Table of Contents

1 introduction ... 3
1.1 Motivation ... 3
1.1.1 The importance of spatial social network ... 3
1.1.2 The status of spatial social network ... 3
1.1.3 Selection bias ... 4
1.1.4 State of the art ... 5
1.2 Task ... 5
1.2.1 The tool/app should be able to pull data from Twitter Streaming API and geolocated tweets on the map ... 5
1.2.2 The tool/app should be able to visualize spatial social network on the map ... 5
1.2.3 The tool/app should be able to geolocated users and visualization on the map 6
1.2.4 The tool/app should be able to sampling user location and visualization on the map ... 6
1.2.5 The tool/app should be able to make some centrality measurements based on spatial social network ... 7
1.2.6 The tool/app should be able to analyze the user's movement and visualization on the map ... 7
1.2.7 The tool/app should be able to compare user’s movements in different time ranges and visualization on the map ... 7
2 Basics ... 9
2.1 What is bounding_box in twitter? ... 9
2.2 The status of twitter and the problem of implementation ... 12
2.3 Define the user's movement based on this tool/app ... 15
3 Approach ... 17
3.1 Requirements analysis ... 17
3.2 Implementation Tools ... 18
3.2.1 Introduction to R ... 18
3.2.2 Shiny ... 19
3.2.3 Leaflet ... 20
3.2.4 Some statistics function ... 20
3.3 Structure of the work ... 22
4 Implementation ... 23
4.1 The tool/app should be able to pull data from Twitter Streaming API and geolocated tweets on the map ... 23
4.2 The tool/app should be able to visualize spatial social network on the map ... 27
4.3 The tool/app should be able to geolocated users and visualization on the map ... 29
4.4 The tool/app should be able to sample user location and visualization on the map. 32
4.5 The tool/app should be able to make some centrality measurements based on sample spatial social network ... 36
4.6 The tool/app should be able to analyze the user's movemensts and visualization on the map ... 37
4.7 The tool/app should be able to compare user’s movements in different time ranges and visualization on the map ... 50
5 Evaluation ... 51
5.1 Evaluation of program outcome based on specific data ... 51
6 Summary and Outlook ... 54

6.1 Advantage of the application.....54
6.2 The use scenarios of application.....54
References.....68

1 introduction

1.1 Motivation

1.1.1 The importance of spatial social network

Spatial social networks are a type of social network that includes the location of nodes on a spatial plane. Currently, there is a desperate need to develop a tool or app to correctly analyze spatial social networks and visualize this particular kind of data. Social media is a significant way to understand human activity. There are currently an abundance of sites to support the user to discover their own social situation, for example, how many followers you have, personal information about the user, and the user's prestige and position within the social media bubble. Because of their ubiquity, these platforms can help the researcher to better understand and analyze global events. eg: the U.S. election, earthquakes, and diffusion of disease. People use social media every day to share their experience with others and discuss events that they are interested in interest in [1].

Social media is extremely powerful, one particularly powerful aspect of social media is that it can gather location information when a post is uploaded. When the information posted is equipped with geotagged, can we not only analyze the social relationship among the tweets but also analyze the relationship between geography relationship and social relationship.

Geolocated social media data provides an influential amount of information concerning the place and regional human behavior. Since Twitter can provide a continuously-updated stream of data, the data can be used to predict human behaviors and modeling populations. Aligning data with the particular geolocation from which it originated creates a powerful tool for modeling geographic phenomena, such as tracking the flu, predicting elections, or observing linguistic differences between groups[2].

1.1.2 The status of spatial social network

There is a range of tool and app support for social network analysis and visualization,

however, a significant amount of them will not support visualization spatial social network relevant function

. function tool	Locate nodes(tweets or users) on a spatial social network	Sampling nodes on a spatial social network	Comparing spatial social network over time	Track user
ucinet	no	no	no	no
gephi	no	no	no	no
visone	no	no	no	no
GeoSocialApp	no	no	no	no

Table 1.1 Disadvantages of existing applications

1.1.3 Selection bias

Selection bias is presented through the selection of individuals, groups or data for analysis, in such a way that correct randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed[4]. This is often referred to as the selection effect. The phrase "selection bias" most frequently refers to the distortion of statistical analysis, as a result of the method used to collect samples. If the selection bias is not taken into account, then a number of conclusions of the study may not be entirely accurate[3].

Twitter is a great source of research on social media, however, it too creates bias during data collection. If the goal of computational social science is to research and better understand society at scale, then the data that is studied must provide an accurate reflection of society[1]

Now, most of social network analysis websites don't have effective measures to reduce selection bias.

1.1.4 State of the art

1 Matthew J. Silk [5] wrote that there is no one best way to sub-sample social networks, you should choose your method according to your goal and situation to sub-sampling social networks. So our application should have different strategys according to the researcher’s goal to choose different sub-sampling.

2 Fred Morstratter et al. [1]. introduced a situation in which we have bias and how we could reduce that bias in social media.

3 Johnson et al. [7] identified four localness assessment techniques in the computing literature: nDays, Plurality, GeometricMedian, and LocationField. Also, Ankit Kariya et al. [8] formalize the approaches for localness assessment.

4 Matteo Manca et al. [9] Give some state of the art survey and case study about mining urban mobility patterns.

5 Antonio Lima and Mirco Musolesi [10] Give us some definition of spatial information dissemination metrics.

Therefore, in this particular sense, it is highly necessary to have a visualization application for spatial social networks and the application should have some function to decrease bias, resulting in the outcome from the application being more precise.

1.2 Task

1.2.1 The tool/app should be able to pull data from Twitter Streaming API and geolocated tweets on the map.

The tool/application should according to user’s purpose to pull data in a specific area and specific time fragment from Twitter Streaming API,then the tool/application should according to geotagged tweets geolocated tweets on the map.

1.2.2 The tool/app should be able to visualize spatial social network on the map

After geolocated tweets on the map, the tool/app should be able to mining relationship among the tweets. The relationship including mention, reply, and retweet. If any two tweets have the relationship, should have a connection between them on the map.
After connected the geolocated tweets according to the relationship, get a spatial social network.

1.2.3 The tool/app should be able to geolocated users and visualization on the map

Before discussing geolocated users, it should be made clear what is meant by ‘local’ and ‘Geolocated user’ . Local does not mean the home of the user, it is a concept depending on different situations. Ankit Kariya at al. identified three concepts: ‘local’ is where someone currently lives, where someone currently votes, and the places with which someone is very familiar. So ‘local’ depends on your purpose, if you want to know where someone is very familiar, you'd better use a strategy that can figure out where the user spends most of his time. This location is user’s ‘local’.
‘Geolocated user’ means using various techniques, according to the user’s location history to assign a user a localness.

As previously mentioned Johnson et al[24]. identified four localness assessment techniques: ndays, plurality, locationfield, and geometricmedian.

The next sections will explain three of the above-mentioned techniques and a new technique named ‘geometric center’.And the difference between my solution in the applicaiton and these four technique.

1.2.4 The tool/app should be able to sampling user location and visualization on the map

There is a certain amount of selection bias in data collecting, resulting in missing value (the amount of geotagged tweets is extremely low, the question is, how do we decrease this bias? Matthew stated[9] that it is better to sample data according to what you want. Therefore, if we prepare some sampling strategies, the user can select different strategies according to their requirements. The following are two

kinds of sampling strategies.

1.Random Sampling

According to the user's input number n to random select n samples.

2.Grid Sampling

According to the user's input number n and bounding box to dynamically generate a grid on the map, the grid should include an exact n point. I regard the tweets which have the minimum distance to the point as a sample, repeat this process n times, so at last, can get n sample tweets.

1.2.5 The tool/app should be able to make some centrality measurements based on spatial social network

Antonio Lima and Mirco Musolesi [10] Give us some definitions and formulas of spatial information dissemination metrics. In the application, I will according to two formulas provided by Antonio Lima and Mirco Musolesi to compute spatial social centrality.

1.2.6 The tool/app should be able to analyze the user's movement and visualization on the map

The tool/app should be able to track users at a localness and a specific time point or specific time fragment, and track users most recent location and determine the relationship between user movement and a specific area. If the user in a specific time fragment make some movements, the application should visualize the movements.

1.2.7 The tool/app should be able to compare user's movements in different time ranges and visualization on the map

If the researcher not only wants to know where the user is in specific time point or specific time fragment, they may also want to compare two-time points or two-time fragment of the user's movement. Therefore, the application should be able to

compare the user's movements in different time ranges.

2 Basics

2.1 What is bounding_box in twitter?

Twitter offical website stated that[11]:

“*bounding_box*

A series of longitude and latitude points, defining a box which will contain the Place entity oints are grouped into an array per bounding box. Bounding box arrays are wrapped in one additional array to be compatible with the polygon notation. Example:

"coordinates":
-74.026675,
40.683935

-74.026675,
40.877483

-73.910408,
40.877483

-73.910408,
40.3935

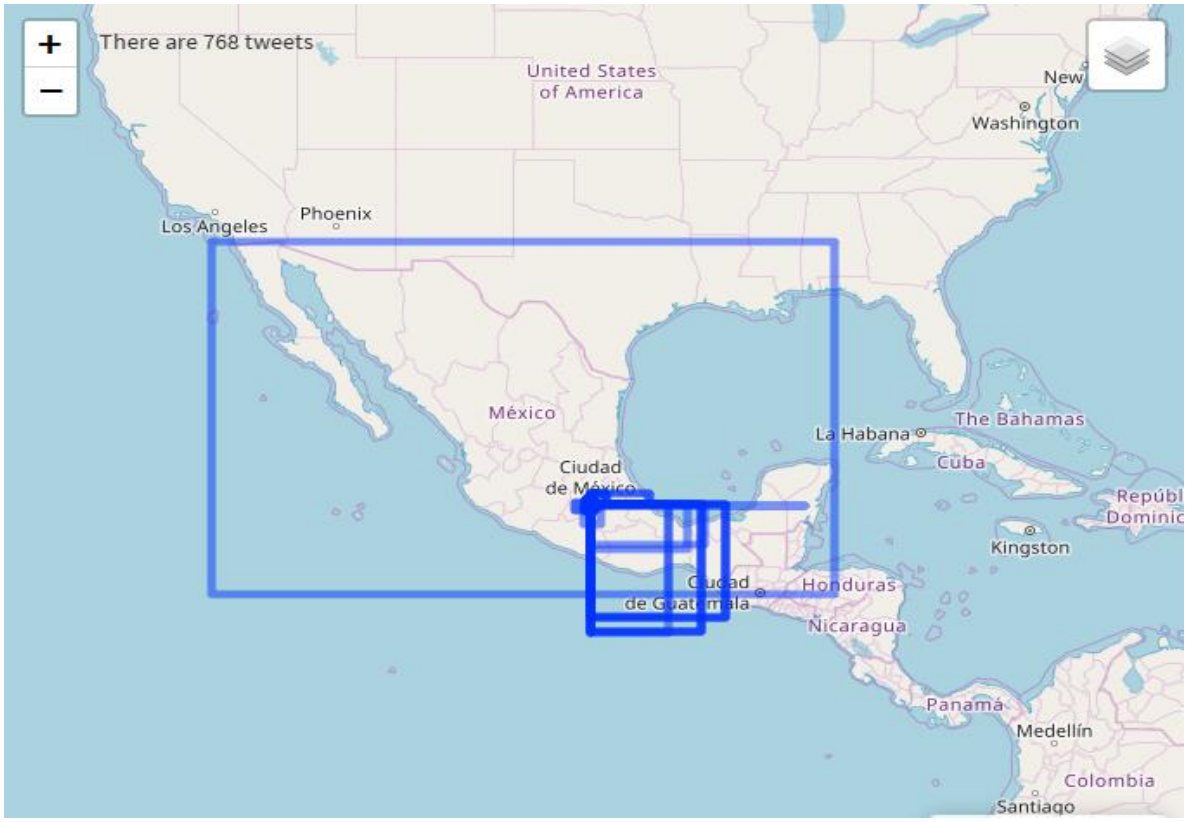
and there is a relevant concept named ‘place object’, twitter’s offical website wrote that*[11]:*

