DIGITAL EPIDEMIOLOGY OF DISEASES THROUGH DATA MINING

DATA MINING

CSE 3013

SLOT - G2

J Component Project Report

submitted by

ADHYA DAGAR -16BCI0160

SONAL SINGH -16BCE0411

NIKUNJ BHATIA -16BCE0838

VIT

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Table of contents

ABSTRACT

Infectious diseases have been on the rise in the past few decades. Every year over a million people in India die due to these illnesses. As our country develops, and no doubt that healthcare has improved over the past years, but at the same time more serious infections have also emerged. Just in this decade we saw Swine-Flu and Ebola claiming millions of lives. The large outbreaks or epidemics have become more common than ever costing our national economy a great deal too. This calls for *a better surveillance system* to detect such outbreaks at an early stage thus providing health care workers with a chance to track, contain and prevent and cure them.

Today social media has completely has brought about a new wave to traditional communication. And this change in social media has enabled data scientist to use the collected information to make various predictions. In our project we discuss how this can be applied to healthcare.

In our project we further try to investigate how social media can be used to prevent outbreak of diseases. The data that is made public on the web, which is on these social media handles, can be used to track and predict epidemics. As suggested above, search engine queries provide with a lot of data but for the sake of our project we'll deal exclusively with information procured from social media platforms. The difference between the kinds of the data we get from these two situations is that, through social media qualitatively different data is collected and this individuality helps us to better understand and analyse the trend.

Through this project we try to develop a system that helps us easily visualize trends by mining tweets.

1. INTRODUCTION

1.1 SYSTEM OVERVIEW

Through this project we have created a web portal or an interface where we can the geo-spatial location of a particular diseases that has been trending and take precautionary measures to prevent an epidemic.Information is gathered by mining data through tweets and crawling the web for news articles that'll give us more insights on the diseases.

1.2 OBJECTIVE:

86% of the world's population lives under mobile cellular network coverage and mobile phone networks routinely register data that can be used to track the location of all active mobile phone users, a fact that is critical in predicting the spread of a disease. Studying the magnitude and trends of population movements, allows us for close to real-time monitoring of the disease outbreak.

Outbreaks can also be tracked through searches and social media. According to Harvard Medical School epidemiologist John Brownstein, an estimated 37–52% of Americans seek health-related information on the Internet each year, mainly using search engines to find details about symptoms and treatments. We can analyze users' search terms, along with the location information encoded in their computers' Internet protocol addresses, for insights into current disease trends.

The process can be enhanced if we have more contextual information. This is possible through Twitter. Frequently, people who are experiencing symptoms or have been confirmed to have a specific disease post about it on social media.

Mining news Websites and web data can also aid in the process of predicting breakthroughs. We plan to capitalize on the trend that public now a days shares the daily whereabouts of their lives and routines. By scanning the web for keywords to track sentiments

We plan to mine data through these sources, similar to what a web crawler does, use *sentiment analysis on tweets* and *gather information locally pertaining to general health*.
Through the gathered information and studying the history of a particular geo location we plan to predict whether the diseases will be trending or not.

1.3 LIMITATIONS:

Social media collects unofficial information which is *unstructured and difficult to mine*. Thus it's not created for public health purposes. Algorithms designed to pick up "fever", for instance, may detect false positives such as "Bieber fever". So we need well-constructed algorithms for data mining. Therefore, technology is required to filter out background noises.

There's also a possibility of *people passing about false information*, especially in the comments section, hence making this data vaguely credible. Many a times these are just personal experiences that cannot be accounted officially for. So, in ought to be careful when dealing with data from social platforms.

