based lens world vexed moldy sees principles marxist view economic revolution

Topic 15: lgbtq gay gays executed saudiarabia thought cruzcrew cruz cr despicable sex marriage matter think doesnt throw stands whos overboard youd

Topic 16: clinton sanders contest poll nbc close debate wi nbcmadison hashing stillsanders york new ahead speeches dings wallst appleton enoughisenough forthepeople

Topic 17: camps vortexes antimuslim inside takes hate amp cruz trump says coat hangers follow friends ryan dumptheestablishment difference paul hes running

Topic 18: people support days gone kkk disavowed called african predators hillary clinton imwithher know brockspin spew hillaryclinton running prowar cnnmsnbc amazing

Topic 19: amnesty plans watch lying fox times bombshell wiprim caught lies exposed shocking lyinted maga cruz camera amp trump repurposed antihillary

Topic 20: zone wall trump main aims donations shell street tax want says itsingingin david bear build mind hasselhoff years targets trumpcaachinggggg

Topic 21: bernie dear voted superdelegate kimmetcalfe alaska sorry endorses htt hillary best notmeus years biggest rightbernies supporters wisconsin dishonest story heart

Topic 22: camp hillarys freaking viral video democrat proud goes read speech parent student ucla teacher scenes interested paula jones crimes corey

Topic 23: punishment hes manager form unable comment respond trumps campaign jail let friends dumptheestablishment ryan ad nekkid follow time abortion paul  
Topic 24: trump cruz kasich wont ted votetrump conservative lets media day endorsement mention start scarlet manufacturing vote gop say win hear

**Topic Analysys:**

Following is the topic analysis for each of the topic:

topic 0: support for Bernie

topic 1: Bernie sander attacked by david brock

topic 2: hillary versus trump

topic 3: corruption in america

topic 4: Wisconsin union

topic 5: negative commentc on cruz regarding violating pledge

topic 6: support for hillary

topic 7: cruz scandal

topic 8: misleading newyork election mailer

topic 9: Tone gate disaster Hillary clinton

topic 10: Women abortions

topic 11: paul ryan wisconsin primary

topic 12: team dl hughley

topic 13: cruz under pressure

topic 14: economic revolution in US

topic 15: same sex marriage

topic 16: Rally at Madison, wisconsin

topic 17: antimuslim

topic 18: prowar candidates

topic 19: antihillary and pro cruz

topic 20: wall street primary

topic 21: Kim Metcalfe in Alaska primary

topic 22:Paula jones

topic 23: punishment for abortion

topic 24:GOP wins