2016 US Presidential election prediction from Twitter sentiments using MapReduce framework and ArcMap

-Nandan Nayak

Introduction

We are fortunate enough to live in a world where we get unbounded amount of information from the social media which adds vibrantly to the perception of our world. This project aims at capturing one figment of the social media to perform analysis to solve some of the problems of the world.

The aim of this project is to predict the ultimate candidate from the US presidential candidates of 2016. The prediction is based on the positive comments that each candidate receives on Twitter. The comments are classified as positive and negative based on the sentiment score that each tweet receives which is calculated using the MapReduce technology. The tweets are then located on the Map in ArcMap using the location of the tweets.

Study Area

The study area would be the entire contiguous US. This area was chosen as it would give a more realistic and holistic view about the prediction.

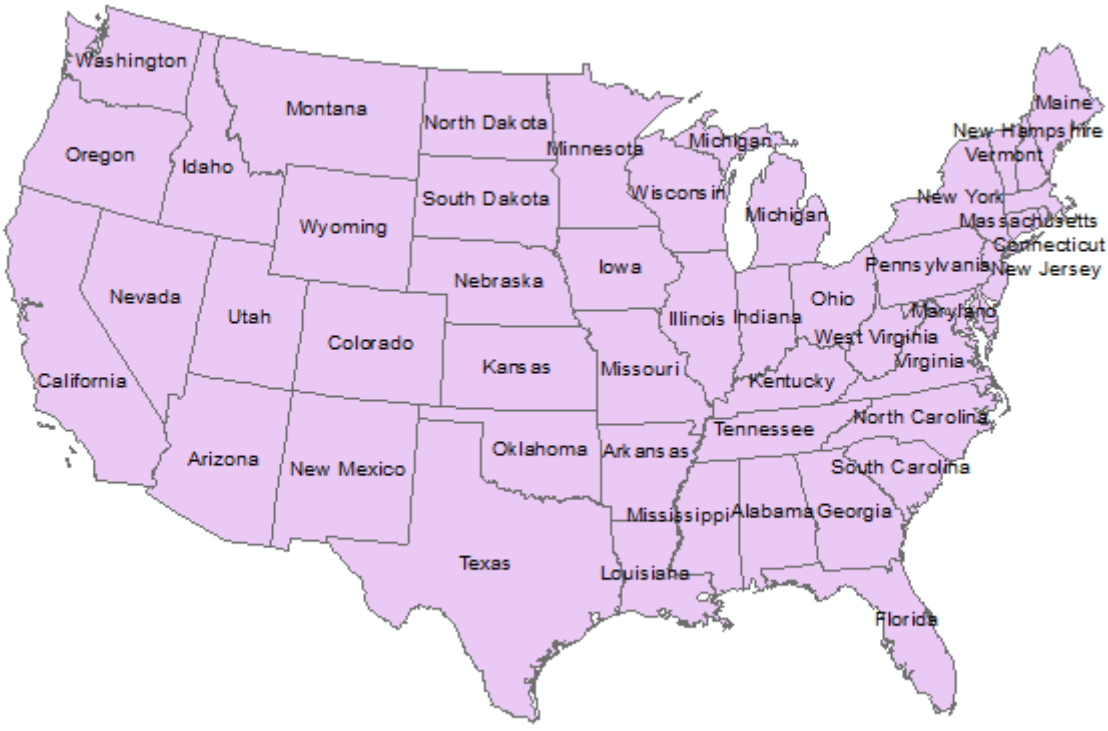
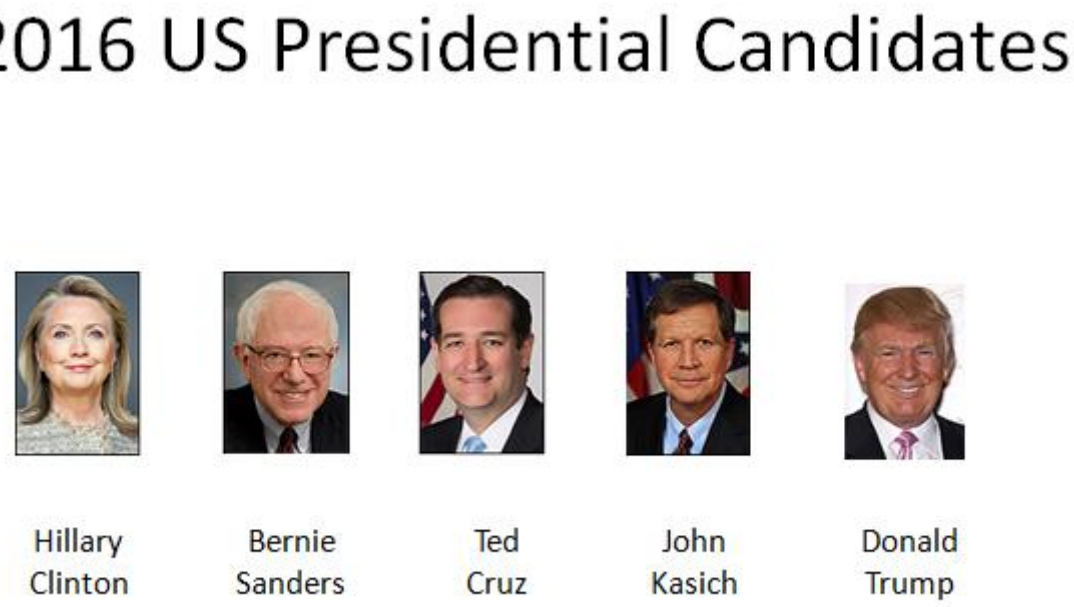


