Shefali Sharma
April 5, 2018

Twitter Data Extraction and Analysis

- Data Collection and Exploratory Data Analysis

Overview

Data Collection and Exploratory Data Analysis that involves replicating professional data analysis on a topic of current interest, and extending the data exploration to include another public data source.

Objectives

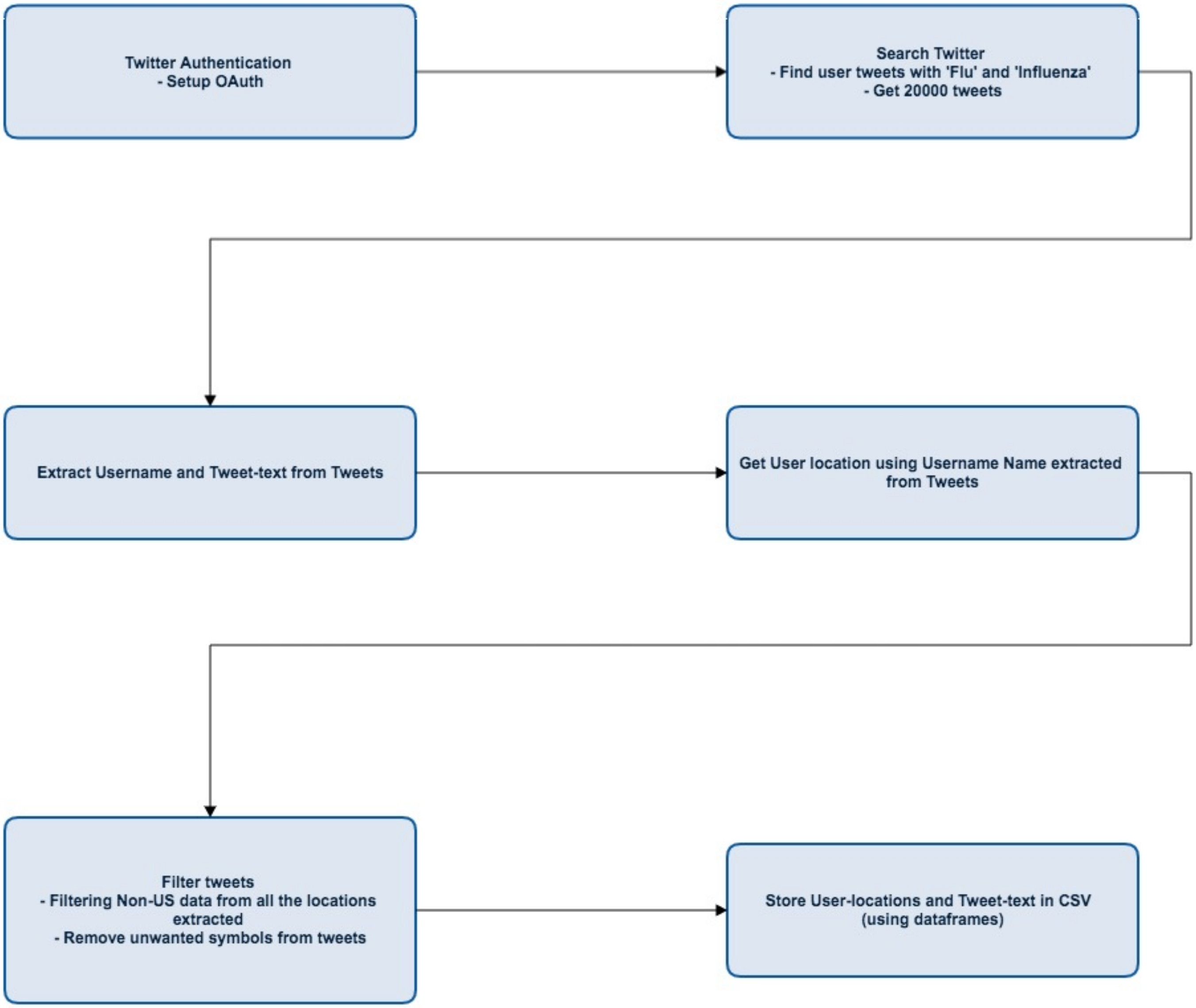
- Explore the real flu data, and replicate and learn from the analysis performed by experts in Center for Disease Control (CDC) [1] and related organizations.
- Collect data by querying Twitter REST API [2]. You will have to get a developer account on twitter and also get the credentials for your application (the twitter client) that you will be writing. Good query word related to “flu” gets you good data.
- Process data using twitteR [3] library package of R
- Visualize geo spatial information extracted from the tweets using geo-map libraries of R: ggplot2, ggmap, maps,and maptools [4]. Maps and geo codes are supported by Google map API.
- Compare CDC flu map with your own home-brewed flu map of the USA derived from the twitter data you obtained.