“*Place object*

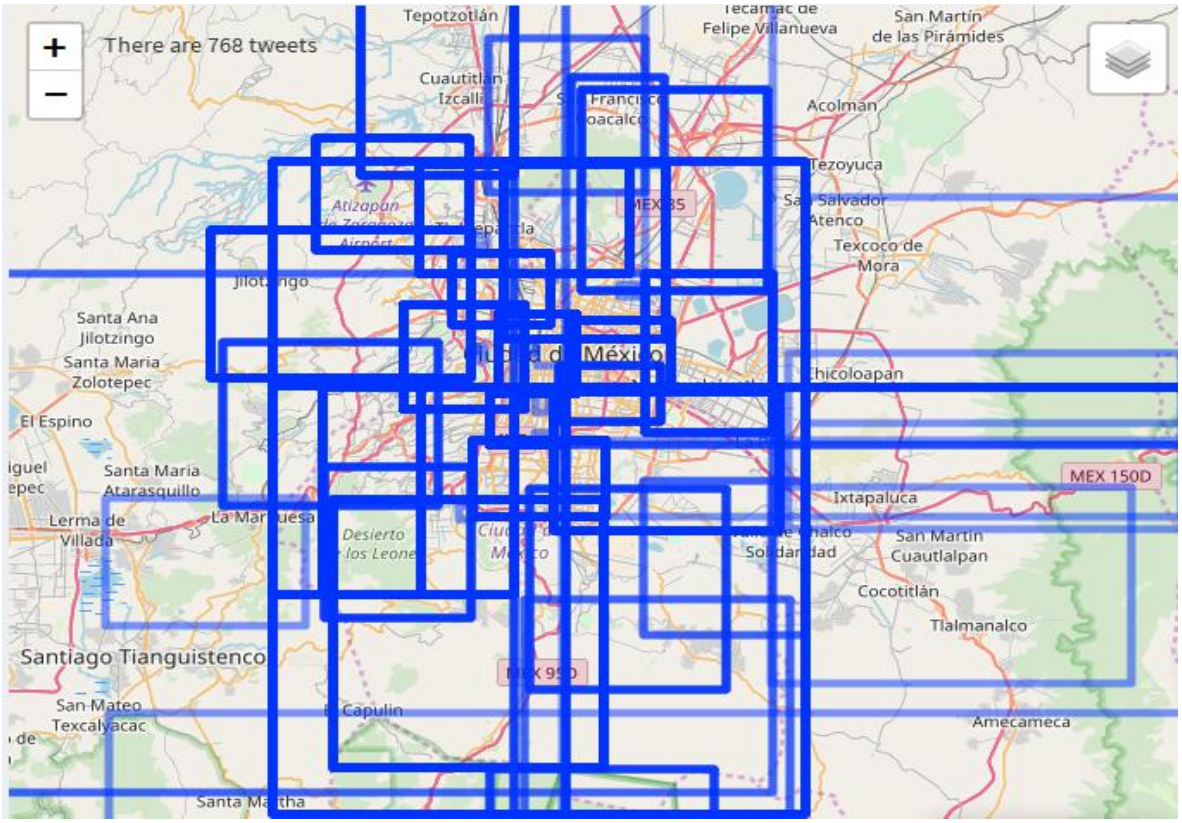
Places are specific, named locations with corresponding geo coordinates. When users decide to assign a location to their Tweet, they are presented with a list of candidate Twitter Places. When using the API to post a Tweet, a Twitter Place can be attached by specifying a place_id when posting the Tweet. Tweets associated with Places are not necessarily issued from that location but could also potentially be about that location.”

This is an explanation of what was discussed above from the twitter official website.

In general, when you want to post a tweet, you will write a certain amount of information and you can choose whether to place a label on your location, the label is enclosed in the bounding_box. For example, if you were in the library of Duisburg-Essen University and you want to post a tweet and the content is“The library is very nice”, and the place label you can choose as “Duisburg Deutschland” or “L-library uni”.



(a)



(b)

Figure 2.1.1 Different zoom level of bounding_box in Mexico City

I use leaflet packet addRectangles() function according to the bounding_box which has four pairs of coordinates plot polygon on the map of Mexico City as Fig 2.1.1 show. We can regard each blue polygon map as a bounding_box. Following those are what I summarize as some points about bounding_box, these points will help your understanding of what a bounding box is on twitter.

1. Different place labels map different bounding_box and same bounding_box map one place label.
2. bounding_box corresponding geographic areas are not the same size, they vary from small to large.
3. bounding_box corresponding geographic areas may overlap.
4. Which place label map which bounding_box and what size of bounding_box is predefined by twitter.

2.2 The status of twitter and the problem of implementation

After analyzing the tweets come from the Twitter Stream API, I found that not all tweets have specific coordinates, but because we use tweet2r function to collect tweets, so all the tweets at least have bounding_box coordinates.

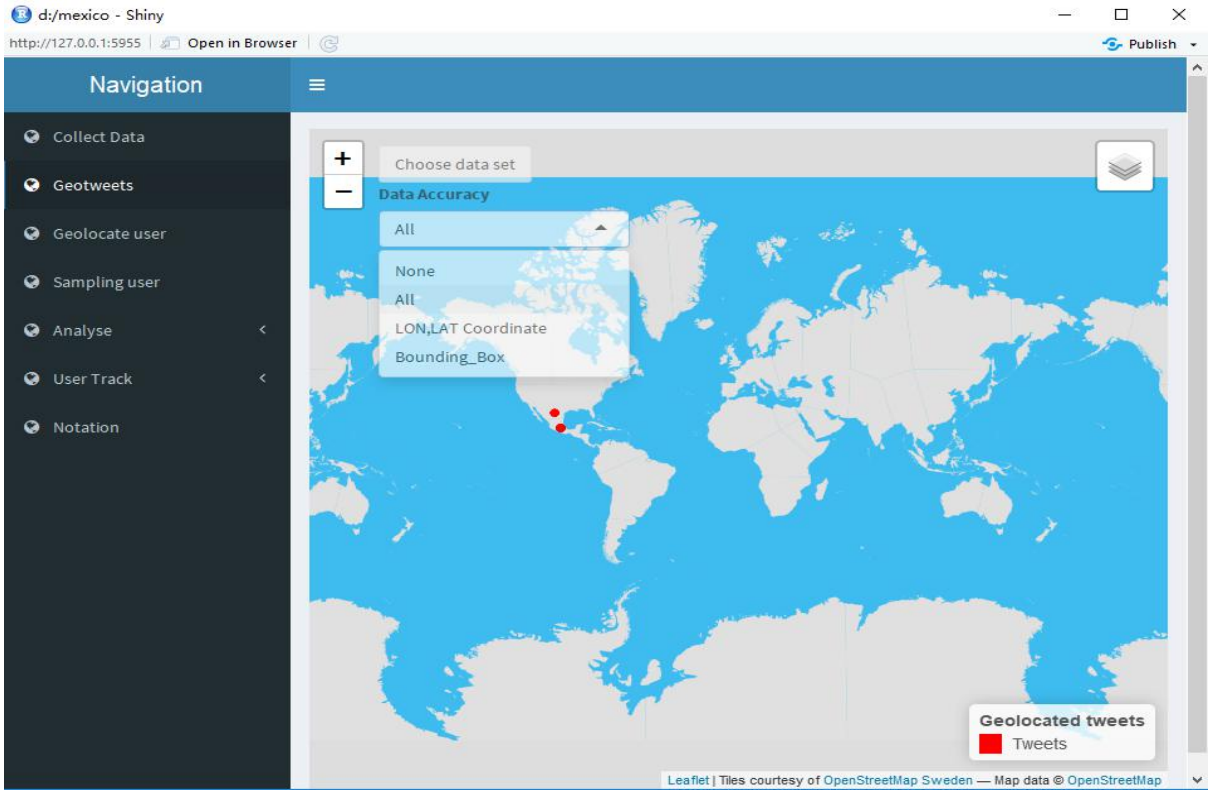
Problem

We wish to assign tweets to a position instead of a bounding box area. If not all tweets have a specific coordinate, how can we plot the tweets on map according to the specific coordinate.