Fig1. Contiguous US Map

Presidential Candidates Considered [1]



Application Development

The development of this application involved the techniques like Map Reduce[2] framework, Pattern Matching which are extensively used in Data Mining.

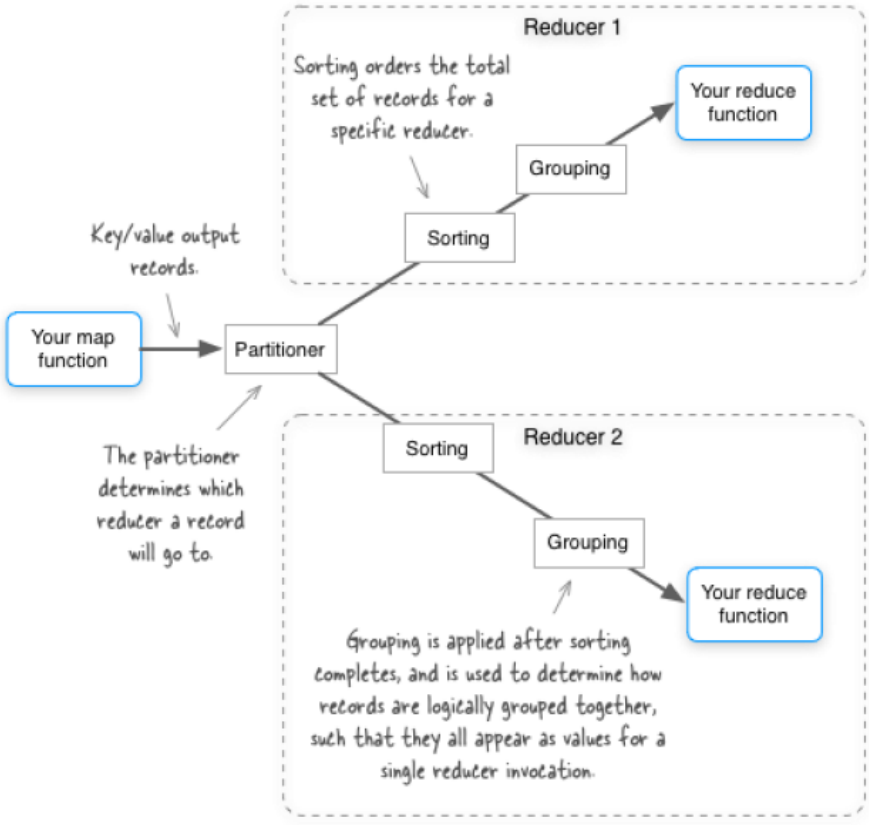


Fig2.MapReduce Framework

Fig2. Captures the idea of MapReduce Framework. In a Distributed File System, the data is divided into chunks as distributed across multiple nodes. Each node then works on only a chunk of main data. The node then performs the Mapper / Reducer functionality. This concept is widely used in Big Data technologies. The data is first fed to the Mapper task which produces output in the form of key value pairs. The partitioner then determines which reducer a record will go to. The key-value pairs are sorted and grouped according to the keys and fed to the reducer task. The reducer then aggregates the values in each key and outputs the result.