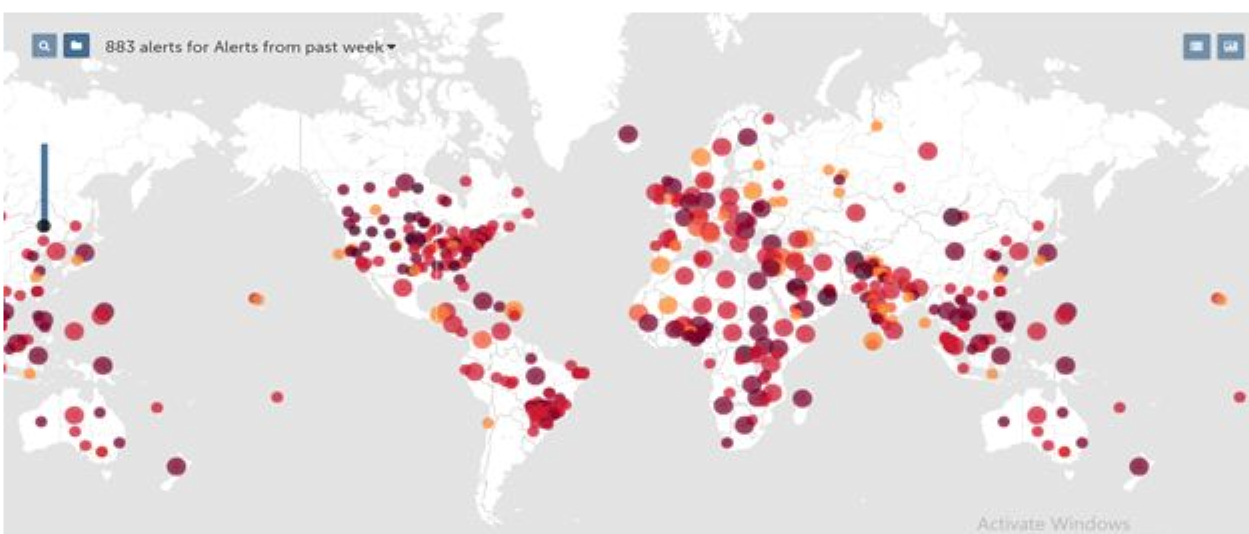
2. LITERATURE SURVEY

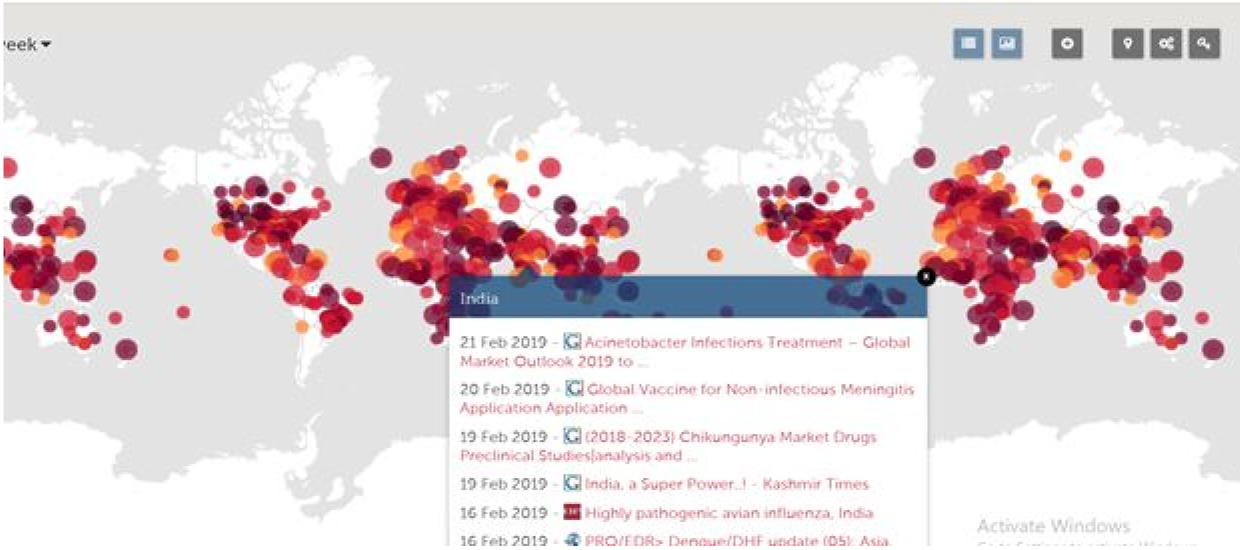
[3]There has been a growing interest in developing statistical models for detecting infectious disease outbreaks to enable effective control measures to be taken in a sufficiently timely fashion. Most early approaches relied on highly specialized data, including medical records or environmental time series. Recently, however, there has been a growing interest in monitoring disease outbreaks using publicly available data on the Web, including news articles, blogs, search engine logs and micro-blogging services, such as Twitter. Due to their volume, ease of availability, and citizen participation, such open source indicators have been shown to be effective at monitoring disease emergence and progression. Most prior work focuses on detecting outbreaks of common diseases, such as influenza, by discovering temporal patterns over predefined groups of keywords. However, many infectious diseases are rare with only a few incidences being reported in open sources. Forecasting outbreaks of rare diseases raises several challenges. We use a real-world scenario to illustrate these challenges.

[4]Early detection and tracking mechanisms are critical in reducing the impact of epidemics and preventing the epidemics from becoming unmanageable by making a rapid response. For example, the cholera epidemic killed over 100,000 people worldwide, and sickened 35 million people during the year 2010.