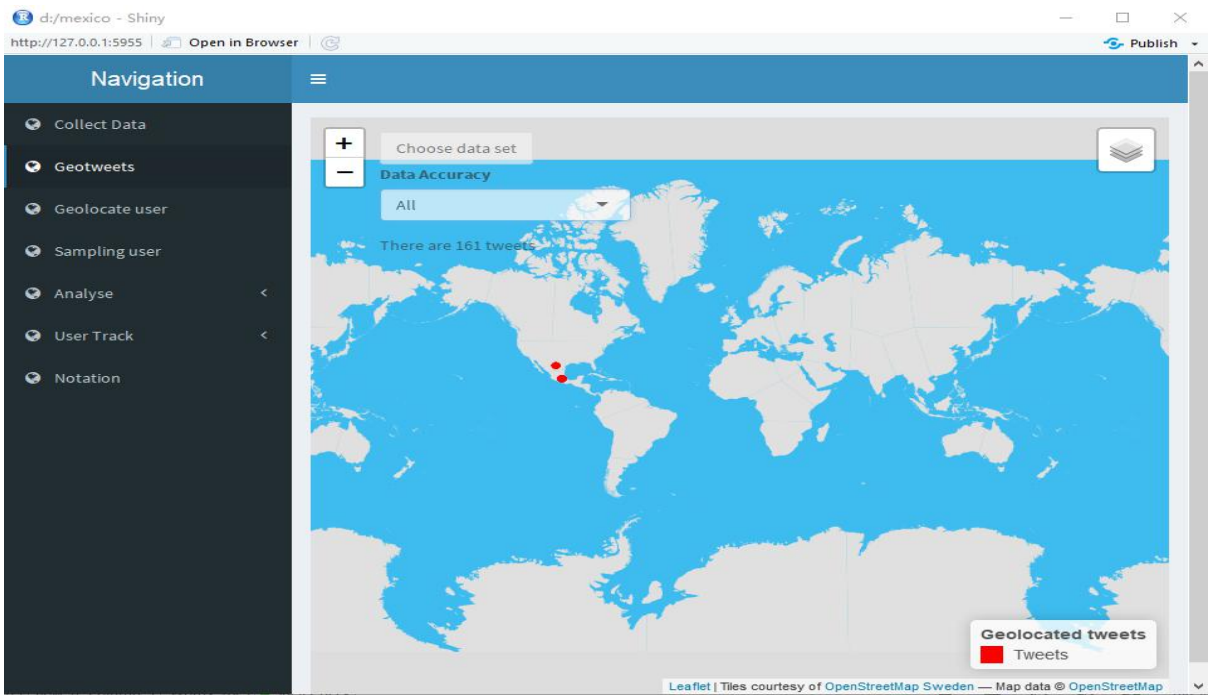
My solution

If the tweets have specific geographical coordinates, we geolocated tweets according to the specific geographical coordinate, if the tweets only have a bounding_box then we assume their positions are the geometric center point of the bounding_box and geolocated tweets. However, this will cause a problem. If a user posted some tweets in different locations but in the same bounding_box area, and these tweets only have bounding_box, after geolocated these tweets, would in the same location.

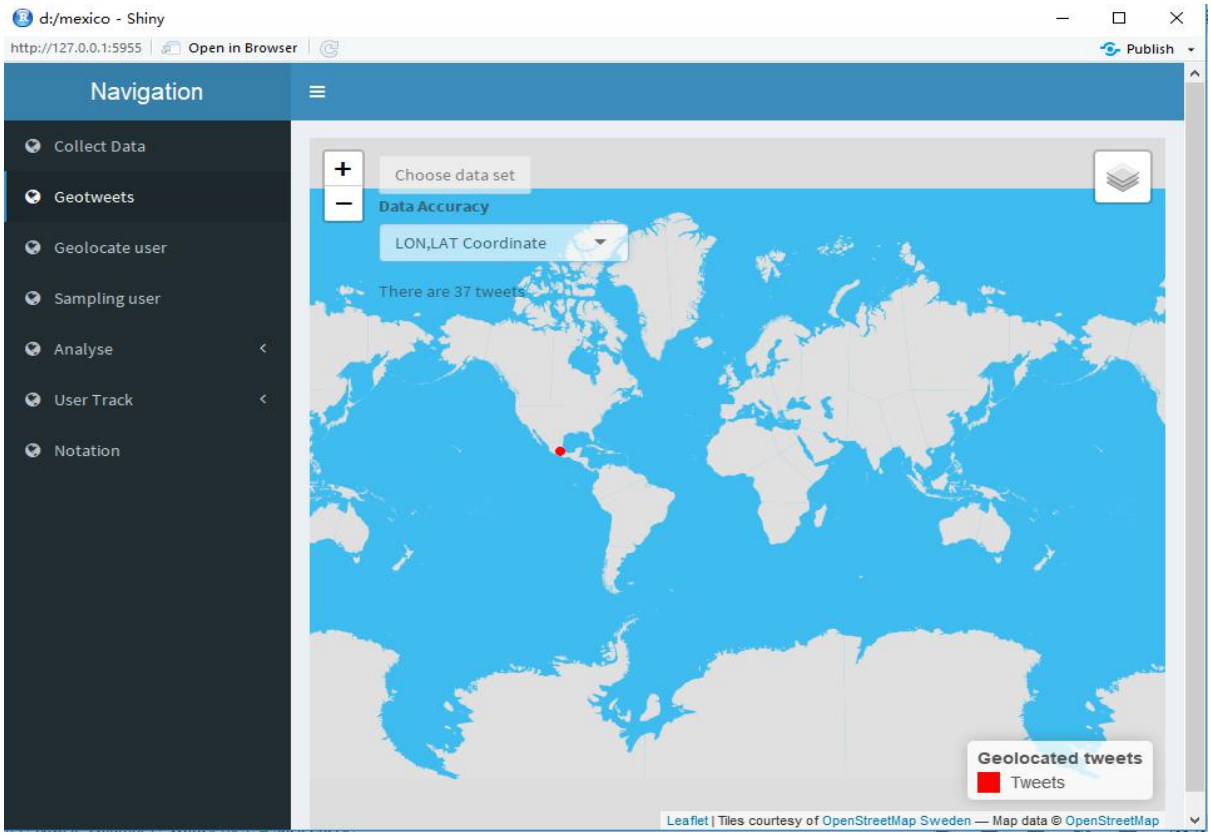
So, I divided these tweets data into three groups in Fig 2.2. They are LON LAT Coordinates group, Bounding_box group, and All group. Tweets in LON, LAT coordinates group are all only have Geographical coordinates. Tweets in bounding_box are all only had bounding_box. Tweets in all group have either Geographical coordinates or bounding_box. Users can make applications display different data groups according to their purpose.


d:/mexico - Shiny
http://127.0.0.1:5955
Open in Browser
Publish
Navigation
Collect Data
Geotweets
Geolocate user
Sampling user
Analyse
User Track
Notation
Choose data set
Data Accuracy
All
None
All
LON,LAT Coordinate
Bounding_Box
Geolocated tweets
Tweets
Leaflet | Tiles courtesy of OpenStreetMap Sweden — Map data © OpenStreetMap

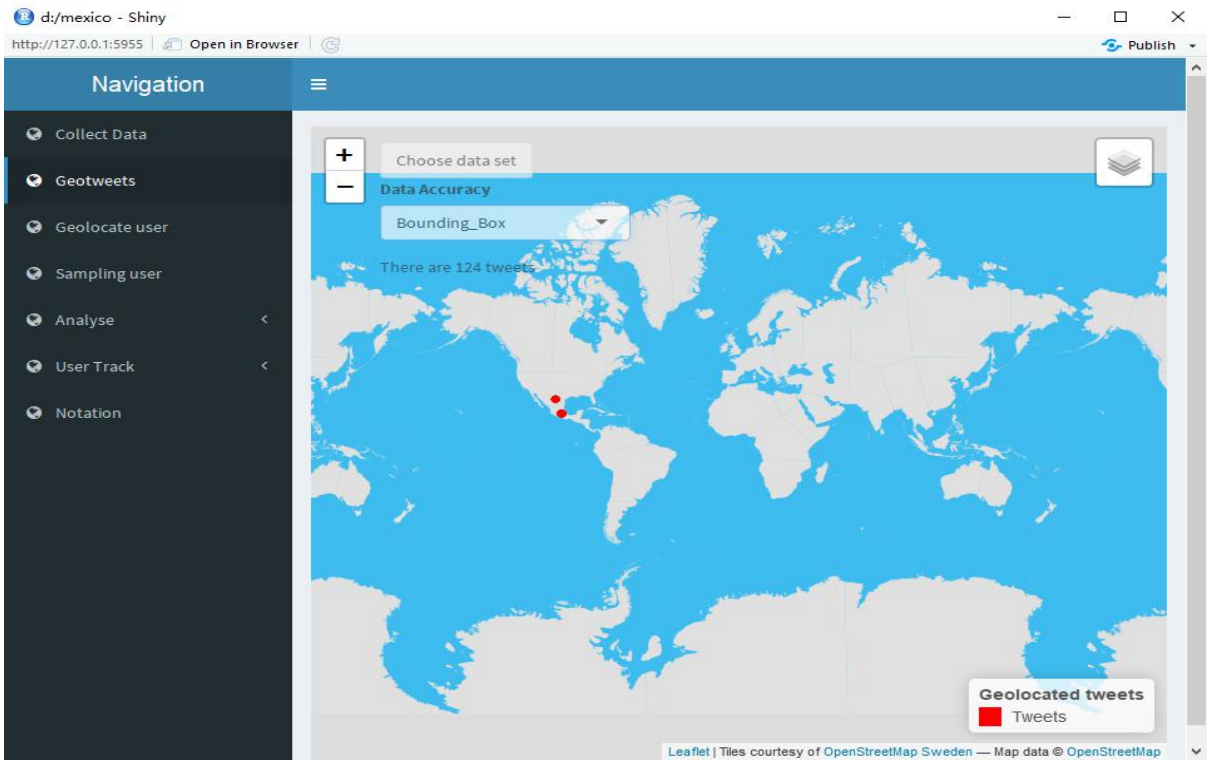
(a)


d:/mexico - Shiny
http://127.0.0.1:5955
Open in Browser
Publish
Navigation
Collect Data
Geotweets
Geolocate user
Sampling user
Analyse
User Track
Notation
Choose data set
Data Accuracy
All
There are 161 tweets
Geolocated tweets
Tweets
Leaflet | Tiles courtesy of OpenStreetMap Sweden — Map data © OpenStreetMap

(b)



(c)



(d)

Figure 2.2.2 Different Tweets Group

Fig 2.2.2 (a) show three group: LON,LAT Coordinates group , Bounding_box group,

and All group.
Fig 2.2.2 (b) show All group ,there are 161 tweets ploted on the map.
Fig 2.2.2 (c) show Geographical coordinates group, system said that there are only 37 tweets ploted on the map.
Fig 2.2.2 (d) show Bounding_box group,and there are 124 tweets ploted on the map.
Obviously, group LON,LAT coordinate plus group Bounding_box is equal to group All.
Limitations
Only when users post tweets in different bounding_box are we able to see different points on the map if user chooose All group or Bounding_box group.

2.3 Define the user's movement based on this tool/app

If we want to visualize a user's mobility patterns, it is necessary to plot a path or trajectory on the map. According to our limitation and twitter status, we define the user's movement as follows:

- User's movement in LON,LAT coordinate group.

In a specific time fragment user posted at least two tweets and the tweets in different positions. We connect the different geolocated tweets on map and the lines are the movements of the user.

- User's movement in Bounding_box coordinate group.

In a specific time fragment user posted at least two tweets and the tweets in different bounding_box areas. We connect the different geolocated tweets on map and the lines are the movement of the user.

- User's movement in All group.

In a specific time fragment user posted at least two tweets. We connect the different geolocated tweets on map and the lines are the movement of the user.

2.4 Filter bounding_box

When I plot the bounding_box on the map, I found that the varied bounding_box are not the same size. Some of their sizes are very large. For example in Mexico City, mostare in the same size level but the bouning_box named Mexico is much bigger than others. It will cause bias when we want the statistics to which bounding_box is

more popular for which user. So it is necessary to filter bounding_box and omit the bounding_box which the size is more varied from others.