Collection of Data and its processing

The first task in this project was to collect the twitter feeds. The library used for this process was Oauth2[3]. The library requires a user's API key, API secret, Token key, Token Secret. This can be obtained by logging into **https://dev.twitter.com/apps** by using the twitter credentials. By executing the library, the feeds can be collected. The tweets are in JSON format. A sample tweet would look like this as shown below.

{"created_at":"Tue Apr 26 04:44:21 +0000 2016","id":724821394589138944,"id_str":"724821394589138944","text":"@tberg15 I just can't be a 12 like you though","source":"\u003ca href=\"http:\/\/twitter.com\/download\/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c\/a\u003e","truncated":false,"in_reply_to_status_id":724819425996738560,"in_reply_to_status_id_str":"724819425996738560","in_reply_to_user_id":406504318,"in_reply_to_user_id_str":"406504318","in_reply_to_screen_name":"tberg15","user":{"id":328631651,"id_str":"328631651","name":"Turtles","screen_name":"nickychris11","location":null,"url":null,"description":null,"protected":false,"verified":false,"followers_count":529,"friends_count":512,"listed_count":4,"favourites_count":11278,"statuses_count":11992,"created_at":"Sun Jul 03 18:53:11 +0000 2011","utc_offset":-14400,"time_zone":"Eastern Time (US & La canada)","geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_color":"C0DEED","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_background_images\/839253051\/a95f92f8d670584869c1dfe49947c8a0.jpeg","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile_background_images\/839253051\/a95f92f8d670584869c1dfe49947c8a0.jpeg","profile_background_tile":true,"profile_link_color":"0084B4","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/684159590917902337\/exORCaft_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/684159590917902337\/exORCaft_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/328631651\/1460481688","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null}}

The collected tweets are then processed to remove hyperlinks, special characters like smileys, hashtags and punctuations using pattern matching technique. The pattern matching was achieved using Regular Expression module in Python. The processed tweets are then fed to the Mapper. The mapper first checks if the tweet is about any presidential candidate contesting for 2016 election. If so, it then compares each word in the tweet to a dictionary which contains words and their sentiment scores and generates key-value pairs of the format. The key-value pairs are then fed to the Reducer task which calculates the sentiment score of the tweet, by summing up the scores of each word contained in it. The final sentiment score of a tweet is in the format – [Tweet_ID, [sentiment_score, presidential candidate's name]]. The example string would look like this as shown below.

```
[869, [-5.0, "ted"]]
[1473, [6.0, "cruz"]]
[1555, [4.0, "trump"]]
[1639, [3.0, "trump"]]
[1657, [6.0, "ted"]]
[1925, [-4.0, "donald"]]
[2230, [1.0, "ted"]]
[2258, [-3.0, "trump"]]
[2414, [-1.0, "hillary"]]
```

Fig3. Sample MapReduce output

The sentiment score is used to determine if the tweet regarding a particular Presidential candidate is positive or negative. Positive tweets are considered to be potential votes for the candidate and the candidate which highest percentage of votes is considered to be the winner of the election.

Another script is written in Python is collect the location information of the tweet and plot the same on ArcMap. The location for each presidential candidate is consolidated in a file.

In ArcMap, two Data Frames were used. One is for US States and the other is US cities. Using US States as the base map, only the cities from where the tweets originated were plotted using select by attribute method.

ER diagram

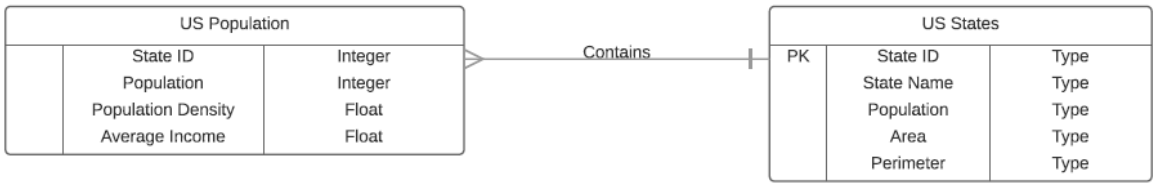


Fig4.ER Diagram

The ER Diagram states that the US population is contained in US States. The US States table has the State ID as the primary key.