[5]Online Social Networks (OSNs) have rich information that can be utilized to develop an almost real-time pandemic surveillance system. The outcomes of OSN surveillance systems have demonstrated high correlations with the findings of official surveillance systems. However, the limitation in using OSN to track pandemic is in collecting representative data with sufficient population coverage. This challenge is related to the characteristics of OSN data. The data are dynamic, large-sized, and unstructured, thus requiring advanced algorithms and computational linguistics. OSN data contain significant information that can be used to track a pandemic. Different from traditional surveys and clinical reports, in which the data collection process is time consuming at costly rates, OSN data can be collected almost in real time at a cheaper cost. Additionally, the geographical and temporal information can provide exploratory analysis of spatiotemporal dynamics of infectious disease spread. However, on one hand, an OSN-based surveillance system requires comprehensive adoption, enhanced geographical identification system, and advanced algorithms and computational linguistics to eliminate its limitations and challenges. On the other hand, OSN is probably to never replace traditional surveillance, but it can offer complementary data that can work best when integrated with traditional data.

[1]Health Map is one the very first social media tools for disease surveillance. This was launched by Brownstein, who collaborated with Children's hospital Boston for his project. This websites retrieves its data from new websites, government alerts, individual personal accounts and numerous other such sources from around the world. All this data collected is then analyzed and these cases are aggregated on a global map displayed in real time. This team has been constantly working on predicting disease outbreaks and have been developing better technologies that gives access to anyone around the world to this information. Recently, they launched, Flu Near you and Outbreaks Near Me, IPhone applications that can be used by individuals to track diseases around them and their severity.





Although social media based surveillances are rapid, their accuracy can't be strongly spoken for. For rapid detection of epidemics, a tradeoff is required between speed and data validity. This type of data collection isn't completely going to eliminate the existing traditional systems but enhance it to improve the capacity of detecting outbreaks timely. Once a rapid signal is acquired, public health authorities can then investigate and confirm the epidemic, and traditional surveillance can take over.

Having these types of clues can help government agencies, and independent epidemiologists, understand how to allocate resources necessary for fighting and preventing illness and disease.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM :

There are some systems developed for digital epidemiology. Those systems use different approach like OSN models, ML algorithms, Big Data etc. One of the existing system used OSN model to predict disease spread in a particular area. The data are dynamic, large-sized, and unstructured, thus requiring advanced algorithms and computational linguistics. OSN data contain significant information that can be used to track a pandemic. Different from traditional surveys and clinical reports, in which the data collection process is time consuming at costly rates, OSN data can be collected almost in real time at a cheaper cost. Additionally, the geographical and temporal information can provide exploratory analysis of spatiotemporal dynamics of infectious disease spread. However, on one hand, an OSN-based surveillance system requires comprehensive adoption, enhanced geographical identification system, and advanced algorithms and computational linguistics to eliminate its limitations and challenges.

3.2 PROPOSED SYSTEM :

The system that we propose uses data mining techniques and uses the twitter data set to get updated information and locates the disease hotspot. It tries to prevent widespread of harmful diseases. It also fetches scholars updates new articles which also contribute to the geographical location of the disease. It is different from existing system is numerous ways. It takes updated data and information and then only shows the results. It uses proper mining algorithms to search articles. It also shows geographical hotspot on a website once the program runs.

4. REQUIREMENTS SPECIFICATION

Hardware Requirements : Our project is at a very early stage so for now we just need a laptop which can run the below mentioned softwares. When we expand our reach we may need servers to our website.

4.1 Software Requirements :

1. Google cloud services to access google maps
2. VS CODE : to run our python code
3. GDealt Project API : To get updated news articles
4. Tweepy library : To get twitter data of diseases

5. SYSTEM DESIGN SPECIFICATION

5.1 SYSTEM ARCHITECTURE

We have divided our project into 3 modules-
The first part consists of programs that help us retrieving tweets that contain information about various trending diseases.

The second part consists of plotting the locations that have been retrieved through the downloaded tweets on a dynamic real world map.

The third part consists lists of relevant articles about the diseases sorted according to the diseases searched.

All these three modules have been integrated and can be viewed through an interface which has been hosted online on a free hosting app.

5.2 MODULE DESCRIPTION

MODULE 1-RETRIEVING TWEETS

Text mining is the application of natural language processing techniques and analytical methods to text data in order to derive relevant information.

In order to access Twitter Streaming API, we need to get 4 pieces of information from Twitter: API key, API secret, Access token and Access token secret.

We will be using a Python library called Tweepy to connect to Twitter Streaming API and downloading the data.

To make sure that the geolocation and the coordinates also get downloaded along with the contents of the tweets we use the following command `csvWriter.writerow([tweet.user.location, tweet.text])`

By following the above approach we get a CSV file that contains a table with location,Twitter Handler's identity and tweet content.

MODULE 2- PLOTTING ON MAPS

Through this module we plan to convert our locations to actual coordinates on maps.We use the following command to do so-
```
gmap3 = gmplot.GoogleMapPlotter(30.3164945,
                                78.03219179999999, 13)
```
First we form a list of available latitudes and longitudes and within these coordinates our code converts the locations in the CSV file to a pair of latitudes and longitudes and shows it on the map.