3 Approach

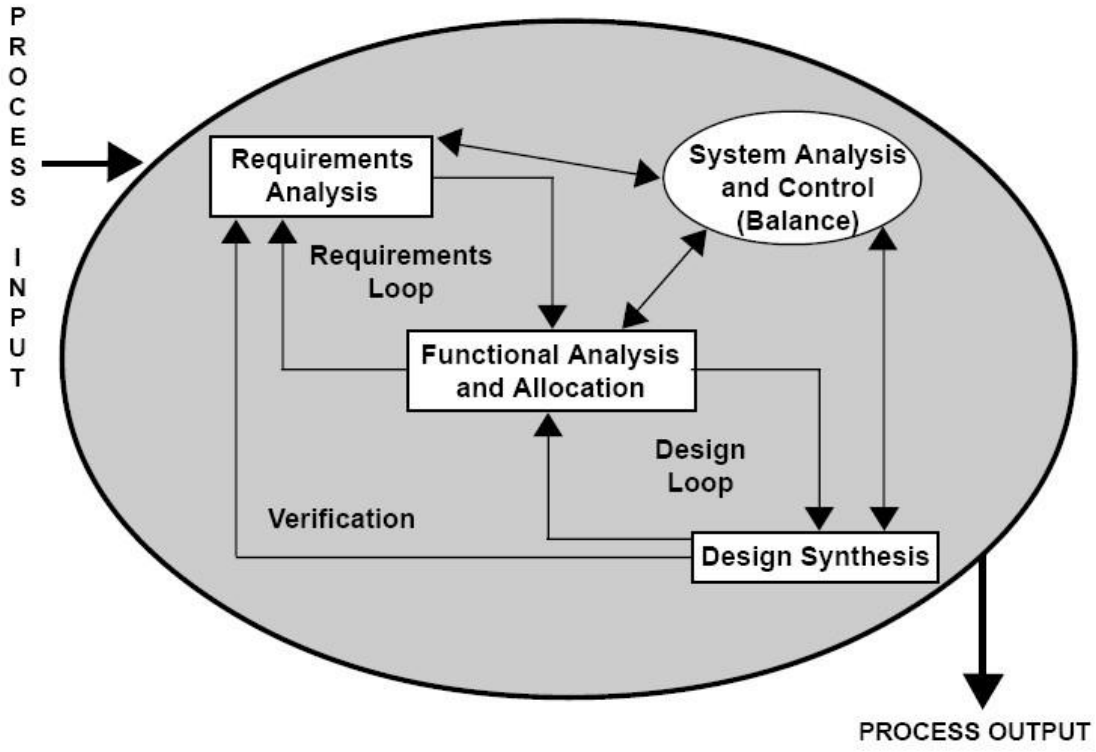


Figure 3.1 A systems engineering perspective on requirements analysis [12]

3.1 Requirements analysis

Requirements analysis plays a crucial role in the success of software development[3], so we made some requirements analysis, after requirements analysis, came up with tasks.

1. The tool/app should be able to pull data from Twitter Streaming API and geolocated tweets on the map.
2. The tool/app should be able to visualize spatial social network on the map.
3. The tool/app should be able to geolocated users and visualization on the map
4. The tool/app should be able to sample user location and visualization on the map
5. The tool/app should be able to make some centrality measurements based on the sample spatial social network.
6. The tool/app should be able to analyze the user's movemensts and visualization on the map.

7. The tool/app should be able to compare user's movements in different time ranges and visualization on the map.

I according to the main tasks write an application.

3.2 Implementation Tools

3.2.1 Introduction to R

R offical website give the introduction to R[13]:

"R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories

..."

The R environment

"R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

...."

3.2.2 Shiny

Shiny offical website give the introduction of Shiny[14]:

"Shiny is an R package that makes it easy to build interactive web apps straight from R. You can host standalone apps on a webpage or embed them in R Markdown documents or build dashboards.

...".

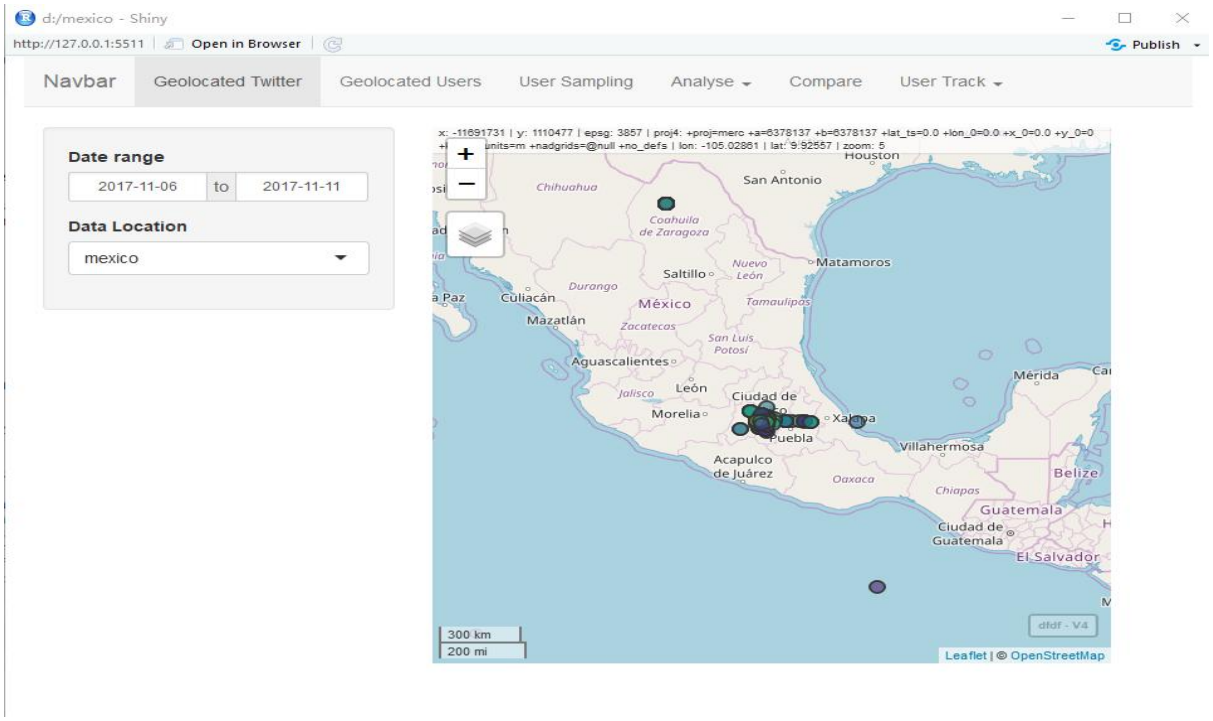


Figure 3.2 Screenshots of Shiny App

Fig 3.2 is a screenshot of the Shiny app, which i wrote.

3.2.3 Leaflet

Leaflet offical website give follow introduction[15]:

"Leaflet is the leading open-source JavaScript library for mobile-friendly interactive maps. Weighing just about 38 KB of JS, it has all the mapping features most developers ever need..."

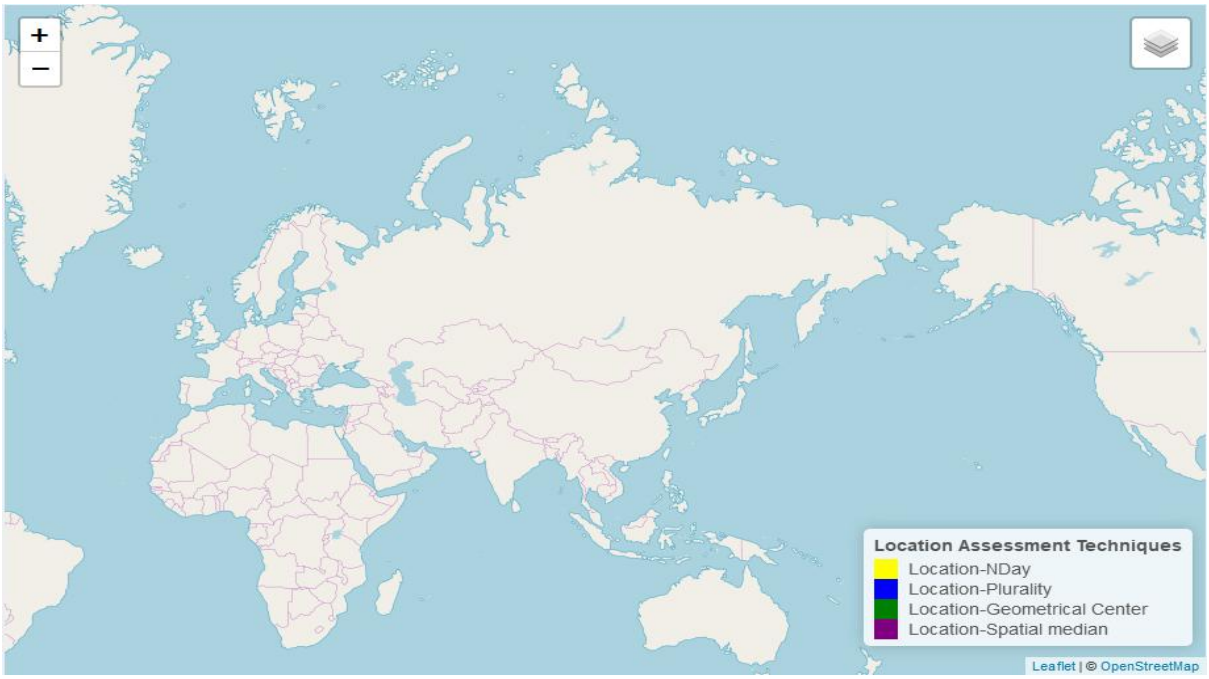


Figure 3.3 leaflet map

I used the leaflet function to visualize map on Shiny.

3.2.4 Some statistics function

- **Merge()**

Merge two data frames by common columns or row names, or other versions of database join operations[16].

- **Subset()**

Subsetting Vectors, Matrices And Data Frames.
Return subsets of vectors, matrices or data frames which meet conditions[17].

- **Data.frame()**

The function data.frame() creates data frames, tightly coupled collections of variables which share many of the properties of matrices and lists, it is used as the fundamental data structure by most of R's modeling software[18].

- **Table()**

Cross Tabulation And Table Creation.
Table uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels[19].