Results

1. The Most Talked about US Presidential Candidate

```
*****************************************
Most Talked about US Presidential Candidates
*****************************************
Trump : 41.36%
Cruz : 24.69%
Clinton : 14.81%
Sanders : 10.49%
Kasich : 8.64%
```

2. US Presidential Forecast (based on positive comments)

```
*****************************************************
US Presidential Forecast (based on positive comments)
*****************************************************
Trump : 40.50%
Cruz : 26.45%
Clinton : 14.88%
Sanders : 12.40%
Kasich : 5.79%
```

3. Share of comments for each presidential candidate

{'Sanders': {'neg': 4, 'pos': 15}, 'Cruz': {'neg': 12, 'pos': 32}, 'Clinton': {'neg': 5, 'pos': 18}, 'Trump': {'neg': 25, 'pos': 49}, 'Kasich': {'neg': 1, 'pos': 7}}

4. Location of Tweets

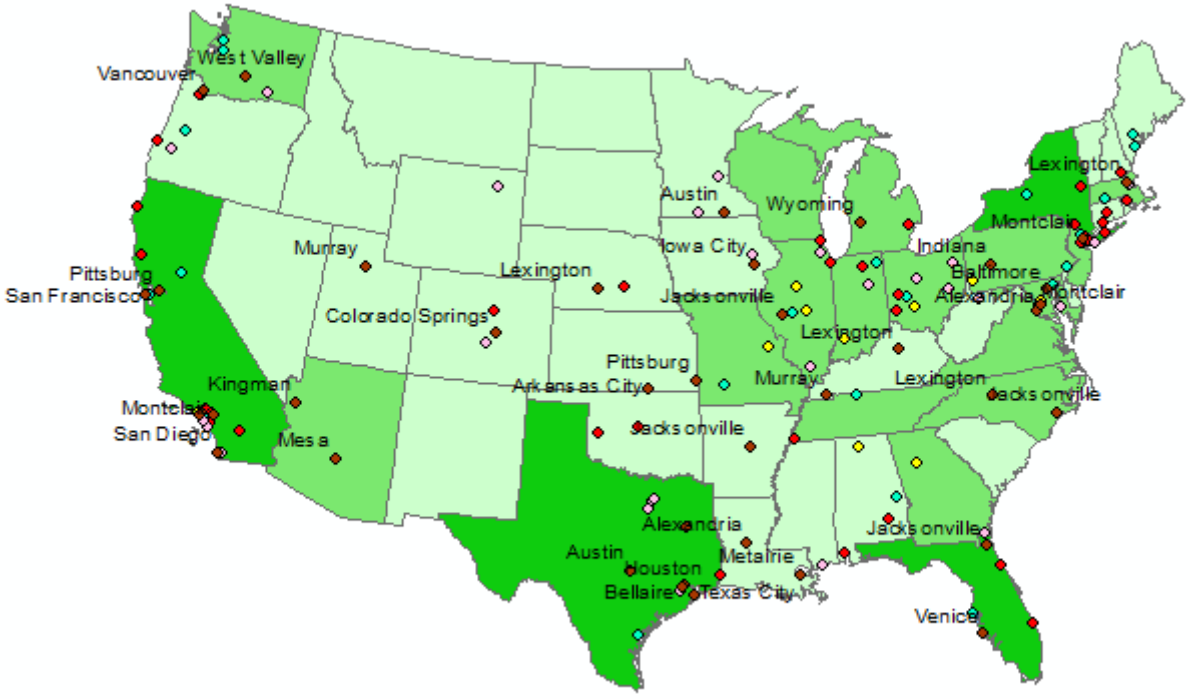
{**'Sanders'**:
{u'edgewood': 1, u'oakland': 1, u'portland': 1, u'coos bay': 1, u'az usa': 1, u'jacksonville': 2, u'pasco': 2, u'los angeles': 1, u'auburn': 2, u'largo': 2, u'seattle': 1, u'west covina': 1, u'springfield': 1},
'Cruz':
{u'houston': 1, u'alexandria': 1, u'pittsburg': 1, u'west valley': 1, u'murray': 2, u'wyoming': 4, u'san francisco': 1, u'venice': 2, u'inglewood': 3, u'san diego': 1, u'bellaire': 2, u'arkansas': 1, u'kingman ': 2, u'texas': 3, u'colorado springs': 1, u'iowa': 2, u'metairie': 1, u'indiana': 1, u'mesa': 1, u'jacksonville': 2, u'vancouver': 2, u'lexington': 1, u'montclaire': 1, u'los angeles': 1, u'austin': 1, u'baltimore': 1}, 'Clinton': {u'cedar hill': 1, u'addison': 2, u'cambridge': 1, u'bostonia': 1, u'gillette': 1, u'canon city': 1, u'richardson tx': 1, u'san diego': 1, u'pasco': 2, u'minneapolis': 1, u'fairmont': 2, u'long beach': 1, u'kingsland': 1, u'newport beach': 1, u'marion': 1, u'missouri': 1, u'alliance': 2, u'roseburg': 1, u'jacksonville': 2},
'Trump':
{u'cedar hill': 1, u'palm desert': 1, u'santa monica': 1, u'alliance': 2, u'canon city': 1, u'edmonton': 1, u'washington': 1, u'west valley': 1, u'united states': 2, u'troy': 1, u'indianana': 2, u'bellaire': 1, u'san francisco': 1, u'warsaw': 1, u'venice': 1, u'pearl city': 2, u'richardson tx': 1, u'munster': 1, u'londonderry ': 1, u'denver': 1, u'palm coast': 1, u'providence': 2, u'coos bay': 1, u'new york': 2, u'arkansas': 2, u'orange': 1, u'indiana': 1, u'indianana ': 2, u'texas': 1, u'colorado springs': 1, u'grand island': 2, u'middle town': 2, u'millburn': 2, u'metairie': 2, u'la canada': 2, u'mesa': 1, u'beaverton': 1, u'memphis': 1, u'bostonia': 4, u'mobile': 1, u'az usa': 1, u'coram': 1, u'eureka': 1, u'long beach': 1, u'uk ia': 1, u'stuart': 3, u'river': 1, u'gurr ne': 1, u'elk city': 1},
'Kasich':