Using Google Map Plotter we can visualize our data on a custom map using Google My Maps.
We have also used gmap3 query which lets you customize maps with your own content and imagery for display on web pages and mobile devices. The Maps JavaScript API features four basic map types (roadmap, satellite, hybrid, and terrain) which you can modify using layers and styles, controls and events, and various services and libraries.

MODULE 3- NEWS DEMONSTRATION (WEB CRAWLING)

For the benefit of the reader we are complementing our system with additional news articles related to the trending disease.This will help the viewer to get more information about the disease.

We are using the following API to crawl through articles all over the web-
```
api.gdeltproject.org/api/v2/doc/doc?query=domain:washingtonpost.com&query=flu&sourcelang:english&maxrecords=250&format=json&startdatetime=20190101000000&enddatetime=20190401000000")
```

The GDELT Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world.

We use the gdelt project API to crawl through news article and the we display the collected new information and display it.The information can be filtered by location,data of creation,topic etc.

5.3 DATABASE DESIGN

Our backend relies on MySQL server and to carry this out we downloaded XAMPP server for MAC which manages Apache web server and phpMyAdmin.For hosting purpose we are using a free web hosting service called webhostingapp.The main purpose of the phpMyAdmin tool is to manage your databases.
The articles retrieved are stored in the database.

6. SYSTEM IMPLEMENTATION

By joining these three individual components into a single interface we can view a working system which shows trending diseases like flu,tuberculosis and dengue.Apart from this we can also see related news articles which will complement this system and will let us know more about the subject.

7. CONCLUSION AND FUTURE ENHANCEMENTS

The already existing system to track diseases in our country is extremely traditional in its ways. It involves formal reporting to the government by doctors and hospitals. Although this data collected is of high accuracy and can give us insights to how effective certain control strategies may be, the only problem is that it’s very slow. And when we’re dealing with diseases in a country like India, time is of the essence, as the rate of spread of such infections can increase exponentially. The earlier these epidemics are detected, the easier it is to contain them. Therefore, these traditional ways needed evolution.

Health care has been evolving dramatically over time and now as we enter the age of data, it’s no surprise that the leading idea of the time is to base *developments in healthcare on the results of data analysis.* Today, as social media becomes an integral part of everyone’s life, it’s a revolution for data sciences as everything you post on Instagram and Facebook and other social media handles, the hashtags that are trending, the search engine queries, all of this together provides scientists with predictors for various things, one such being illness and diseases in real time.

To further enhance this project we can use strong and dynamic info-graphics and data visualisation maps that can give us more accurate information.To ease the user, options like searching within a specific geo-spatial location can also be provided.

8. APPENDICES

APPENDIX 1 - SAMPLE SOURCE CODE

Code for retrieving Tweets and putting them in a CSV file

```python
import tweepy
import csv
import pandas as pd
####input your credentials here

consumer_key = "kEE3DM1v28wb5eNkxOR8oyA38"
consumer_secret = "ZDuorWdbbl9HzTw0wnQpvXI3NyJcZbxsncRiCnjJL9RJsOepFn"
access_token = "877193643341135872-IZ4ZQ29dZr3USnAyUgCCSwg88P0Ng3K"
access_token_secret = "gIO13HApWBf24ey50MvTG7G5whUGIFoyLa3PP76PxmkdM"

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth,wait_on_rate_limit=True)
#####United Airlines
# Open/Create a file to append data
csvFile = open('ua6.csv', 'a')
#Use csv Writer
csvWriter = csv.writer(csvFile)

for tweet in tweepy.Cursor(api.search,q="#flu",count=1000, lang="en",
since="2019-01-01").items():
   print (tweet.created_at, tweet.text)
   print(tweet.user.location)
   print("\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\")
   csvWriter.writerow([tweet.user.location, tweet.text])

import tweepy
 # Fill the X's with the credentials obtained by
# following the above mentioned procedure.
consumer_key = "kEE3DM1v28wb5eNkxOR8oyA38"
consumer_secret = "ZDuorWdbbl9HzTw0wnQpvXI3NyJcZbxsncRiCnjJL9RJsOepFn"
access_key = "877193643341135872-IZ4ZQ29dZr3USnAyUgCCSwg88P0Ng3K"
access_secret = "gIO13HApWBf24ey50MvTG7G5whUGIFoyLa3PP76PxmkdM"
 # Function to extract tweets
```