Process

Part 1:

This part focuses on:

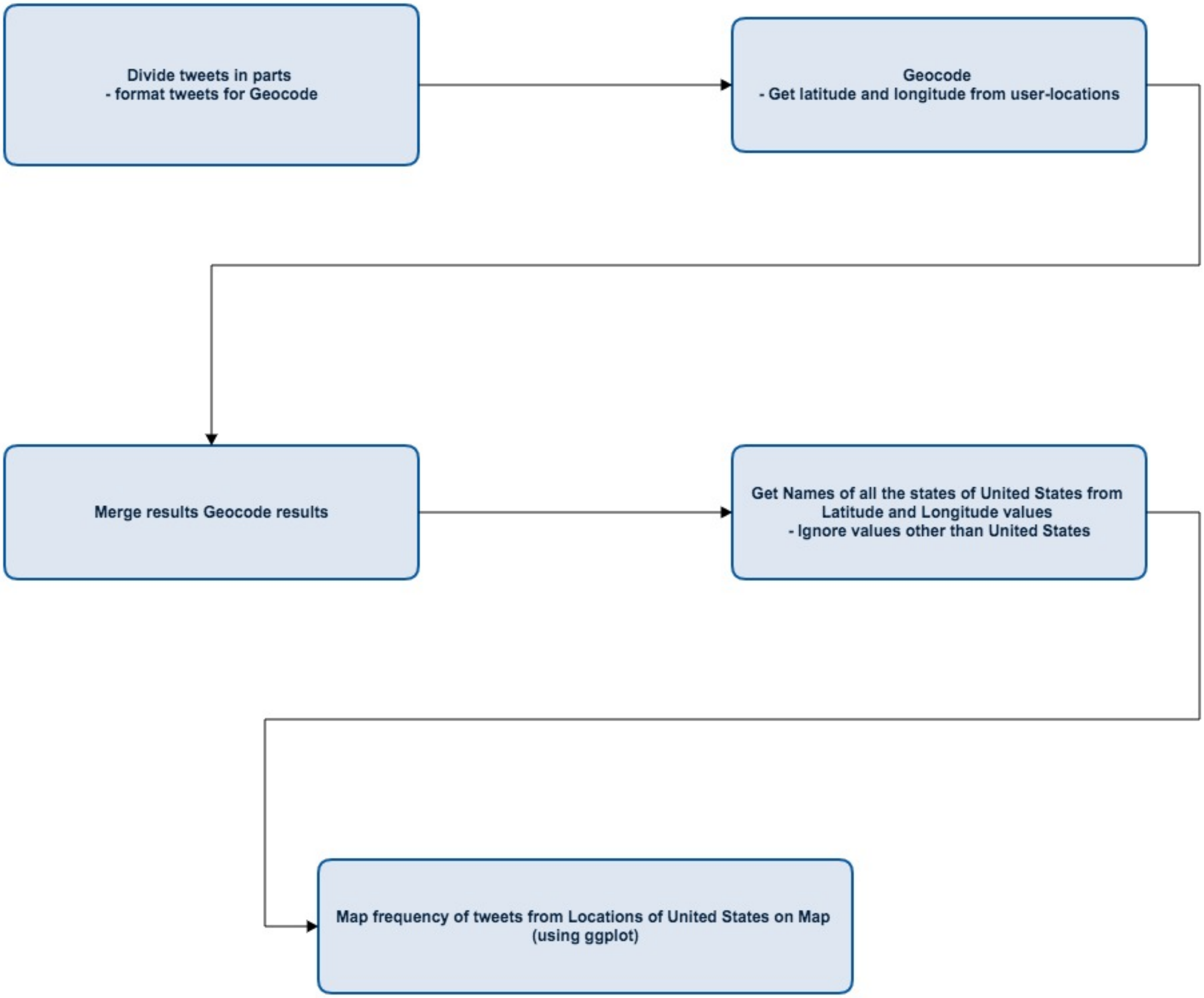
- Collecting relevant tweets
- Gathering locations from Users who tweeted
- Filter tweets to remove irrelevant and corrupt data


Twitter Authentication
- Setup OAuth
Search Twitter
- Find user tweets with 'Flu' and 'Influenza'
- Get 20000 tweets
Extract Username and Tweet-text from Tweets
Get User location using Username Name extracted from Tweets
Filter tweets
- Filtering Non-US data from all the locations extracted
- Remove unwanted symbols from tweets
Store User-locations and Tweet-text in CSV (using dataframes)

