Modeling the Spread of Zika Through Twitter® Analysis

The recent outbreak of the Zika virus, originating in Brazil, has rapidly spread throughout the world, with thousands of people affected worldwide. Transmitted sexually, through birth, and by mosquitoes, Zika is suspected of causing microcephaly and Guillain-Barré syndrome, and currently has no known cures. Because major efforts are going into awareness and prevention, a novel computational model was created in this study that can analyze human movement in Zika-exposed regions and predict where the virus might spread to. Using Twython as a wrapper, a Python program was created to scrape tweets from Twitter® pertaining to Zika by using keywords and geolocation. The authors of tweets were grouped into cohorts and their subsequent tweets were collected to monitor their movement. This data was plotted onto a map of *Aedes aegypti* and *Aedes albopictus* mosquito populations in order to predict locations where Zika might spread to next. The model's predictive potential was illustrated by its mapping of at-risk users to South Korea, Thailand, the Philippines and Spain before the recent outbreaks of Zika. The predictions provided from this model could enable at-risk communities to take preventative measures to reduce the mosquito population in their areas and stop the spread of Zika.

Introduction

The Zika virus is a mosquito borne flavivirus central to a currently ongoing epidemic in the Americas. Historically, Zika virus symptoms have been constrained to small clusters of patients until 2007, when the first major outbreak occurred in the Yap Islands of Federated States of Micronesia and 73% of the population was infected (Plourde, 2016). Since then, the Zika virus has spread and the most recent outbreak began in May 2015 when multiple cases of microcephaly were reported in Brazil.

From the family *Flaviviridae,* Zika is related to several other mosquito borne viruses such as Dengue, yellow fever, and the west nile virus (Kuno, 1998). The Zika virus primarily spreads through its mosquito vectors *Aedes aegypti* and *Aedes albopictus* and the virus can be transmitted from humans *to* mosquitoes as well as the reverse, which greatly increases the impact of human travel on the spread of the virus (Messina, 2016) (CDC Transmission and Risks, 2016). Furthermore, recent studies also indicate an even wider array of human-related transmission risks. There is now evidence that Zika can be transmitted from a pregnant mother to her fetus if infection occurs during pregnancy or near childbirth (Johansson, 2016). Zika is also transmitted sexually and there have been investigations into the presence of the virus in blood transfusions, with two such transmission cases being reported in Brazil (D'Ortenzio, 2016), (Motta, 2016).

Once acquired, the Zika virus has been shown to lead to a host of deleterious conditions with an impact range stretching from infants to adults. In particular, recent studies show that fetuses in pregnant women exposed to Zika are at-risk of developing microcephaly (Johansson, 2016). Yet other studies also indicate that large numbers of people infected with the Zika virus also show signs of having Guillain–Barré syndrome, a disorder in which the immune system attacks the peripheral nervous system (Dos Santos, 2016). Current studies of the virus in mouse models also suggests that Zika can affect adult brain cells, and in particular populations of cells responsible for learning and memory, in a manner very similar to Alzheimer's disease (Li, 2016). Therefore, with new research bringing newer and more detrimental effects of Zika to the forefront and with the full range of effects unknown, it is imperative that along with developing a cure, research must focus on protecting people in infected areas and preventing transmission of the disease.

Along with the fact that there is no current treatment for the Zika virus, the World Health Organization announced that research should prioritize diagnosis and prevention over treatment (World Health Organization, 2016) and major companies such as Google are investing data scientists into determining where the next outbreak will occur (The Baltimore Sun, 2016).

If a government knows in advance that they are at-risk of having a future Zika outbreak, they can take many preventative measures that will limit the spread of the virus and protect already vulnerable populations. These measures include distributing condoms to prevent sexual transmission and spraying areas with mosquitoes to reduce their numbers (CDC, 2015). Therefore, an accurate model capable of predicting locations vulnerable to future Zika outbreaks is imperative in responding to Zika and reducing its detrimental effects.

The research project, therefore, is aimed at creating a model of Zika transmission using social media data on the movements of potentially infected humans in order to predict future locations where the Zika virus may spread. Billions of people use social media such as Twitter® to discuss virtually all aspects of their lives, thus presenting researchers with vast amounts of accessible data. Furthermore, analysis of social media inputs not only provides a way to study human interactions and matter of public health, but also human movement.

Due to the easy availability of Twitter®'s API, there has been an explosion in the research concerned with using 'big data' in analysis (Batrinca, 2015). There is a vast wealth of literature showing the use of Twitter® analysis in assessing global mobility patterns (Dredze, 2016) and this data has been used in a variety of applications, including disaster response and recovery after major events such as Hurricane Sandy (Wang, 2015).

Thus, the goal of this study was to develop a program that obtains Twitter® data from users around known Zika-infected regions and refine those tweets in order to create cohorts of users with geolocation identifiers. Many people of varied nationalities are moving into and out of Rio de Janeiro between August 5, 2016 and August 21, 2016 due to the Summer Olympic Games. This not only provides a large number of cases to analyze, but also brings many people into and out of a Zika-infected country, thus giving the model, which collected tweets in this time range, a global sample size. This model can then be assessed relative to current data on Zika outbreaks in order to show the potential predictive capacity of the program and its usefulness in preventing further spread of Zika and identifying regions vulnerable to future outbreaks.

Methods

All work described below was completed by Competition Entrants.

In order to create a collection of tweets and map geolocation, it was necessary to create a database of tweets with geographical coordinates. Python 3.5.2 32-bit and Twython, a Python wrapper used to collect tweets using the application program interface (API) provided by Twitter®, were downloaded and installed. The Twython source code was installed into the site-packages directory.

In order to access data from Twitter®, a Twitter® app was created, and a Consumer Key and Consumer Secret were obtained. Twython was imported, and using the App Key and App Secret, the Access Token was obtained (**Figure 2**).