3.3 Structure of the work

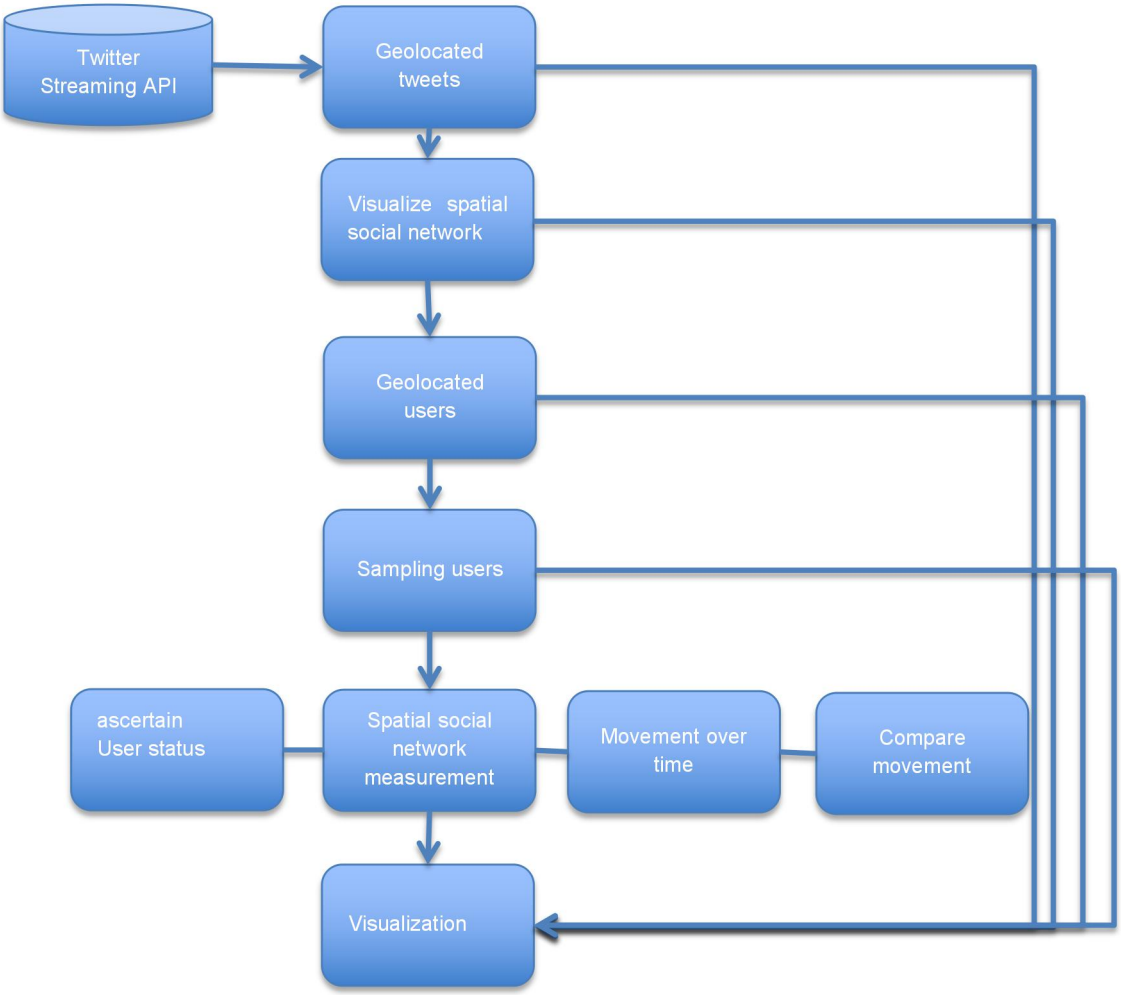


Figure 3.4 Structure of the application

Fig 3.4 Each step should be able to be Visualization, and each step is the essential for the next step.

4 Implementation

4.1 The tool/app should be able to pull data from Twitter Streaming API and geolocated tweets on the map.

1.You need to use OAuth from Twitter

1. Log into the Twitter Developers section.
2. Get Consumer Key & Consumer Secret, you have to create an app in Twitter via https://apps.twitter.com/app/new.
3. Then you'll be taken to a page containing Consumer Key & Consumer Secret.

```
# you need to use your own key, which can be obtain from tweeter
# api_key <- "xxxxxxxxxxxxxxxxxxxxxxxxx"
# api_secret <- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
# access_token <- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
# access_token_secret <- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)
```

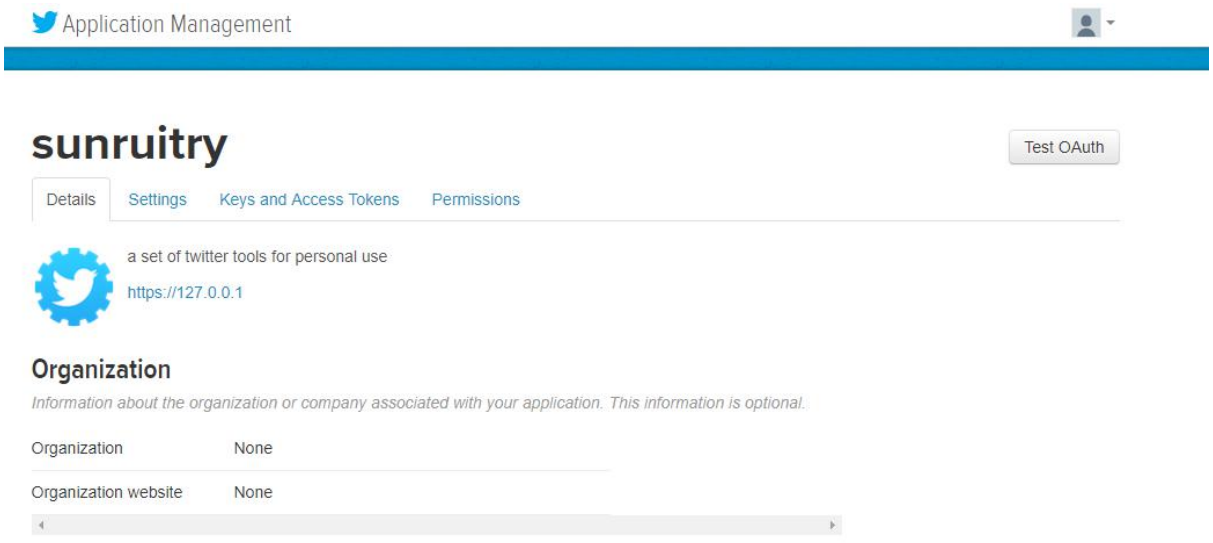


Figure 4.1.1 Screenshot of twitter application.

2.Use twitter2r or stream_tweets function to pull tweets data from Twitter Streaming API

- **What is twitter2r function?**

Tweet2r sets up parameters to file streaming and stores tweets in a JSON file using Twitter Streaming API [23].

- **What is stream_tweets function?**

  - Returns public statuses via one of the following four methods:[24]
  - 1. Sampling a small random sample of all publicly available tweets
  - 2. Filtering via a search-like query (up to 400 keywords)
  - 3. Tracking via vector of user ids (up to 5000 user_ids)
  - 4. Location via geo coordinates (1-360 degree location boxes)

In the application, use number 4 method to pull data.

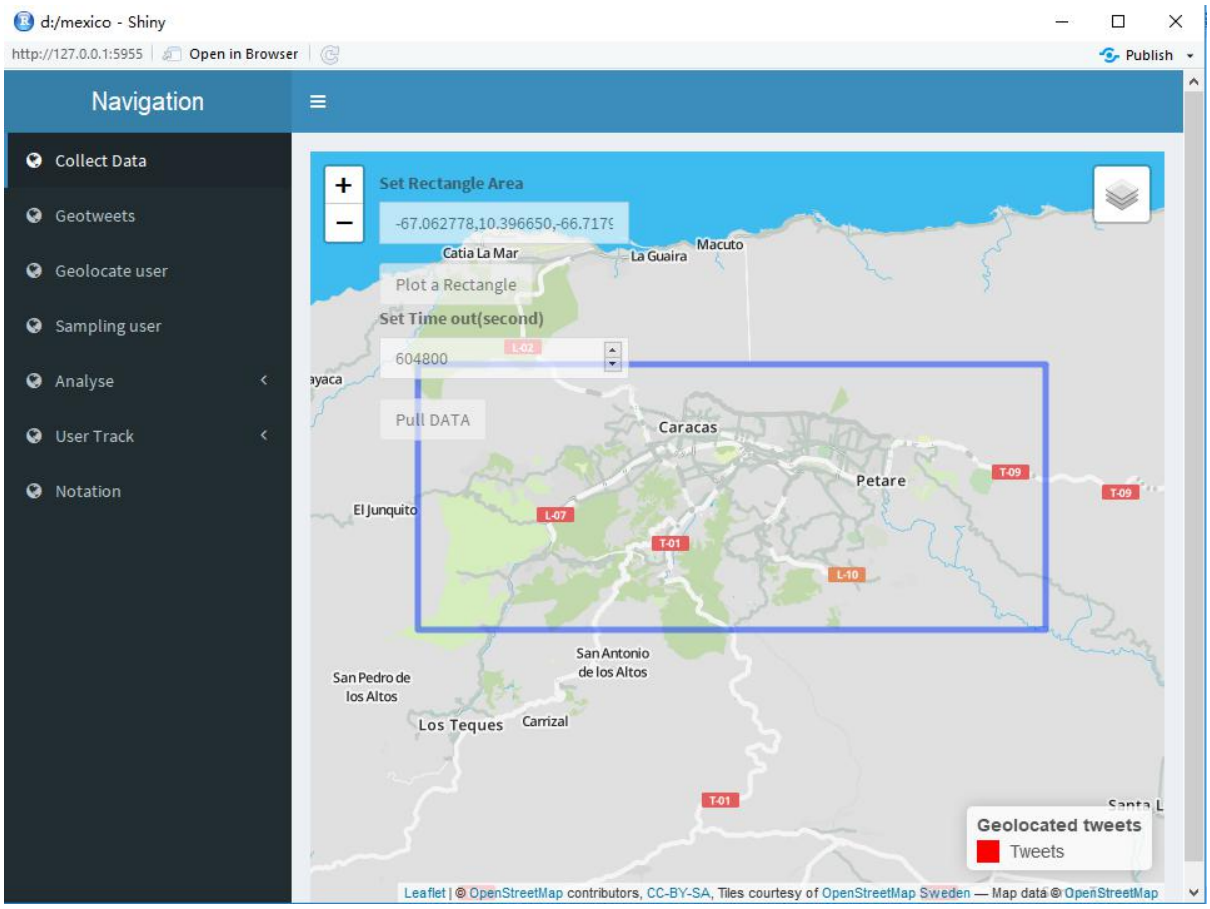


Figure 4.1.2

The application integrates the stream_tweets function to pull data from Twitter Streaming API. As Fig 4.1.2 show, the user can set a rectangle area and time range to pull tweet data from Twitter Streaming API.

'Set Rectangle Area' means pull tweet data from the Rectangle Area. 'Set Time Out' means during the time fragment from the rectangle area to pull tweet data.

In the Fig 4.1.2 I set the rectangle area :-67.062778,10.396650,-66.717953,10.540133. They are two pairs of Geographical Coordinates, any pair of diagonals of a rectangle.