Part 2 :

This part focuses on:

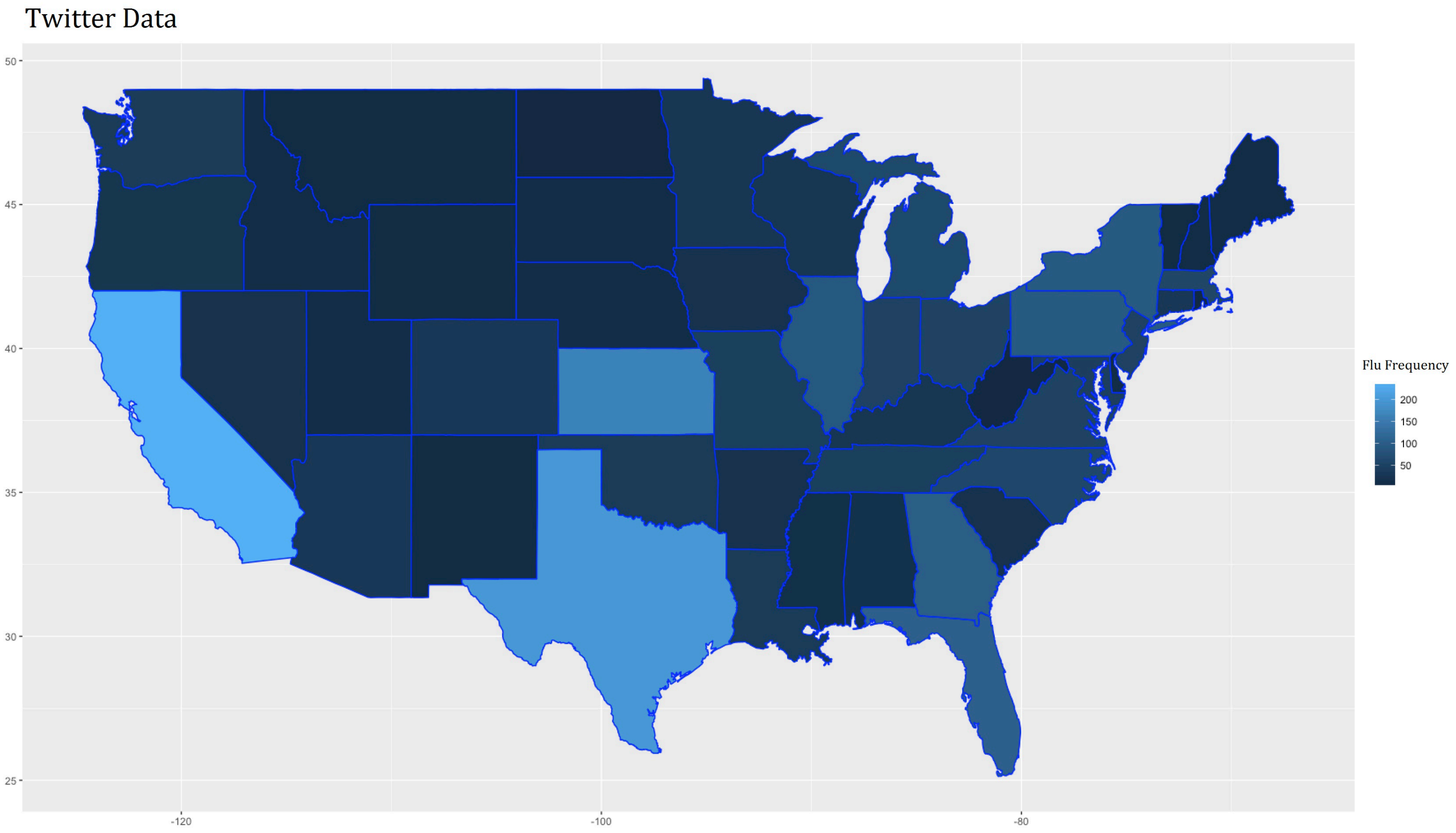
- Use locations to get latitude and longitude values - using Geocode
- Creating map to represent frequency of tweets from different states of the United States



Divide tweets in parts
- format tweets for Geocode
Geocode
- Get latitude and longitude from user-locations
Merge results Geocode results
Get Names of all the states of United States from Latitude and Longitude values
- Ignore values other than United States
Map frequency of tweets from Locations of United States on Map (using ggplot)