Tweets were collected from three main locations: Rio de Janeiro (within a 500 mile radius), Miami (within a 100 mile radius), and New York City (within a 100 mile radius) based on the geolocation the tweets were sent from, and whether or not the tweets had at least one of the search terms chosen to focus on Zika cases (**Figure 5**). These three locations were chosen because they had many cases of Zika and were large cities with human movement (CDC, Case Counts in the US, 2016). The keywords were searched in English and Portuguese. A maximum of 200 tweets were scraped per keyword-geocode-date pairing, a limitation set by Twitter®. The code was run to collect the maximum number of tweets starting on August 18, 2016, which allowed tweets from August 10, 2016 to be collected, as Twitter® limits callback to nine days. Early August was chosen as the starting point because of the large amount of movement into and out of Rio de Janeiro for the Summer Olympic Games, which began on August 5, 2016 and ended on August 21, 2016. The date the tweet was created, the unique ID attached to each tweet, the unique ID attached to each user/author of the tweet, each user's screen name, the text in each tweet, whether or not geolocation was enabled by each user, and the geographical data of each tweet were collected from each tweet.

In order to store and process the tweets, Microsoft SQL Server 2008 Express and pyodbc were installed. A new database was created, and a table called Twitter_Scrape was made with 15 columns (**Figure 3**). Users that had geo enabled allowed the location they were tweeting from to be displayed, which was required for the movement of users to be monitored. The column

Parsed_GEO would be filled to contain the coordinates corresponding to each tweet, but to the tenths decimal place instead of the millionths, as changes in the tenths digit reflect movement from one city to another.

After the Twitter_Scrape table was set up, the code for inserting scraped tweets into the table was created (**Figure 1.1**). The full code can be accessed in a public GitHub respository at https://github.com/HLAP2016/Zika-Twitter-Scrape/blob/master/OAuth2AccessTwitterScrape.py . The insert statement was printed for every tweet scraped to check that the statement was running properly. Within the Twitter_Scrape table were entries both with and without geo enabled. In order to focus on users that would show movement through their geolocation in their tweets, the entries without geo enabled, or where "Geo_enabled <> 'False'", needed to be removed. In a query, a new table, GEO_Twitter_Scrape was created in the database with some of the same columns and properties as those in Twitter_Scrape, and tweets with geo enabled were duplicated into GEO_Twitter_Scrape (**Figure 1.2**).

In order to observe the movement of users, the user of each tweet stored in Twitter_Scrape table was tracked, and any tweets the user made subsequently was collected and stored in GEO_Twitter_Scrape by using Since_ID (**Figure 1.3**). The full code can be accessed at https://github.com/HLAP2016/Zika-Twitter-Scrape/blob/master/Since_ID.py. The setup and basis of the initial tweet scraping code was implemented in the the Since_ID code as well. A new cursor was added so that tweets from Twitter_Scrape were grabbed, and new tweet entries were inserted into GEO_Twitter_Scrape.

Because users could be constantly publishing more tweets, the greatest (and thus most recent) tweetID from each user in GEO_Twitter_Scrape was set as the since_id. A third .py code, Max_Since_ID, was created (**Figure 1.4**), and can be accessed at https://github.com/HLAP2016/Zika-Twitter-Scrape/blob/master/Max_Since_ID.py. The maximum tweetID for each user was selected and run as the since_id to scrape subsequent tweets from the user. A try-except exception was also added because it was possible for users to delete their tweets. For further searches of subsequent tweets from each user, Max_Since_ID would be run, both grabbing and inputting tweets into GEO_Twitter_Scrape.

When both the initial tweets containing geolocation and the subsequent tweets by the same group of users were in GEO_Twitter_Scrape, the changes in coordinates needed to be analyzed. The coordinates provided by Twitter®, however, contained four decimal places, causing the slightest change in location to still be identified when running a query in SQL Server. Thus, the coordinates of each tweet were parsed to one decimal place, enough to distinguish one large city from a neighboring large city. In a query, everything except the latitude and longitude was removed, and a new column called parsed_geo was created to store the revised coordinates.

A new table called Processed_GEO_Twitter_scrape was created containing columns TwID, UserID, Username, crawlingDT, Cont, CreateDT, Geo, and parsed_geo with the same characteristics as the corresponding columns in GEO_Twitter_Scrape. Only tweets with geolocation were selected from GEO_Twitter_Scrape and placed into Processed_GEO_Twitter_scrape (**Figure 1.5**). The separate latitude and longitude coordinates for each parsed_geo were placed in corresponding columns.

From Twitter_Scrape, the data for the initial map was extracted. This data would contain the three initial tweet scraping points (Rio de Janeiro, Miami, and New York), and the number of users in each of these three cohorts (**Figure 1.6**). The results of the query were exported to an Excel spreadsheet (**Figure 4**).

A new table, Ready4Map, was created, and the distinct tweets and their information were selected from Processed_GEO_Twitter_scrape to Ready4Map (**Figure 1.7**). On these maps, the locations of those who had some movement was required, so only the entries in Ready4Map where latitude or longitude changed were needed(**Figure 1.8**). Then, the latitude-longitude coordinates showing movement, the initial cohort the users came from, and the userid were selected, grouped by the latitude, longitude, SearchCity, and userid. From there, a new column called OriginalCity was created and the SearchCity was stored under OriginalCity. Along with SearchCity/OriginalCity, latitude and longitude were selected, and for each distinct latitude-longitude pairing, a counter was set up (**Figure 6**).

In order to identify the effectiveness of each keyword used, the tweets stored in Twitter_Scrape with geolocation were queried by the keyword that retrieved the tweets in the original code. The content of the tweets were read through, and the number of tweets in each region with each keyword that indicated an exposure to Zika was noted.

Once cohorts were identified for each of Rio de Janeiro, Miami, and New York, the ScribbleMaps API was used in order to visually represent the location of Twitter® users that were followed. First, the Twitter® search parameters were used in order to create radii of 100 miles around the coordinates of Miami and New York and a radius of 500 miles around the coordinates of Rio (**Figure 7**). These radii were large due to all three cities being coastal, which required a greater radius but meant that a lot of the area was the ocean. Each population was overlaid with the number of users starting out in that location and then mapped for the initial time point (**Figure 7**). Then, users were monitored and the truncated geocoordinates (**Figure 6**) were used to create a separate map for each starting city for a total of three overall (**Figures 9-11**). In order to increase the specificity of regional variations in users' movement, several clusters of geocoordinates were also mapped individually to highlight areas of large density.