Code for retrieving Tweets

```python
def get_tweets(username):

       # Authorization to consumer key and consumer secret
```

```
        auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
         # Access to user's access key and access secret
        auth.set_access_token(access_key, access_secret)
         # Calling api
        api = tweepy.API(auth)
         # 200 tweets to be extracted
        number_of_tweets=50
        tweets = api.user_timeline(screen_name=username)
         # Empty Array
        tmp=[]
         # create array of tweet information: username,
        # tweet id, date/time, text
        tweets_for_csv = [tweet.text for tweet in tweets] # CSV file
created
        for tweet in tweets:
            print(tweet.entities.get('hashtags'))
            print(tweet.text)
            print("\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\")
        for j in tweets_for_csv:
             # Appending tweets to the empty array tmp
            # print(j)
            # print("\n\n")
            # #print("Name:", tweet.author.name.encode('utf8'))
            # #print("Screen-name:",
tweet.author.screen_name.encode('utf8'))
            # print("Tweet created:", tweet.created_at)
            # print("Tweet:", tweet.text.encode('utf8'))
            # print("Retweeted:", tweet.retweeted)
            # print("Favourited:", tweet.favorited)
            # #print("Location:", tweet.user.location.encode('utf8'))
            # print("Time-zone:", tweet.user.time_zone)
            # print("Geo:", tweet.geo)
            # print("//////////////////")
            tmp.append(j)
         # Printing the tweets
        #print(tmp)
```

```
# Driver code
if __name__ == '__main__':
    # Here goes the twitter handle for the user
   # whose tweets are to be extracted.
   get_tweets("@UNICEFIndia")
```

Code for plotting tweets on Map

```
import gmplot
 latitude_list = [ 30.3358376, 30.307977, 30.3216419 ]
```

```
longitude_list = [ 77.8701919, 78.048457, 78.0413095 ]
 gmap3 = gmplot.GoogleMapPlotter(30.3164945,
                                78.03219179999999, 13)
 # scatter method of map object
# scatter points on the google map
gmap3.scatter( latitude_list, longitude_list, '# FF0000',
                              size = 40, marker = False )
 # Plot method Draw a line in
# between given coordinates
gmap3.plot(latitude_list, longitude_list,
          'cornflowerblue', edge_width = 2.5)
gmap3.draw( "Macintosh HD\\Users\\adhyadagar\\map13.html" )
```

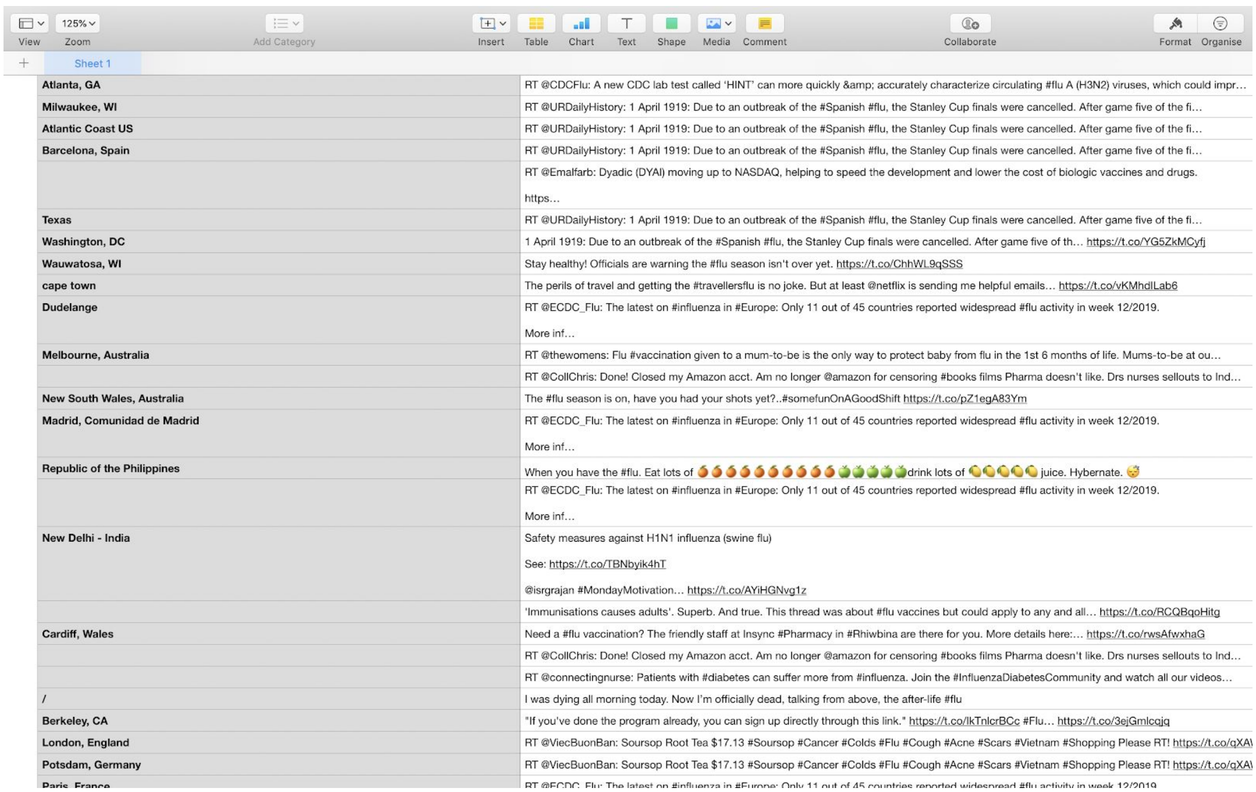
Code for getting articles through gdelt API

```
import requests
import json
from newspaper import Article
import csv
#import time
#get page
response =
requests.get("https://api.gdeltproject.org/api/v2/doc/doc?query=domain:wa
shingtonpost.com&query=flu&sourcelang:english&maxrecords=250&format=json&
startdatetime=20190101000000&enddatetime=20190401000000")
data = response.json()

articles_d = data['articles']
articles=[]
for i in range(len(articles_d)):
   articles.append(articles_d[i]['url'])
print(articles)

with open("C:\\Users\\Intel\\Desktop\\Python 2019\\washingtonpost.csv",
'w',newline='') as csvfile:
   writer = csv.DictWriter(csvfile,fieldnames=["Body ID","articleBody"])
   writer.writeheader()
   ID=0
   for i in range(len(articles)):

       #i = articles[0]
       link = articles[i]
       #    print(link)
       article1=Article(link)
       article1.download()
#             while article1.download_state != 2:
#ArticleDownloadState.SUCCESS is 2
```

```
#                       time.sleep(1)
        article1.parse()
        a=article1.text
        writer.writerow({"Body ID":ID,"articleBody":a})
        ID+=1
        #     article1.nlp()
        #     print(article1.keywords)
```