{u'santa monica': 1, u'pittsburg': 2, u'metairie': 2, u'inglewood': 2, u'washington': 1, u'bellaire': 1, u'montclaire': 1, u'arkansas': 1, u'indiana': 1, u'mesa': 1, u'decatur': 1}
}

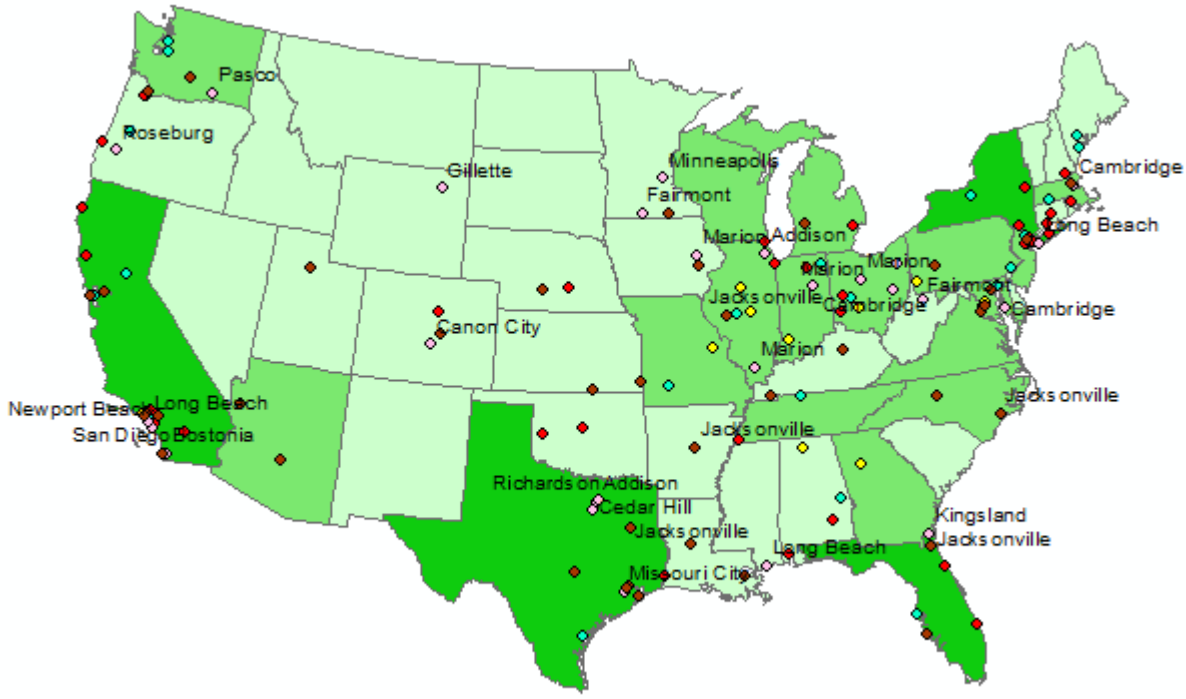
5. Labelled Locations for each candidate