These four maps were used to analyze worldwide trends in human movement and then compared to published data on global populations of *Aedes aegypti* and *Aedes albopictus* mosquitoes (**Figure 8**) (Kraemer, 2015) in order to determine places with an increased chance of future outbreaks. Finally, these locations were analyzed relative to Zika data published by agencies such as the CDC and Boston Children's Hospital (healthmaps.org). Major outbreaks reported by healthmaps.org were found by searching for the Zika virus and each region's earliest major outbreak was used as a comparison. These comparisons were then used to determine the predictive capacity of our model and confirm it's accuracy.

Results and Discussion

The initial volume of tweets relating to the search parameters resulted in 38,810 tweets from 38,567 different users, which were then filtered in order to find 281 different users to be monitored for the duration of the study. Once cohorts were determined and the truncated geocoordinates were mapped, an initial-state map was created, as presented in **Figure 7**. The map in this figure shows the radius around each city where tweets were initially collected (**Figure 6**) and is overlaid with the number of users originating within those boundaries that were then tracked for the remainder of the study. Published data on *Aedes aegypti* and *Aedes albopictus* populations was also analyzed, such as the map in **Figure 8**, which showed that the three aforementioned regions from which users originate all lie in close proximity to Zika vectors.

The collected tweets and maps of movement showed that a large percentage of users originally in Miami either stayed in the city or moved within the continental United States. However, some groups within this cohort showed movements to Europe, and more specifically, towards Italy and Portugal, both of which lie in zones with a large number of *Ae. aegypti* and *Ae. albopictus*, as demonstrated by **Figures 9 and 12**. These results highlight the potential vulnerability cities in the southern Iberian peninsula and Italy may have with people returning from this Zika-infected region.

Movement of users whose initial location was New York was mostly concentrated to Boston and coastal cities in Virginia, as seen in **Figure 10b**. According to the map in **Figure 8**, these areas susceptible to a Zika outbreak because of their mosquito populations. Some New York users also moved to Brazil, where a large number of Zika outbreaks has occurred.

Data from users starting in Brazil, as depicted in **Figure 11**, showed a much larger spread in final location, which was most likely due to the Summer Olympics. In particular, a number of users moved to countries in South and East Asia. **Figure 8** showed that these regions, especially parts of Thailand, the Philippines, and South Korea, have a high *Ae. aegypti* and *Ae. albopictus* density, and thus may be particularly susceptible to future outbreaks, as seen in **Figure 12**. Furthermore, users starting out in the Rio de Janeiro cohort also travelled to Greece and the East Coast of the U.S., locations that were identified in other cohorts and contain nearby mosquito populations.

Using this data, a list of cities and regions with a high potential for Zika outbreak, as specified by both the number of users and mosquito population density, can be compiled (**Figure 12**). These regions, which includes Thailand, South Korea, the Philippines, Spain, Portugal, and Italy, are all regions where the model predicts a possible spread of Zika.

When these cities were compared to more conventional agencies in charge of tracking outbreaks, it was found that the locations predicted by the model far in advance ended up being locations of eventual Zika outbreaks (**Figure 13**). In particular, the first report on new Zika cases in Thailand, as indicated by healthmap.org, a disease-outbreak monitoring service created by Boston Children's Hospital, occurred on August 30, 2016 (HealthMap, 2016). In South Korea, the outbreak was reported on August 20, 2016 (HealthMap, 2016). A warning was issued and cases of Zika discovered in the Philippines between September 3 and 5, 2016 (HealthMap, 2016). Finally, healthmaps.org reported the first indication of Zika cases in Spain on August 23, 2016 (HealthMap, 2016). All of these dates occurred after the majority of the data was collected and the model created by our program, within the span of two weeks, accurately predicted the aforementioned countries to have a possible Zika risk based on human movement, a prediction that has since been confirmed.

Finally, from the complete database of tweets, a random selection was taken and independent tweets were read in order to determine the percentage of tracked users who reported direct contact with Zika in the vicinity of their tweeted location (**Figure 4**). This data showed that about 51% of tweets were directly related to the Zika virus and could be used to assess users with, or in the vicinity of, Zika. Furthermore, this analysis also revealed febre, bug spray, médico, and pregnant were keywords that could be used to track the Zika virus with an average of 63% or greater confidence.

Illustrations

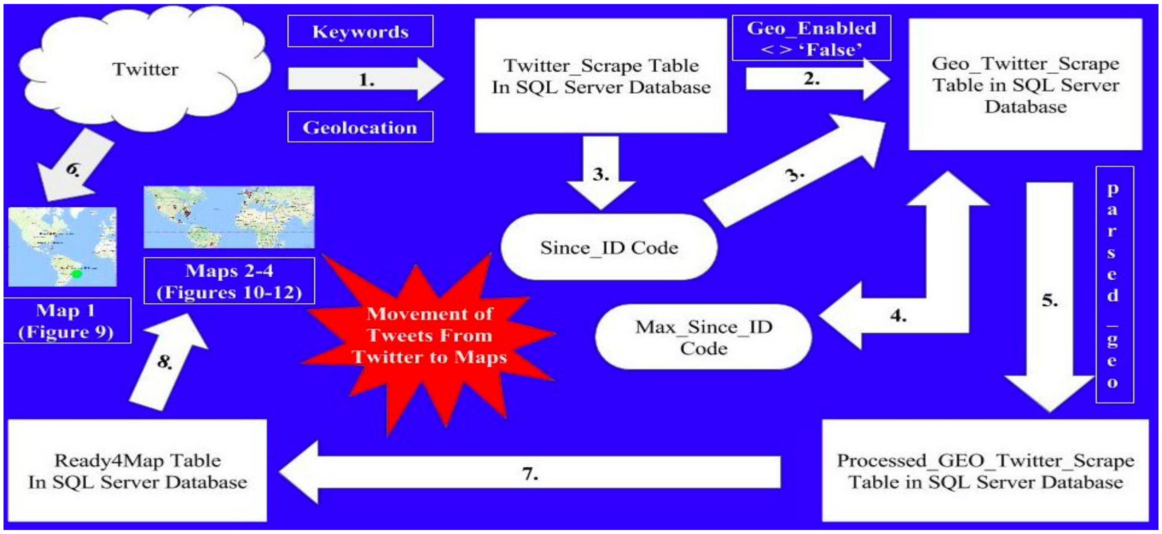

Twitter
Keywords
1.
Geolocation
Twitter_Scrape Table In SQL Server Database
Geo_Enabled < > 'False'
2.
Geo_Twitter_Scrape Table in SQL Server Database
6.
3.
3.
parsed_geo
Map 1 (Figure 9)
Maps 2-4 (Figures 10-12)
Since_ID Code
Movement of Tweets From Twitter to Maps
Max_Since_ID Code
4.
5.
8.
Ready4Map Table In SQL Server Database
7.
Processed_GEO_Twitter_Scrape Table in SQL Server Database