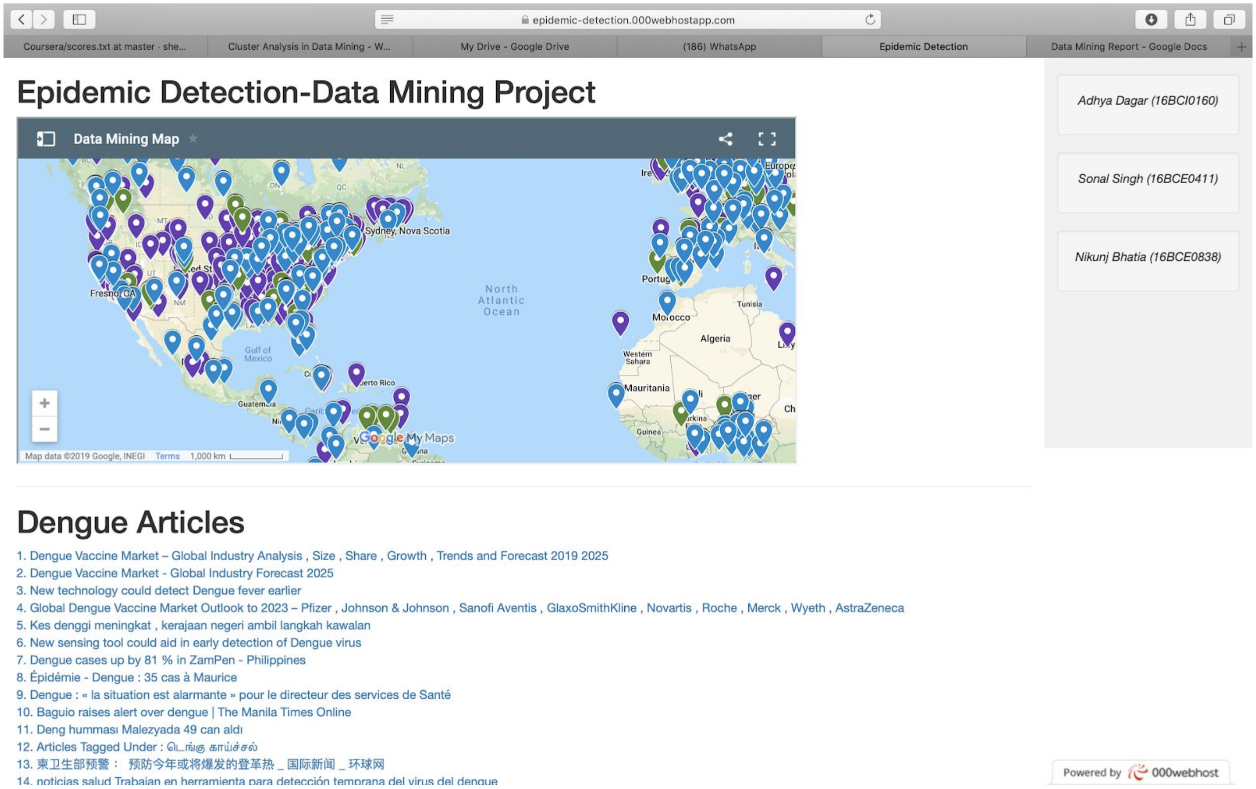
APPENDIX 2 - SCREEN SHOTS /OUTPUTS

Tweets and their locations in the CSV File
eg.CSV file containing tweets about flu