Time Out is 604800s,Time Out unit is second it is calculated from 60*60*24*7.So the time fragment is 7days. User can also manually use twitter2r function to pull data from Twitter Streaming API.

3. Conversion json form to data.frame form

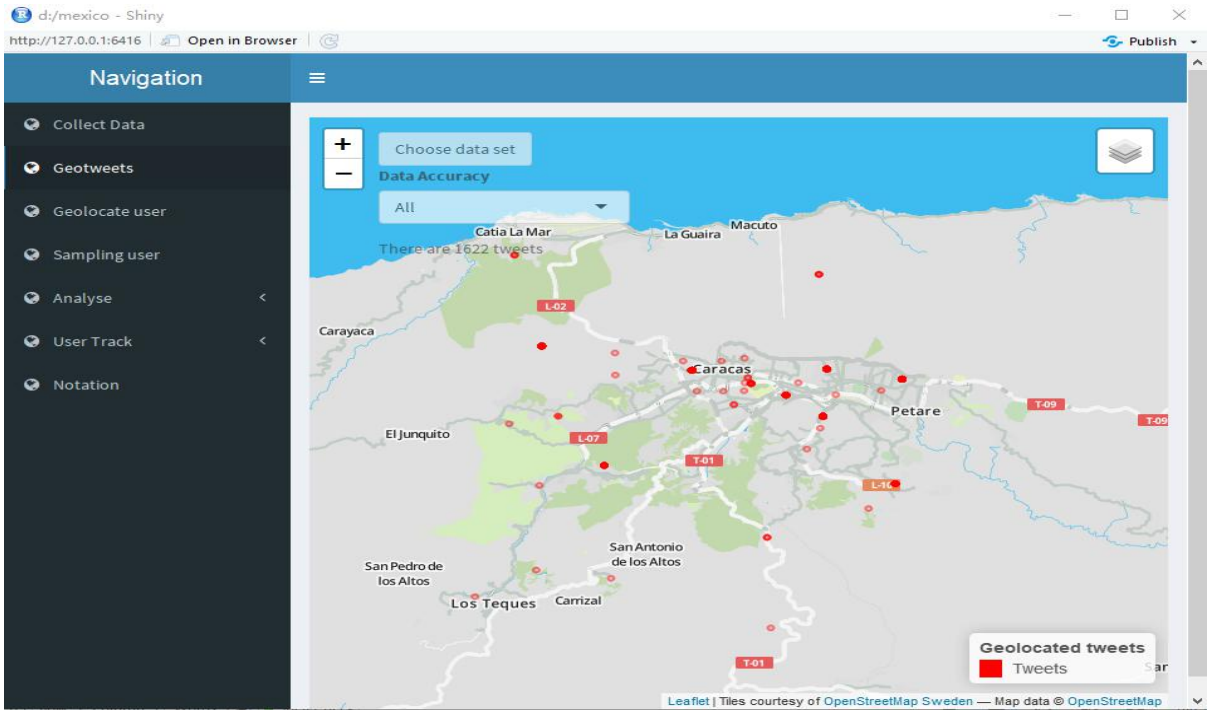
	V1	lon	lat	V4	V5	V6	V7	V8	basetime	flag
1	PairsonnalitesE	-99.1008	19.4593	Gustavo A. Madero	-99.18068	19.44494	-99.05110	19.59238	1510183080	coordinate
2	PairsonnalitesE	-99.1008	19.4593	Gustavo A. Madero	-99.18068	19.44494	-99.05110	19.59238	1510183080	coordinate
3	PairsonnalitesE	-99.1008	19.4593	Gustavo A. Madero	-99.18068	19.44494	-99.05110	19.59238	1510183080	coordinate
4	PairsonnalitesE	-99.0983	19.3344	Iztapalapa	-99.14024	19.28469	-98.96068	19.40088	1510183080	coordinate
5	PairsonnalitesE	-99.0983	19.3344	Iztapalapa	-99.14024	19.28469	-98.96068	19.40088	1510183080	coordinate
6	EPAZROJAS	-99.0965	19.4108	Iztacalco	-99.13717	19.37577	-99.05582	19.42243	1510183080	boundingbox
7	Ozereph	-99.2480	19.3609	álvaro Obregón	-99.32438	19.23223	-99.17164	19.40386	1510183080	boundingbox
8	tlacocondesa	-99.1534	19.4493	Cuauhtémoc	-99.18435	19.39983	-99.12238	19.46584	1510183080	boundingbox
9	CoachingECL	-102.5580	28.1722	Mexico	-118.40386	14.53192	-86.71222	32.71892	1510183080	boundingbox
10	Circuito_mx	-99.1615	19.3924	Benito Juárez	-99.19200	19.35710	-99.13097	19.40412	1510183080	boundingbox
11	javiexcesos	-99.1534	19.4493	Cuauhtémoc	-99.18435	19.39983	-99.12238	19.46584	1510183080	boundingbox
12	pooky0_0	-99.2507	19.4111	Huixquilucan	-99.40417	19.30597	-99.23269	19.44218	1510183080	coordinate
13	zntznt	-102.5580	28.1722	Mexico	-118.40386	14.53192	-86.71222	32.71892	1510183080	boundingbox
14	Bourge11	-99.2338	19.5091	Naucalpan de Juárez	-99.41377	19.41203	-99.20709	19.53556	1510183080	coordinate
15	trendinaliaMX	-99.1329	19.4319	Cuauhtémoc	-99.18435	19.39983	-99.12238	19.46584	1510183080	coordinate
16	trendinaliaMX	-99.1329	19.4319	Cuauhtémoc	-99.18435	19.39983	-99.12238	19.46584	1510183080	coordinate
17	trendinaliaMX	-99.1329	19.4319	Cuauhtémoc	-99.18435	19.39983	-99.12238	19.46584	1510183080	coordinate

Table 4.1.1 Part of data.frame

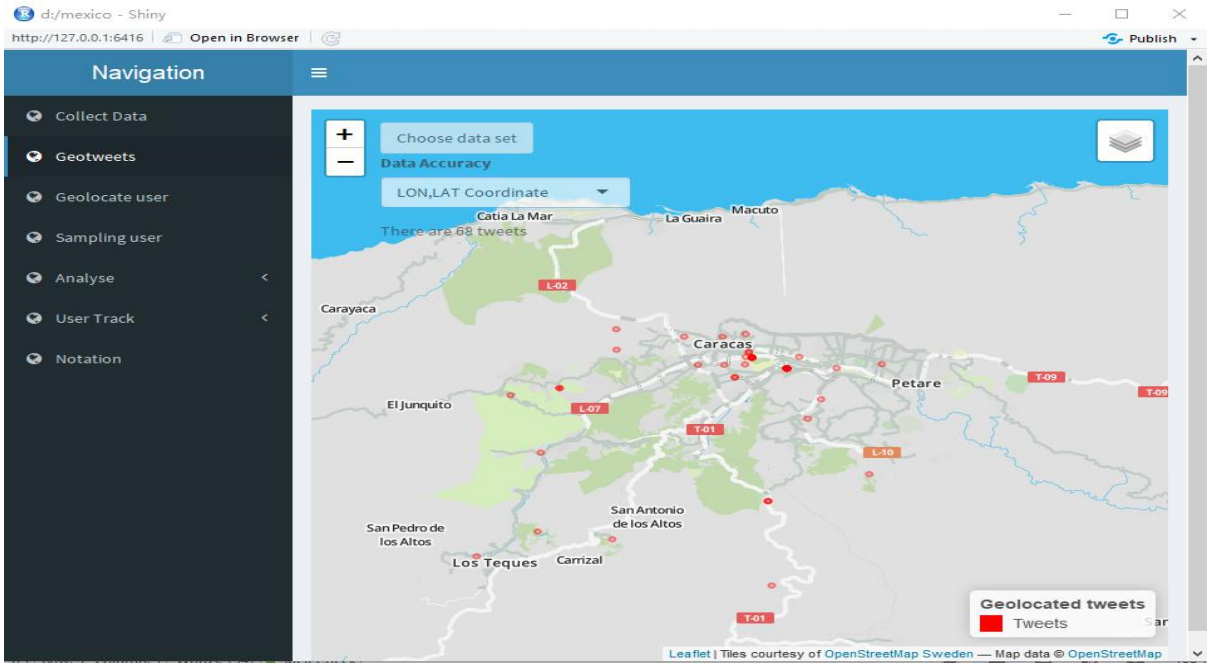
Stream_tweets give us JSON form, we have to convert the JSON form to data.frame form. Table 4.1.1 is the data.frame from which convert from JSON form. Each row is a tweet relevant information. Columns V1 is the user screen_name,lon, lat are user's location coordinates, V4 is the user's place, V5-V8 are the user's bounding_box, basetime is the time point which when user created the tweets, flag is the

Geographical coordinates kind.Use columns flag to visualize different groups: All group, LON,LAT group, Bounding_box group.

4. Visualization on map according to location coordinates in data.frame



(a)



(b)

Figure 4.1.3 Visualization different group of tweets on map

4.2 The tool/app should be able to visualize spatial social network on the map.

1 Mining relationship from the tweets data.

In the tweets data, there are two items called [“in_reply_to_screen_name”] and [“entities_user_mentions”], you can according to these two items to build mention relationship social network and reply relationship social network.

	a	b
1	EPAZROJAS	AGUCDMX
2	trendinaliaMX	trendinaliaMX
3	trendinaliaMX	trendinaliaMX
4	trendinaliaMX	trendinaliaMX
5	trendinaliaMX	trendinaliaMX
6	trendinaliaMX	trendinaliaMX
7	trendinaliaMX	RadioWera
8	Reporte311	inmamablemejia
9	_Suheyl	CrayolaDeUva
10	superchiva1968	blancafelixc
11	avrildotdeb	Xoluco
12	ArmandoGarciaC	teop4
13	alex_treto	monicagarzag
14	Corasoun	arbolfest
15	TheNorthFaceMx_	UCS_CDMX
16	hpbellamalfoy	betun_Ds

Table 4.2.1 part of mention relationship

In table 4.2.1, user in Columns ‘a’ mention columns ‘b’