Figure 1. Flowchart of Procedure from Twitter® to Maps

1.1 By implementing for-loops and lists, every keyword was searched for in each of the three cities. Using since = ‘’ and until = ‘’ functions restricted the date on which tweets were created, and these functions were run everyday, starting from August 10, 2016. A connection was opened connecting the code and the results returned, and a cursor was created to insert each tweet and its information into a new row in the database. For full code, visit https://github.com/HLAP2016/Zika-Twitter-Scrape/blob/master/OAuth2AccessTwitterScrape.py**.**

1.2 A new table, GEO_Twitter_Scrape was created by selecting all entries into GEO_Twitter_Scrape from Twitter_Scrape where an impossible statement, in this case 1 = 2, was true. The columns GEOenable, SearchKey, SearchGEO, SearchCITY, Parsed_GEO, latitude, and longitude were deleted from GEO_Twitter_Scrape. All tweets where Geo_Enabled did not equal, or ‘< >’, ‘False’ was selected from Twitter_Scrape and duplicated into GEO_Twitter_Scrape.

1.3 A SQLCommand was written to select distinct TwID and UserID from Twitter_Scrape where geo ‘< >’, or did not equal, ‘None’. The first cursor executed the SQL Command, and with a while loop, grabbed the TwID and UserID of each of the desired tweet entries. The since_id command and insert statement were written, a second cursor then carried out the insert string, a second connection committed the new entries, and the process was repeated in a for loop. For full code, visit https://github.com/HLAP2016/Zika-Twitter-Scrape/blob/master/Since_ID.py.

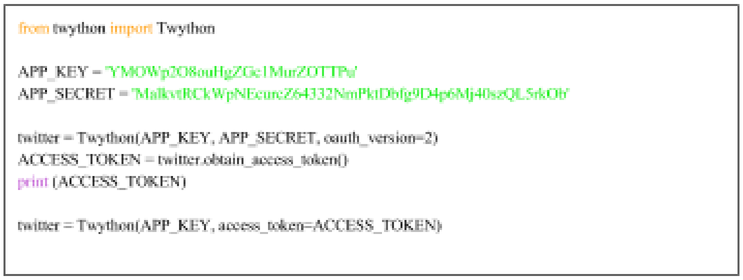
1.4 The code for Max_Since_ID was the same as that for Since_ID, except that the max tweetID for each user was selected and run as the since_id when scraping. Max_ID was used because the greater the TwID for each user’s tweets, the more recent the tweets would be. Because it is possible for users to delete their tweets, if an exception occurred, “exception error” was printed, and the code continued past the exception. Further searches of subsequent tweets from each user ran Max_Since_ID For full code, visit https://github.com/HLAP2016/Zika-Twitter-Scrape/blob/master/Max_Since_ID.py.

1.5 Everything in GEO_Twitter_Scrape was selected, and a substring was run removing everything up to and including the square brackets from the geo information of each tweet, which was in the format of: “{ coordinates : [latitude, longitude], type : Point }. The parsed coordinates were stored in a new column created called Parsed_geo. A new table, Processed_GEO_Twitter_scrape, was created containing columns TwID, UserID, Username, crawlingDT, Cont, CreateDT, and Geo, and parsed_geo, and the tweets with geolocation in GEO_Twitter_Scrape were placed into Processed_GEO_Twitt\er_scrape. The separate latitude and longitude coordinates for each parsed_geo were placed in separate columns.

1.6 From Twitter_Scrape, the number of different users from each SearchCity with geocoordinates was counted and grouped by SearchCity, displaying the results shown in **Figure 4**.

1.7 All the tweets in Processed_GEO_Twitter_scrape were transferred to a new table, Ready4Map.

1.8 Only tweets of the same user where the new latitude or longitude did not equal the original latitude or longitude, respectively, in Twitter_Scrape were selected, along with the initial cohort the users came from and the userid. For each distinct latitude-longitude pairing, a counter was set up, and the count for each location was stored under a new column called VisitNum. See **Figure 6** for a sample of the data queried.



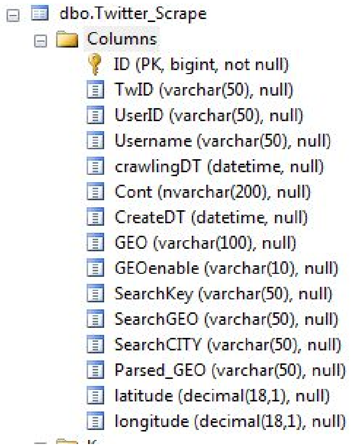
```
from twython import Twython

APP_KEY = 'YMOWp2O8ouHgZGc1MurZOTTPu'
APP_SECRET = 'MalkvtRCkWpNEcunZ64332NmPktDbfg9D4p6Mj40szQL5rkOb'

twitter = Twython(APP_KEY, APP_SECRET, oauth_version=2)
ACCESS_TOKEN = twitter.obtain_access_token()
print (ACCESS_TOKEN)

twitter = Twython(APP_KEY, access_token=ACCESS_TOKEN)
```

Figure 2. Code for Setting up Access Token This code was used to obtain the access token by implementing the Consumer Key and Consumer Secret, which was given with the creation of a Twitter® application. Twython was imported, and the Access Token was printed.



```
dbo.Twitter_Scrape
  Columns
    ID (PK, bigint, not null)
    TwID (varchar(50), null)
    UserID (varchar(50), null)
    Username (varchar(50), null)
    crawlingDT (datetime, null)
    Cont (nvarchar(200), null)
    CreateDT (datetime, null)
    GEO (varchar(100), null)
    GEOenable (varchar(10), null)
    SearchKey (varchar(50), null)
    SearchGEO (varchar(50), null)
    SearchCITY (varchar(50), null)
    Parsed_GEO (varchar(50), null)
    latitude (decimal(18,1), null)
    longitude (decimal(18,1), null)
```