Results

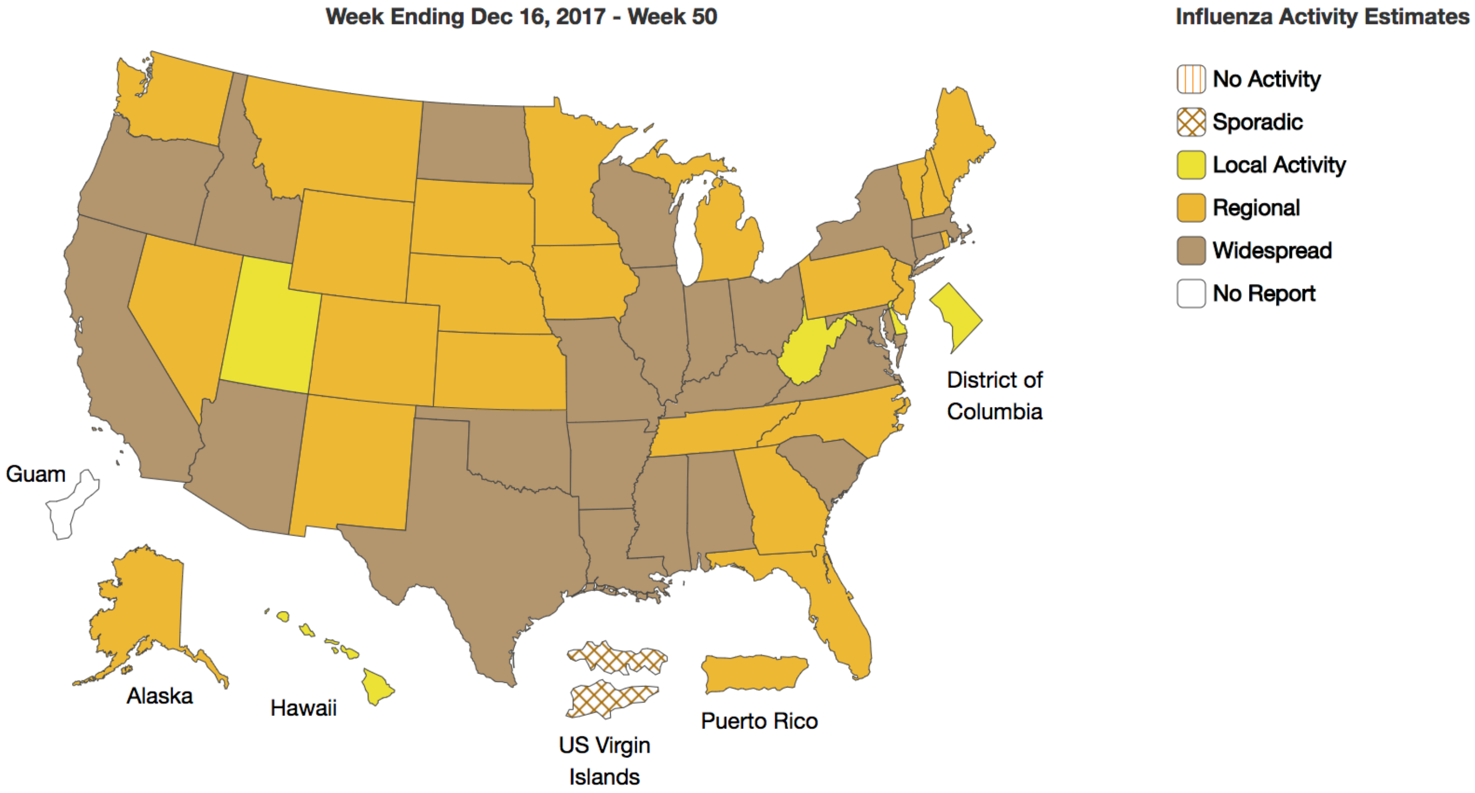
Results generated by collecting user generated data from Twitter for ‘Flu’ :



These results show high spread of flu in states like:

- California
- Texas
- Kansas
- Oklahoma
- And other nearby regions to the above mentioned states

CDC (Centers for Disease Control and Prevention) Results for ‘Flu’[5] :



This is the map depicting the spread of ‘Flu’ provided by CDC. It presents a stark similarity for the results collected from twitter and the actual spread of the disease.

States like Californias, Texas, Oklahoma and others do have a widespread Flu activity.

References

- [1] https://www.cdc.gov/flu/, CDC Weekly report on Flu Activity, last viewed 2018.
- [2] Twitter API. Twitter Developer https://dev.twitter.com/, last viewed 2017.
- [3] TwitteR package. https://cran.r-project.org/web/packages/twitteR/twitteR.pdf, last viewed 2017.
- [4] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal Vol. 5/1, June 2013, https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf
- [5] https://www.cdc.gov/flu/weekly/usmap.htm