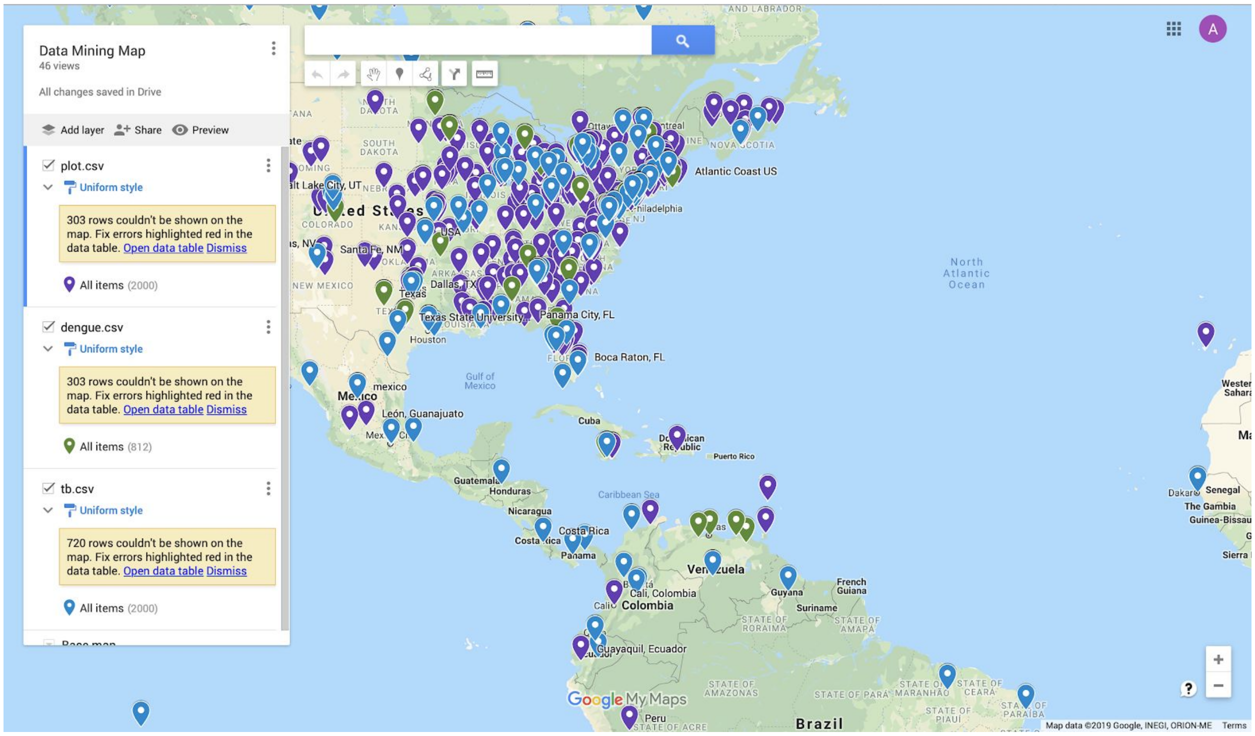
Atlanta, GA	RT @CDCFlu: A new CDC lab test called 'HINT' can more quickly & accurately characterize circulating #flu A (H3N2) viruses, which could impr…
Milwaukee, WI	RT @URDailyHistory: 1 April 1919: Due to an outbreak of the #Spanish #flu, the Stanley Cup finals were cancelled. After game five of the fi…
Atlantic Coast US	RT @URDailyHistory: 1 April 1919: Due to an outbreak of the #Spanish #flu, the Stanley Cup finals were cancelled. After game five of the fi…
Barcelona, Spain	RT @URDailyHistory: 1 April 1919: Due to an outbreak of the #Spanish #flu, the Stanley Cup finals were cancelled. After game five of the fi…
	RT @Emalfarb: Dyadic (DYAI) moving up to NASDAQ, helping to speed the development and lower the cost of biologic vaccines and drugs. https…
Texas	RT @URDailyHistory: 1 April 1919: Due to an outbreak of the #Spanish #flu, the Stanley Cup finals were cancelled. After game five of the fi…
Washington, DC	1 April 1919: Due to an outbreak of the #Spanish #flu, the Stanley Cup finals were cancelled. After game five of th… https://t.co/YG5ZkMCyfj
Wauwatosa, WI	Stay healthy! Officials are warning the #flu season isn't over yet. https://t.co/ChhWL9qSSS
cape town	The perils of travel and getting the #travellersflu is no joke. But at least @netflix is sending me helpful emails… https://t.co/vKMhdlLab6
Dudelange	RT @ECDC_Flu: The latest on #influenza in #Europe: Only 11 out of 45 countries reported widespread #flu activity in week 12/2019. More inf…
Melbourne, Australia	RT @thewomens: Flu #vaccination given to a mum-to-be is the only way to protect baby from flu in the 1st 6 months of life. Mums-to-be at ou…
	RT @CollChris: Done! Closed my Amazon acct. Am no longer @amazon for censoring #books films Pharma doesn't like. Drs nurses sellouts to Ind…
New South Wales, Australia	The #flu season is on, have you had your shots yet?..#somefunOnAGoodShift https://t.co/pZ1egA83Ym
Madrid, Comunidad de Madrid	RT @ECDC_Flu: The latest on #influenza in #Europe: Only 11 out of 45 countries reported widespread #flu activity in week 12/2019. More inf…
Republic of the Philippines	When you have the #flu. Eat lots of 🍊🍊🍊🍊🍊🍊🍊🍊🍊🍊🍏🍏🍏🍏🍏drink lots of 🍋🍋🍋🍋🍋 juice. Hybernate. 😴
	RT @ECDC_Flu: The latest on #influenza in #Europe: Only 11 out of 45 countries reported widespread #flu activity in week 12/2019. More inf…
New Delhi - India	Safety measures against H1N1 influenza (swine flu) See: https://t.co/TBNbyik4hT @isrgrajan #MondayMotivation… https://t.co/AYiHGNvg1z
	'Immunisations causes adults'. Superb. And true. This thread was about #flu vaccines but could apply to any and all… https://t.co/RCQBqoHitg
Cardiff, Wales	Need a #flu vaccination? The friendly staff at Insync #Pharmacy in #Rhiwbina are there for you. More details here:… https://t.co/rwsAfwxhaG
	RT @CollChris: Done! Closed my Amazon acct. Am no longer @amazon for censoring #books films Pharma doesn't like. Drs nurses sellouts to Ind…
	RT @connectingnurse: Patients with #diabetes can suffer more from #influenza. Join the #InfluenzaDiabetesCommunity and watch all our videos…
/	I was dying all morning today. Now I'm officially dead, talking from above, the after-life #flu
Berkeley, CA	"If you've done the program already, you can sign up directly through this link." https://t.co/lkTnlcrBCc #Flu… https://t.co/3ejGmlcqjq
London, England	RT @ViecBuonBan: Soursop Root Tea $17.13 #Soursop #Cancer #Colds #Flu #Cough #Acne #Scars #Vietnam #Shopping Please RT! https://t.co/qXA\
Potsdam, Germany	RT @ViecBuonBan: Soursop Root Tea $17.13 #Soursop #Cancer #Colds #Flu #Cough #Acne #Scars #Vietnam #Shopping Please RT! https://t.co/qXA\
Paris, France	RT @ECDC_Flu: The latest on #influenza in #Europe: Only 11 out of 45 countries reported widespread #flu activity in week 12/2019