Figure 3. Columns and Descriptions of Columns for Twitter_Scrape Table in SQL Server Database The ID column was created as a primary key so that each entry had a unique ID in the table and could not be null, ensuring that every entry would have an ID. The TwID column corresponded to the Tweet ID provided by Twitter®, which was unique to each tweet, and could store up to 50 integers or characters. The UserID column, which corresponded to the unique ID assigned to every Twitter® user, had the same data type as the TwID column. The Cont column stored the content/text in each tweet and could store up to 200 characters of UNICODE and/or multilingual data to be collected per entry. The limit of 200 characters was chosen because Twitter® has a limit of 140 characters per tweet. The columns SearchKey, SearchGEO, and SearchCity stored the keyword, the coordinates of the city that was being searched for when the tweet was scraped, and the city that was being searched for when the tweet was scraped, respectively. The column Parsed_GEO was filled to contain the coordinates corresponding to each tweet to the tenths decimal place. The columns latitude and longitude, which corresponded to the separate coordinates in Parsed_GEO, could contain decimals with up to a length of 18 and one decimal place.

Search City	Number of Users in SearchCity Cohort
Miami	111
New York City	51
Rio de Janeiro	119

Figure 4. Number of Users in Three Cohorts. By running the query in Figure 6b, the number of distinct users from each city that had tweets containing geocoordinates were counted, and the data was exported into an Excel spreadsheet. This data was used to create the first map displayed in **Figure 7**.

English		Portuguese	
pregnant 36% confirmed pregnancy, 67% confirmed in NYC, 100% confirmed in Rio	**joint pain** 0% confirmed pain	**grávida** 11% confirmed pregnancy, 100% confirmed in Rio	**dor nas articulações** 0 results
eye pain 0% confirmed eye pain	**doctor** 5% confirmed visits, 4% in Miami, 15% in Rio	**dor nos olhos** 25% confirmed eye pain, 100% confirmed in NYC	**médico** 62% confirmed visits, 64% in Rio
vomit 0% confirmed vomiting	**doctor appointment** 0% confirmed visits	**vomitar** 0% confirmed	**consulta médica** 0% confirmed visits
muscle pain 0 results	**mosquito bite** 100% confirmed bites 100% in NYC	**dor muscular** 0% relevant	**picada de mosquito** 0 results
mosquito 50% confirmed exposure, 12.5% confirmed bites, 18% confirmed in Miami, 100% confirmed in NYC, 67% confirmed in Rio	**bug repellent** 0 results	**mosquito** 50% confirmed exposure, 12.5% confirmed bites, 18% confirmed in Miami, 100% confirmed in NYC, 67% confirmed in Rio	**repelente de insetos** 0 results
Zika 15% confirmed exposure, 27% confirmed in Miami	**bug repellant** 0 results	**Zika** 15% confirmed exposure, 27% confirmed in Miami	**repelente de insetos** 0 results
feel sick 20% confirmed sickness, 25% in NYC	**mosquito repellent** 0 results	**doente** 36% confirmed sickness, 100% in NYC, 33% in Rio	**repelente de mosquito** 0 results
fever 10% confirmed fever, 8% in Miami, 15% in NYC, 7% in Rio	**mosquito repellant** 0 results	**febre** 100% confirmed fever, 100% in NYC, 100% in Rio	**repelente de mosquito** 0 results
rash 40% confirmed rash, 100% in NYC	**bug spray** 50% confirmed exposure, 20% in Miami, 100% in NYC, 100% in Rio	**erupção** 0% confirmed rash	**repelente de insetos** 0 results

Figure 5. Keywords and their Effectiveness. Keywords were chosen based on the likelihood that their usage would demonstrate an exposure to or a person having Zika. The keywords were scraped for in both English and Portuguese. Confirmed cases were based on possible exposure or relevant symptoms of Zika. The percentages of confirmed within a city was found by the number of confirmed tweets as compared to the total number of tweets from the city.

OriginalCity	latitude	longitude	VisitNum
Miami	-34.3	150.9	1
Miami	-23.6	-46.7	1
Miami	-23.5	-46.7	1
Miami	-22.5	-49	1

Figure 6. Sample of Results from Query in Figure 1.8.
Data was sorted by Original City, and the latitude-longitude coordinates were those that differed from the original coordinates of the tweet scraped and stored in Twitter_Scrape. The VisitNum counted the number of different users who travelled to each distinct latitude-longitude pairing.

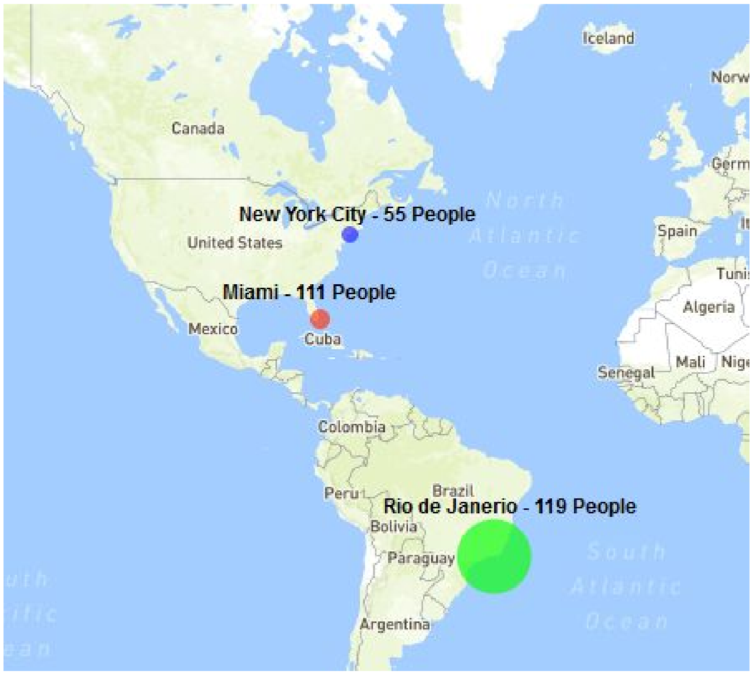


Figure 7. Initial Location of Twitter® Users
Data from Figure 4, the number of users in each cohort, were mapped on each of the three studied cities with radii of 100 miles (Miami and New York) and 500 miles (Rio de Janeiro). The map shows the initial location of users in each of the three cohorts.