Donald Trump

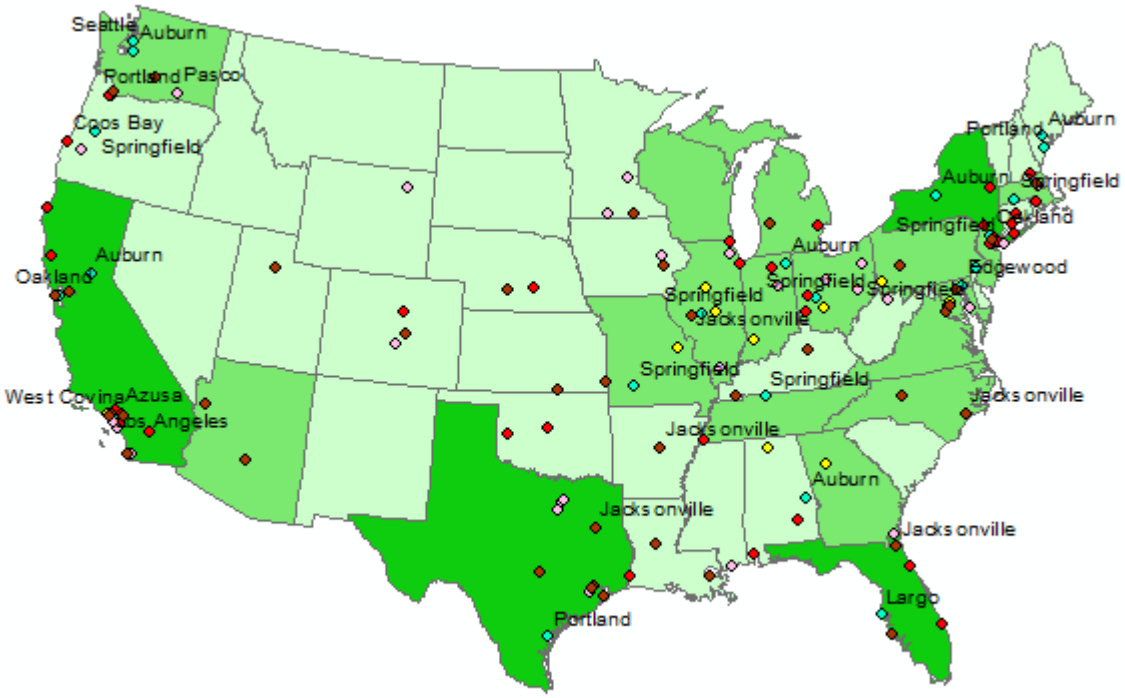
Ted Cruz



Hillary Clinton



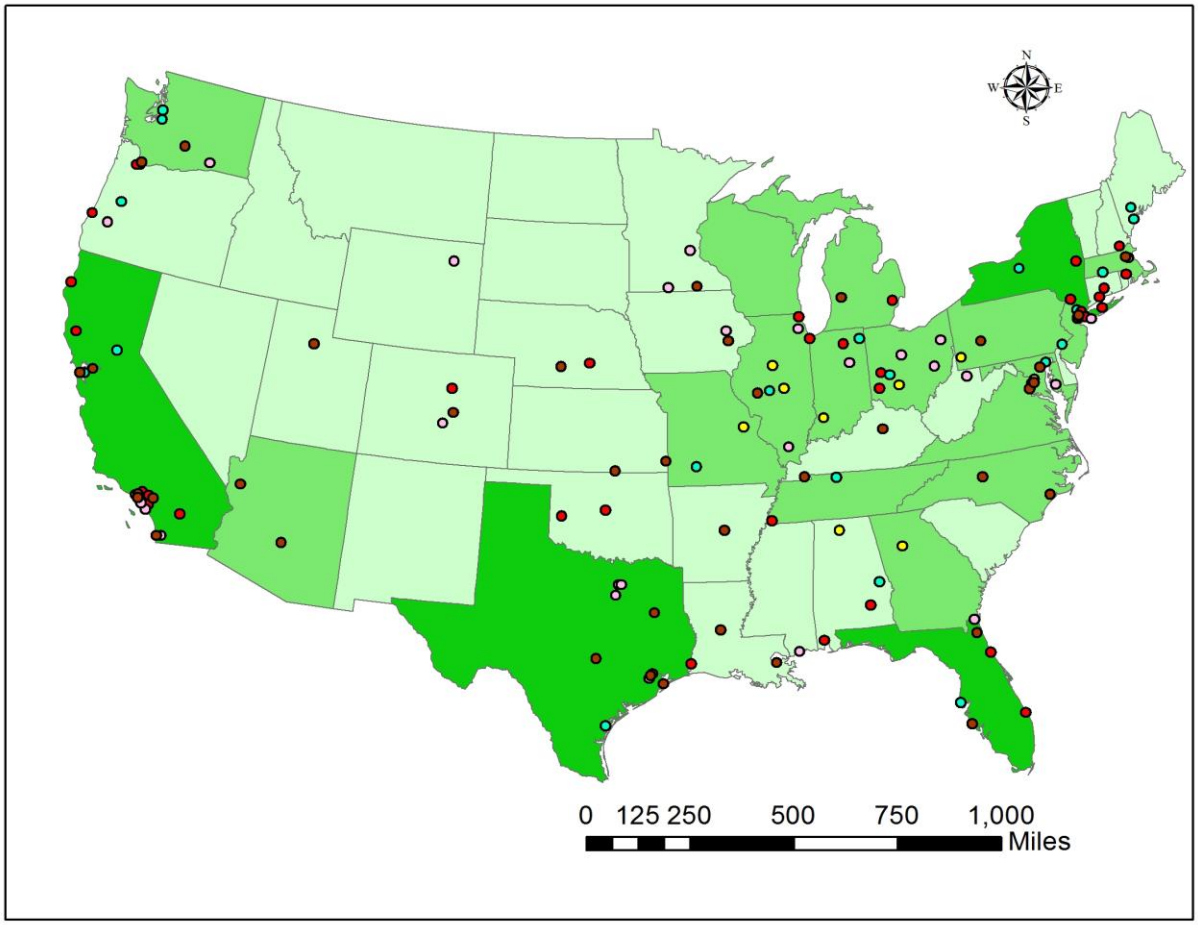
Bernie Sanders

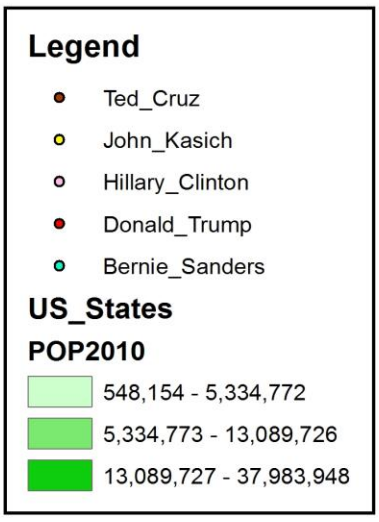


John Kasich


Pittsburg
Washington
Decatur
Washington
Washington
Washington
Pittsburg
Arkansas City
Indiana
Montclair
Montclair
Inglewood
Montclair
Santa Monica
Mesa
Decatur
Decatur
Bellaire
Metairie

2016 US Presidential election prediction from Twitter sentiments using MapReduce framework and ArcMap





Conclusion:

From the share of positive comments, the ultimate winner of the 2016 US Presidential Election can be easily predicted.

There were however some major challenges that had to be overcome. Recently Twitter stopped sharing the geo-locations of the tweets by default. These are only shared if the user has agreed to share his/her geo-location. As a result not too many tweets have the geo-locations in them. Out of 20000 tweets that were collected, only few thousands had them.

This also leads to another problem that is too much data processing. Out of 20000 tweets, only about 1% of the tweets were considered. Too much data processing can crash the system.

References

1. Presidency 2016, http://www.politics1.com/p2016.htm
2. MapReduce, https://en.wikipedia.org/wiki/MapReduce
3. Oauth2 library, https://dev.twitter.com/apps