Web Interface



Retrieved Articles on the topic

Google Map with Trending Tweets



9. REFERENCES:

[1]Using Social Media to Predict and Track Disease Outbreaks by Charles W Schmidt, 2018.

[2]Predicting Infectious Disease Using Deep Learning and Big Data by Sangwon Chae, Sungjun Kwon, and Donghyun Lee*, 2018

[3]SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources Theodoros Rekatsinas* , Saurav Ghosh† , Sumiko R. Mekaru‡ , Elaine O. Nsoesie‡ John S. Brownstein‡ , Lise Getoor§ , Naren Ramakrishnan†

[4] Detecting and Tracking Disease Outbreaks by Mining Social Media Data Yusheng Xie,1,2 Zhengzhang Chen,1,2 Yu Cheng, Kunpeng Zhang,

[5]Social Media Mining For Public Health Monitoring And Surveillance Michael J. Paul,1* Abeed Sarker,2 John S. Brownstein

[6]Mining Twitter Data For Influenza Detection and Surveillance, Kenny Byrd, 2016

[7] Detecting and Tracking Disease Outbreaks by Mining Social Media Data, Yusheng Xie, Zhengzhang Chen, Alok N. Choudhary, Published in IJCAI 2013

Other links:

1. https://bmcnephrol.biomedcentral.com/articles/10.1186/1471-2369-14-114
2. http://www.icmr.nic.in/ijmr/2007/march/0302.pdf
3. https://www.sciencedirect.com/science/article/pii/S1470204514701159
4. https://www.thiemeconnect.com/products/ejournals/abstract/10.1055/s-0038-1634449
5. https://pulse.embs.org/january-2017/tracking-disease/