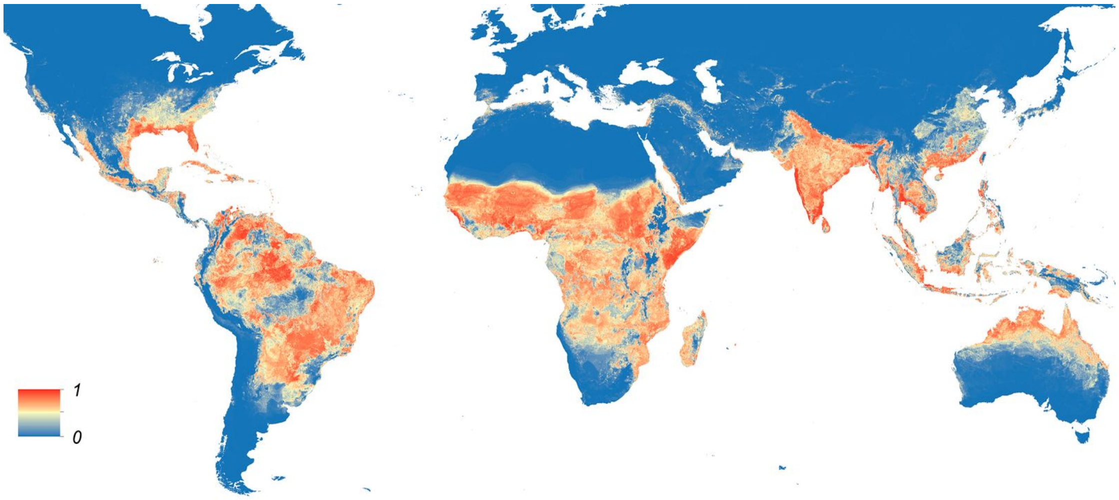


Figure 8. Map of *Ae. aegypti and Ae. albopictus* Population
A map showing the global distribution of *Ae. aegypti and Ae. albopictus,* the mosquito vectors that spread the Zika virus. The probability of occurrence is depicted from blue (0) to red (1) for the mosquito populations. (Kraemer, 2015)

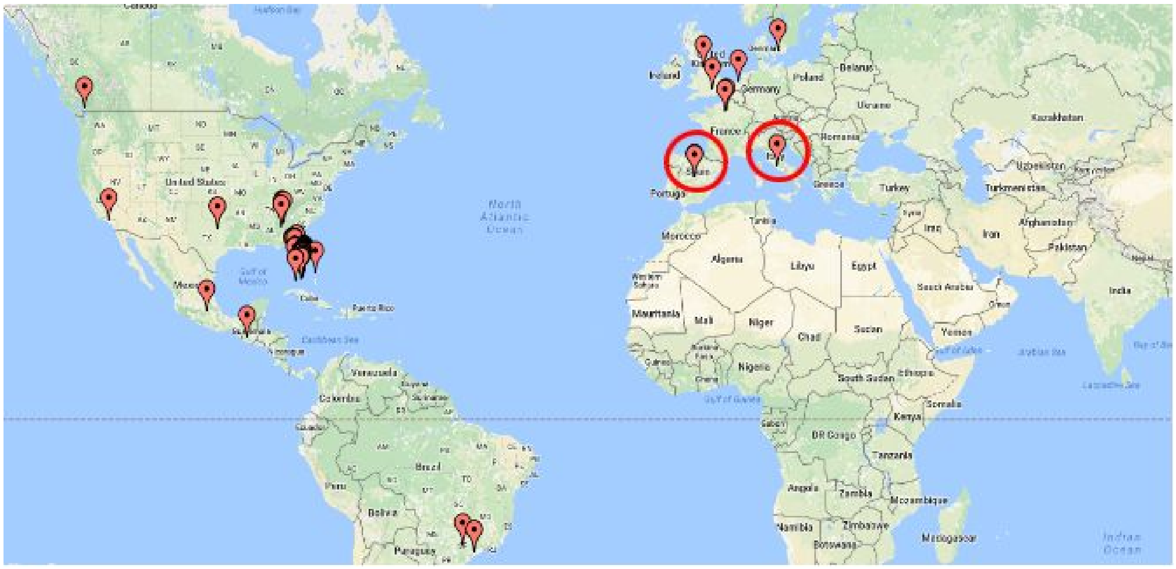


Figure 9. Global Spread Map of Miami Cohort
A global map showing the final location of all users in the Miami cohort whose initial tweet location was within a 100 mile radius of the city (**Figure 7**). Areas of interest include movement into Spain and Italy (circled in red).

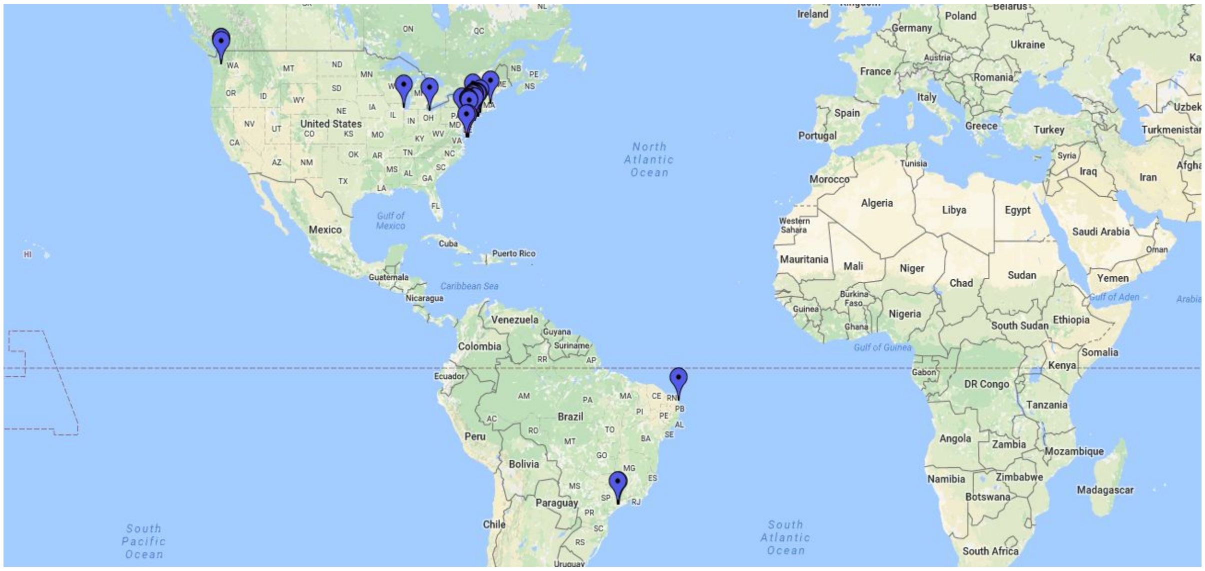
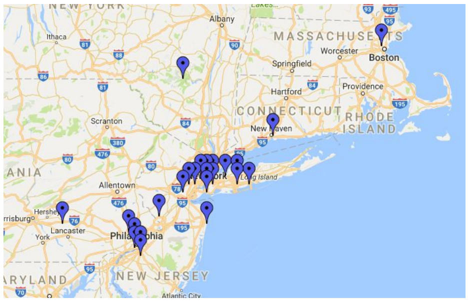