2 Visualization the social network based on the relationship on the map and get a spatial social network.

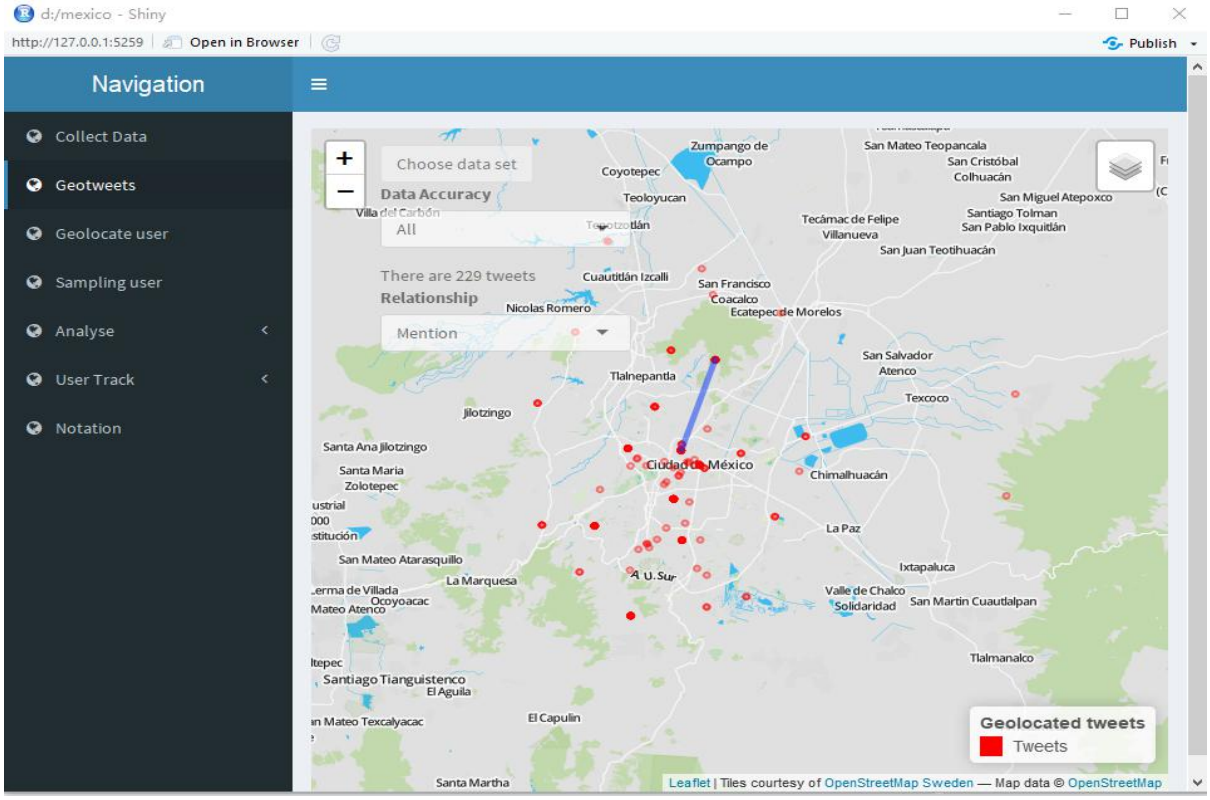


Figure 4.2.1 mention relationship spatial social network.

There is only one mention relationship in the spatial social network on map. But system said that there are four relationships in the tweet data.

	location1		location2	
	lon1	lat1	lon2	lat2
1	-99.1489	19.3641	-99.1489	19.3641
2	-98.7972	19.3961	-98.7972	19.3961
3	-99.1159	19.5555	-99.1534	19.4493
4	-99.2115	19.4517	-99.2115	19.4517

Table 4.2.2 mention relationship

In table 4.2.2 lon1,lat1 mean one tweet location coordinate, lon2,lat2 mean another tweet location coordinate. The whole row means tweet1 at location1 mentions tweet2 at location2.

You can find that row 1,2,4 location1 is the same as location2, that's why only one relationship visualization on map.

Why some location1 are the same as location2?

Because some tweets in the same bounding_box don't have specific geographical coordinates, we assume their coordinates is the geometric center of the same

bounding_box.

Why there are only four relationships in the tweet data?

Because some be mentioned tweets are not in the tweet data, so we don’t know the coordinate of the be mentioned tweets.certainly the relationship will not be considered.

4.3 The tool/app should be able to geolocated users and visualization on the map

1. preprocess the data.frame

We get a data.frame

	Var1	Var2	Freq
1	_gabrielaocampo	álvaro Obregón	0
2	_GreenHouseMx	álvaro Obregón	0
3	_israel75	álvaro Obregón	0
4	_jgomez	álvaro Obregón	0
5	_marieeta	álvaro Obregón	1
6	_natriquelme_	álvaro Obregón	1
7	05Cesgar	álvaro Obregón	0
8	182Azpeitia	álvaro Obregón	0
9	92joss	álvaro Obregón	0
10	abrownglez	álvaro Obregón	0
11	Adanzilla	álvaro Obregón	0
12	adavigevani	álvaro Obregón	0
13	Adondeiria	álvaro Obregón	0
14	aficionadolejos	álvaro Obregón	1
15	agaavee	álvaro Obregón	0
16	agalavizt	álvaro Obregón	0
17	agsalinas	álvaro Obregón	0
18	Al_Ruiiz	álvaro Obregón	0

Table 4.3.1 Part of data.frame

In Table 4.3.1 Var1 is the name of user, Var2 is the place label, meaning where the user posted the tweets, Freq means how many times the user posted tweets in the corresponding place label.

2.According to four localness assessment techniques process data.frame

The ndays technique

Assign a user to a localness, if the user at that localness posts tweets more than ‘n’ times. Prompt: if the user post tweets at different localness all more than ‘n’ times, then the user will have more than one localness.

My solution

Because we also want to be mining uses’ mobility patterns, so it is better to assign user a point-location instead of a localness.

So I modified the technique, after assign a user a localness , the user has already posted more than n tweets in n location-points in the localness, assign user to the point-location, which is the geometric center of a polygon which is a closed polygon consisting of the N tweets location-points.

Pluarilty technique

Assigning a user a localness if the user at the localness post most tweets.

Prompt: plurality technique is only assigned each user a locationness.

My solution

After assign a user a localness, the user must have already posted the most tweets in the localness and we assume the number of user post tweets in the localness is M. The location-point is the geometric center of a polygon which is a closed polygon consisting of the M tweets location-points.

Geometric median technique

Assigning a user to a point let this point minimizes the distance among all localness where the user ever posted tweets, and assign the user localness that the point belong to.

My solution

Omit that the last step, only assign a user a point that minimizes the distance among all localness where the user ever posted tweets.

Geometric center technique

Assigning a user a location-point, the location-point is the geometric center of the Polygon,which is a closed polygon consisting of the L tweets location-points. The L tweets location-points is where the user ever posted tweets.

3. visualization different localness assessment techniques on map

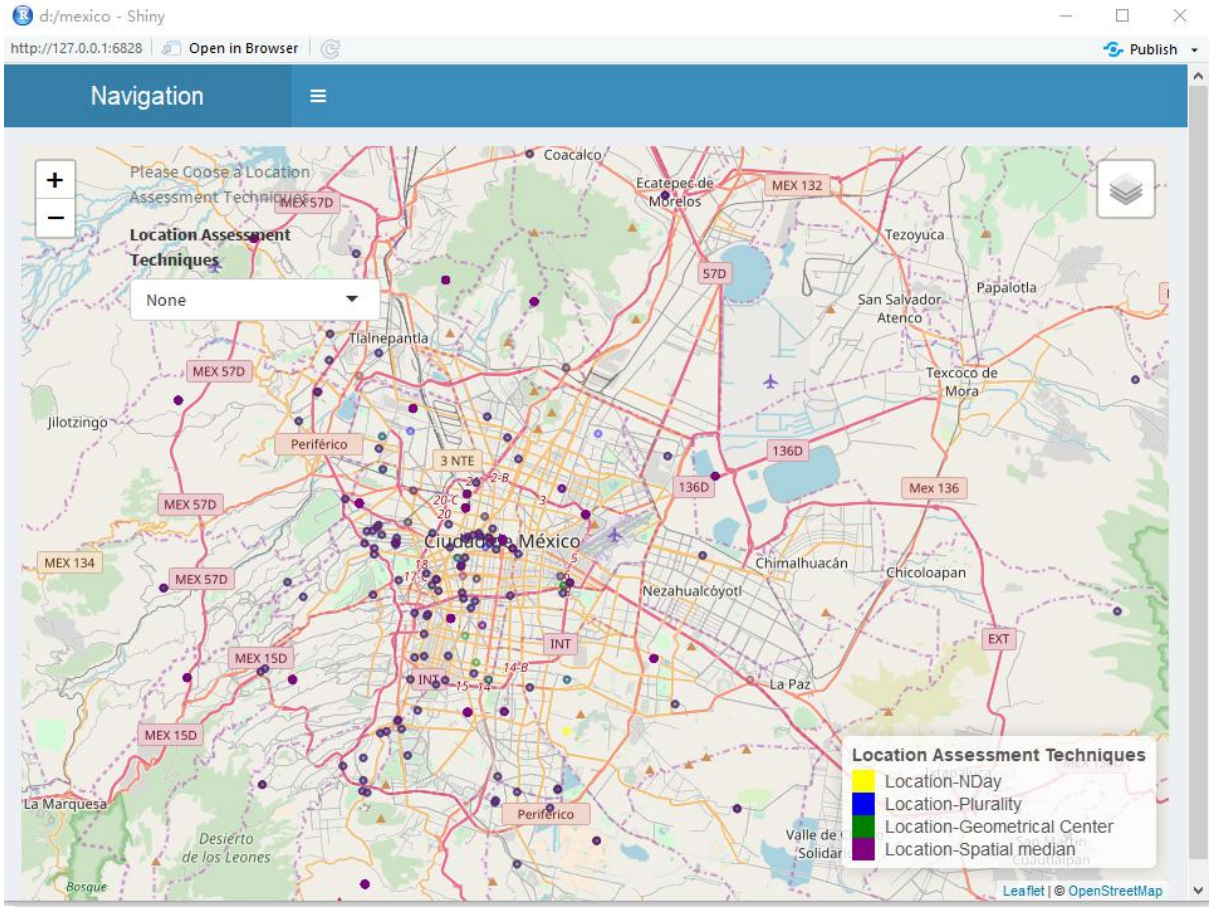


Figure 4.3.1 Visualization of different location assessment techniques

In Fig 4.3.1 different color points represent different location assessment techniques.

4.4 The tool/app should be able to sample user location and visualization on the map

After geolocating the user we get this location data. frame,in Table 4.5 the V1 column is the name of the user, the V4 column is the user's location, lon is the longitudes of the user, and lat is the latitudes of a user. In the application, there are two kinds of strategies: random sampling and grid sampling for sampling location.

	V1	lon	lat	V4
1	zutjmx	-99.1907	19.33970	álvaro Obregón
2	Zurannio	-99.1615	19.39240	Benito Juárez
3	zool_lev	-99.1884	19.33550	Coyoacán
4	zipperdf1	-99.1615	19.39240	Benito Juárez
5	ZaiyagohaN	-99.1615	19.39240	Benito Juárez
6	yuz07	-99.1534	19.44930	Cuauhtémoc
7	Yuvannamontalvo	-99.1333	19.43330	Cuauhtémoc
8	YoSoyChavaPerez	-99.1990	19.48615	Benito Juárez
9	yojamesxd	-99.1537	19.42370	Cuauhtémoc
10	YoChairo	-99.1615	19.39240	Benito Juárez
11	Yo_SoyElOtro	-99.1992	19.43550	Polanco III Sección
12	yazz_no_se_que	-99.3056	19.36190	Cuajimalpa de Morelos
13	Yazelrojo	-99.1615	19.39240	Benito Juárez
14	xeladelsur	-99.1615	19.39240	Benito Juárez
15	WladekNancy	-99.1615	19.39240	Benito Juárez
16	wendyc34	-99.2033	19.42850	Miguel Hidalgo
17	wbcmoro	-99.2480	19.36090	álvaro Obregón
18	vozandante	-99.0965	19.41080	Iztacalco

Table 4.4.1 User location data.frame

1.Random sampling

Description

Random sampling takes a sample of the specified size from the elements of x user either with or without replacement.