Figure 10a. Global Spread Map of New York Cohort
A global map showing the final location of all users in the New York cohort whose initial tweet location was within a 100 mile radius of the city (**Figure 7**).

Figure 10b. Areas of Interest after Movement from New York Cohort
A popout of **Figure 10a** focusing on movement into Northeastern United States .



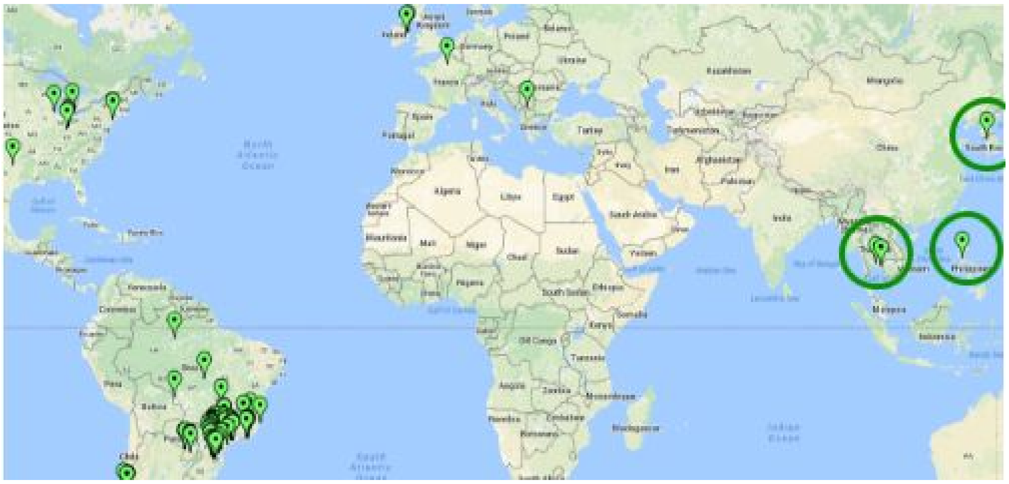


Figure 11. Global Spread Map of Rio de Janeiro Cohort
A global map showing the final location of all users in the Rio de Janeiro cohort whose initial tweet location was within a 100 mile radius of the city (**Figure 7**). Areas of interest include movement into Thailand, the Philippines, and South Korea (circled in green).

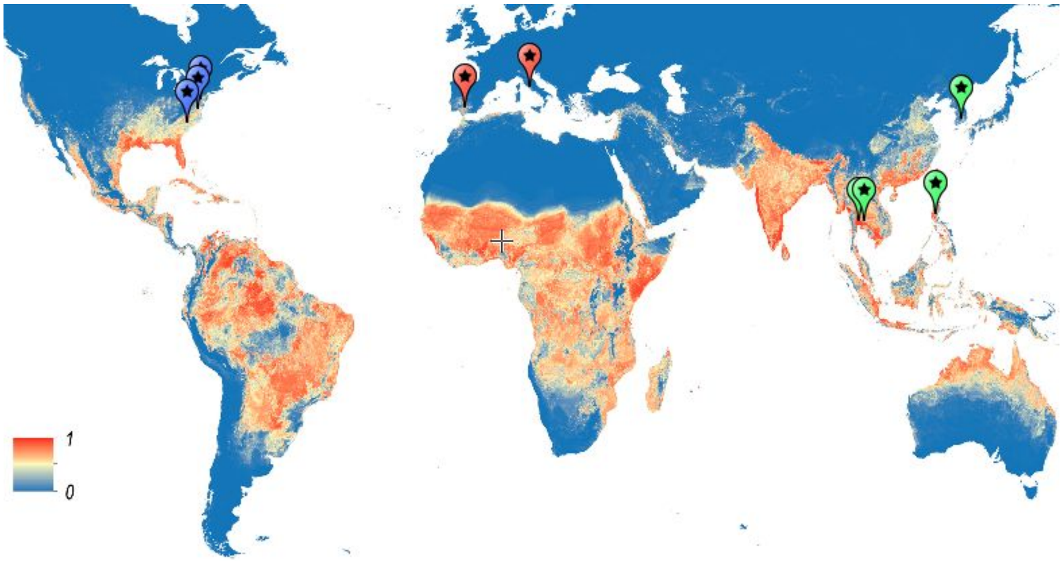


Figure 12. Cities of Interest as Determined Through All Three Cohorts
A map that shows all previously highlighted areas of future interest with regards to a potential Zika outbreak. Each color refers to the cohort from where the movement occurred.

	Date Collected from Twitter-Scraping Model	Date Collected on healthmaps.org
Spain	August 21, 2016	August 23, 2016
Thailand	August 21, 2016	August 30, 2016
South Korea	August 17, 2016	August 20, 2016
The Philippines	August 17, 2016	September 3, 2016

Figure 13. Prediction Efficacy Between the Model and Healthmaps.org The table compares the dates where movement to the at-risk country was detected through the model created using Twitter® data, and the dates where a Zika outbreak occurred within the country. The data shows that the model consistently pinpointed at-risk cities before outbreaks were reported.

Conclusions and Future Work

This project demonstrated that Twitter® can be used to predict the movement of Zika. By implementing Twython, a wrapper for Twitter®, and Python, and storing the scraped tweets in a SQL Server database, 38,810 distinct tweets were collected from 38,567 different users, with 281 different users with geo enabled. Then, those 281 users were monitored, and their following tweets and geolocations were also collected. The resulting 14,906 tweets were mapped using ScribbleMaps API, and from the cohort maps created, it was evident that the model accurately predicted recent outbreaks, especially in Southeast Asia, where recently, Singapore had a major outbreak of Zika. Furthermore, this model has many possible applications because other Flaviviruses with similar transmission methods, such as yellow fever, West Nile, chikungunya, and dengue, be mapped and predicted. Organizations, like the Center for Disease Control and the World Health Organization, and local governments can then use these predictions to focus on preventing transmission.

This goal of this project was to test whether or not a model based on human movement through social media could effectively follow and predict the spread of Zika. This model was built on a small-scale version, with only tweets from three major cities with a significant number of cases of Zika: Rio de Janeiro, Miami, and New York City. Furthermore, the maximum number of tweets that can be collected in one call was 200, a limit set by Twitter®, which in turn limited the scope and accuracy of the model as well. If multiple people were running the Twitter® scraping code on different computers with different Access Tokens, then more tweets could be grabbed with the same keyword, date created, and geolocation. Despite these limits, the model still proved to be accurate as to where users, and thus Zika, were moving, as seen by the appearance of users from Rio de Janeiro in South Korea, Thailand, and the Philippines, and users from Miami in Spain. In the future, this model could be expanded to search for more tweets globally, thus improving the its accuracy.

While the model was easy to use because the data moved in a step-wise manner, from Twitter® through Twython to SQL Server database tables, and then analysis, the structure of the model as it is would make it more difficult to analyze the information collected while continuously collecting Twitter® data. Additionally, streamlining the Twitter® scraping process would make the model easier to use, especially with greater amounts of data.

The model ran starting from the beginning of August because it was expected that many people would move to and from Rio de Janeiro, due to the Summer Olympic Games, but a further and more in-depth analysis could be performed with longer periods of data collection.

Further research and testing of whether other social media platforms, such as Facebook, could be implemented (separately or along with Twitter® data) to track the spread of Zika and other Flaviviruses would be interesting to explore. Another possible study that could be conducted is whether or not other types of diseases or viruses, such as those that are airborne, could be studied through social media.

The model, despite its limitations in scale, showed the ability to predict possible areas of future outbreaks weeks in advance of the news agencies traditionally monitored. Furthermore, many of the cities predicted by the model ended up with significant outbreaks in the coming days and weeks. This study, therefore, shows a novel method to track disease spread and can be used to provide early warning and preparation time for countries around the world during crises such as the Zika outbreak.

References

Batrinca, B. & Treleaven, P.C. Social media analytics: a survey of techniques, tools and platforms AI & Soc (2015) 30: 89. doi:10.1007/s00146-014-0549-4

CDC, Case Counts in the US. (2016, September 19). The Center for Disease Control. https://www.cdc.gov/zika/geo/united-states.html

CDC, Zika Virus Prevention. (2015, June 1). The Center for Disease Control. https://www.cdc.gov/zika/prevention/

CDC, Zika Virus Transmission and Risks. (2016, August 27). The Center for Disease Control. http://www.cdc.gov/zika/transmission/

D'Ortenzio E, Matheron S, de Lamballerie X, Hubert B, Piorkowski G, Maquart M, Descamps D, Damond F, Yazdanpanah Y, Leparc-Goffart I. Evidence of sexual transmission of Zika virus. N Engl J Med. 2016;374(22):2195–8.

Dos Santos T, Rodriguez A. et al. (2016, August) Zika Virus and the Guillain–Barré Syndrome — Case Series from Seven Countries. New England Journal of Medicine. DOI: 10.1056/NEJMc1609015

Dredze M, García-Herranz M, Rutherford A, Mann G (2016, June 20). Twitter as a Source of Global Mobility Patterns for Social Good. arXiv:1606.06343

Haoyu Wang, Eduard Hovy, Mark Dredze. The Hurricane Sandy Twitter Corpus. *AAAI Workshop on the World Wide Web and Public Health Intelligence*, 2015 http://www.cs.jhu.edu/~mdredze/publications/aaai_w3phi_sandy.pdf

HealthMap, The Boston Children's Hospital. Zika Virus Cases. https://www.healthmap.org/en/

Johansson MA, Mier-Y-Teran-Romero L, Reefhuis J, Gilboa SM, Hills SL (2016) Zika and the risk of microcephaly. N Engl J Med. doi:10.1056/NEJMp1605367

Kraemer M, et al. (2015). The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. Elifesciences. http://dx.doi.org/10.7554/eLife.08347

Kuno, G., Chang, G.-J. J., Tsuchiya, K. R., Karabatsos, N., & Cropp, C. B. (1998). Phylogeny of the Genus Flavivirus. Journal of Virology, 72(1), 73–83.

Li, Hongda et al. (2016) Zika Virus Infects Neural Progenitors in the Adult Mouse Brain and Alters Proliferation. Cell, http://dx.doi.org/10.1016/j.stem.2016.08.005

Messina Jane P, et al. (2016). Mapping global environmental suitability for Zika virus. Elifesciences. http://dx.doi.org/10.7554/eLife.15272

Motta I, Bryan S, Suely C, et al. Evidence for Transmission of Zika Virus by Platelet Transfusion. N Engl J Med 2016; 375:1101-1103

Plourde AR, Bloch EM. A literature review of Zika virus. Emerg Infect Dis. 2016 Jul [September 2016]. http://dx.doi.org/10.3201/eid2207.151990

Tribune News Services (2016, March 3). Google donates $1M to help fight Zika virus spread. *The Baltimore Sun.* Retrieved from http://www.baltimoresun.com/health/ct-google-zika-virus-20160303-story.html

World Health Organization, WHO and experts prioritize vaccines, diagnostics and innovative vector control tools for Zika R&D. (2016, March 9). http://www.who.int/mediacentre/news/notes/2016/research-development-zika/en/