2 Grid sampling

According to the input number and the size of the map dynamics, a a rectangle grid

is established including points equal to the input number on the map, and select the closest points to the grid points as a sample.

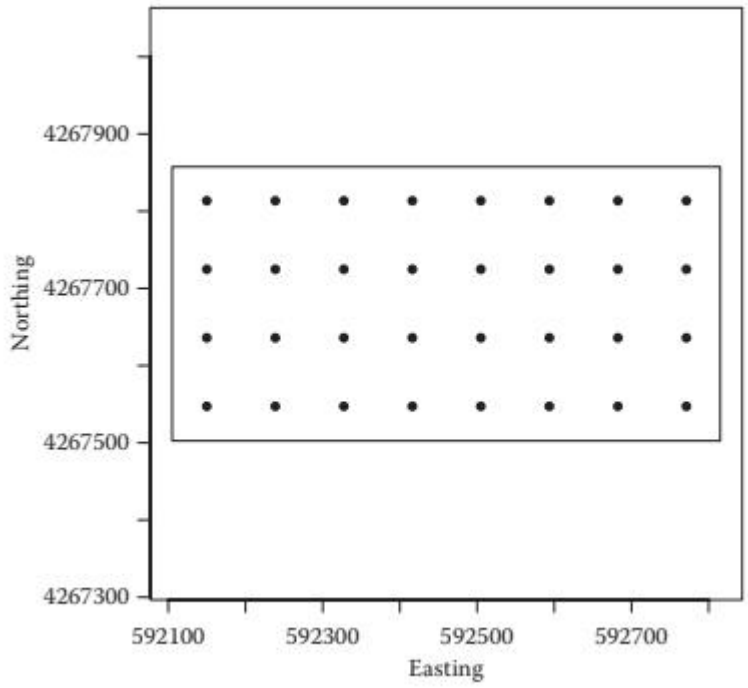


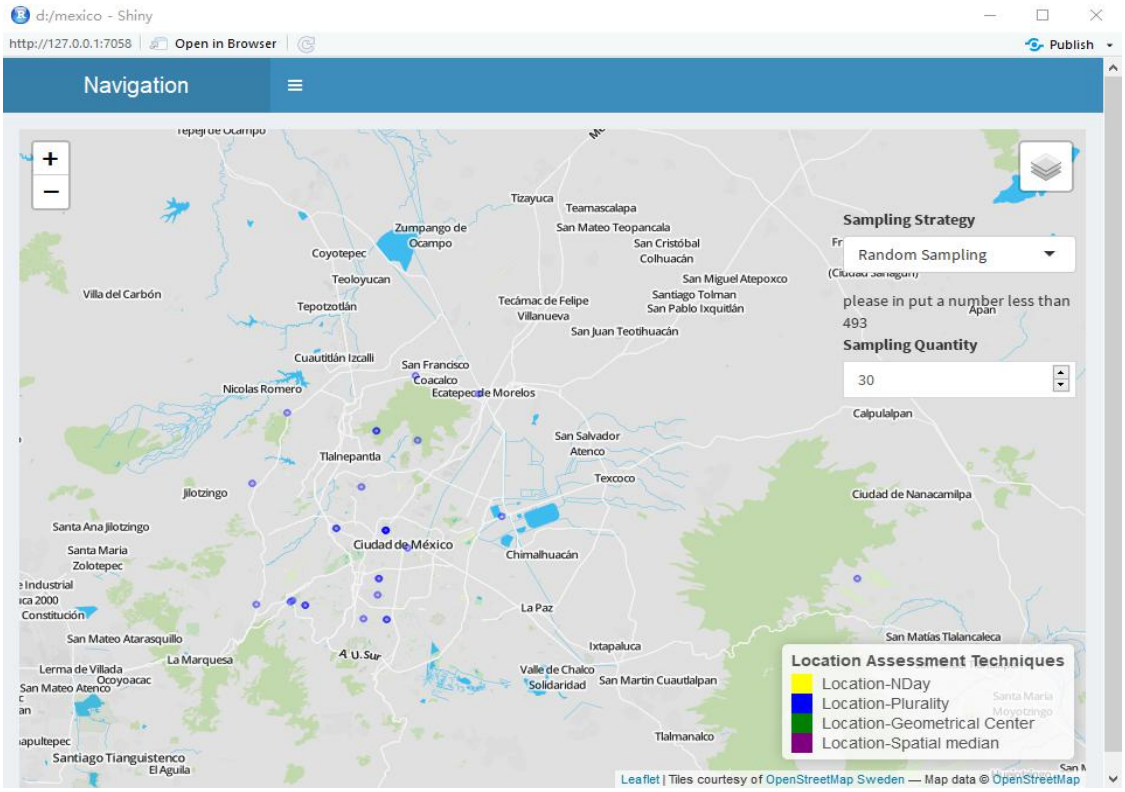
Figure 4.4.1 A regular square grid of 32 points[20].

	V1	lon	lat	V4
295	HectorOAguilar	-99.15340	19.44930	Cuauhtémoc
283	il_sO	-99.15788	19.42365	Benito Juárez
105	reenatasalomon	-99.24800	19.36090	álvaro Obregón
434	azul218azul	-99.11590	19.55550	Gustavo A. Madero
306	gracedjesus	-99.16290	19.37310	Benito Juárez
312	giberbecerra	-99.13590	19.31100	Coyoacán
433	BarryFiz	-99.10200	19.03930	Tepoztlán
118	Pocas_Chichis	-99.05050	19.37180	Iztapalapa
221	licita_alejos	-99.26540	19.36500	álvaro Obregón
25	vicentegtz	-99.16150	19.39240	Benito Juárez
462	AnaCeciGaV	-99.21150	19.45170	Miguel Hidalgo
229	LalitoohYT	-99.16430	19.56660	Tlalnepantla de Baz
207	malafamauc	-99.15230	19.34410	Coyoacán
154	NatalyaCarrasco	-99.21150	19.45170	Miguel Hidalgo
168	MoniPortes	-99.21150	19.45170	Miguel Hidalgo
266	JahelMorga	-99.15230	19.34410	Coyoacán
483	Adanzilla	-99.15340	19.44930	Cuauhtémoc
160	nadia	-99.21150	19.45170	Miguel Hidalgo

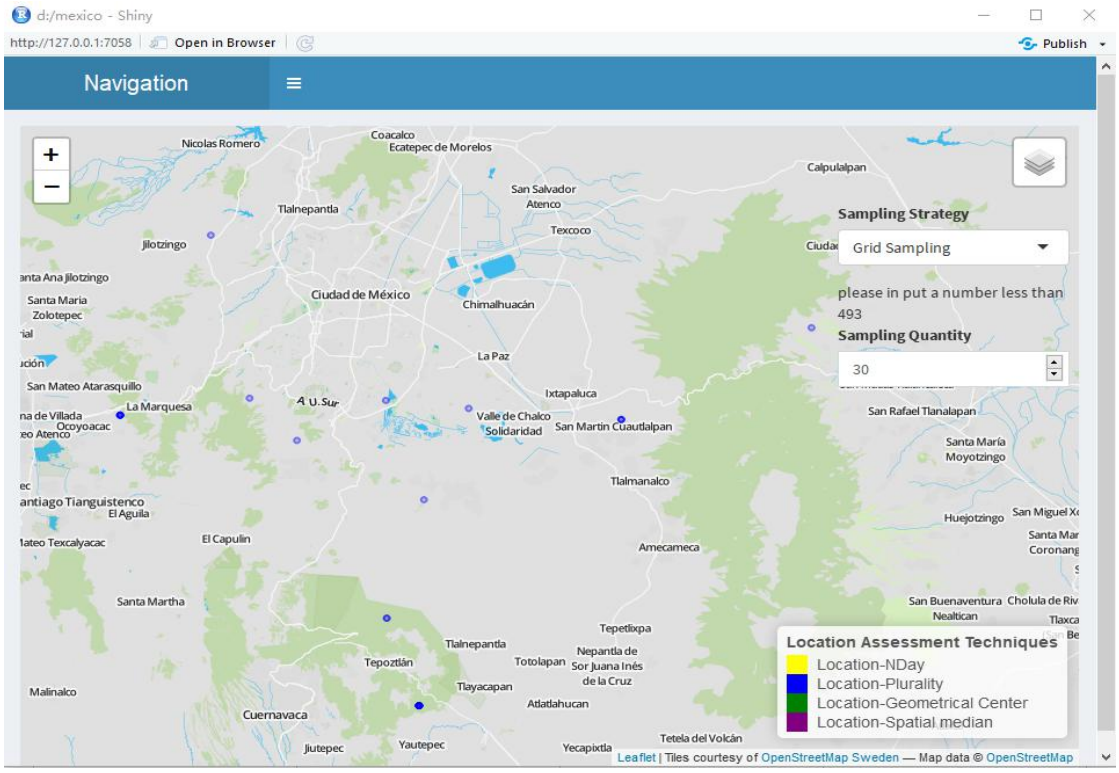
Table 4.4.2 Sample data.frame

Table 4.4.2 is a part of the sample data, I use random sampling to sample 30 data.

Visualization random sampling on map



(a)



(b)

Figure 4.4.2 Outcome of two sampling techniques.

From Fig 4.4.2 we can see a blue point means that in the last step we use the plurality technique to geolocate the user.
Fig 4.4.2(a) uses the random sampling technique Sample space are 30.
Fig 4.4.2(b) uses the grid sampling technique sample space is also 30. After comparing it is clear that the grid sampling technique outcome is dispersed

4.5 The tool/app should be able to make some centrality measurements based on sample spatial social network.

In the application, there are two kinds of spatial centrality.

1. Spatial Degree Centrality

Antonio Lima and Micro Musoles [10] give a definition and formula about spatial degree centrality:

$$C_{i,s} = \sum_{j \in Ni} |P_j \cap S|$$

The formula means that user i has some position Pj, S is the area around the Pj. At each Pj he/she posted some tweets, if he/she mention, reply or retweet others in the tweets he/she will have some relationship with others. Meanwhile others should also in the area S, The sum of relationship at each position Pj is user i spatial degree centrality.
Antonio Lima and Micro Musoles don't specific how big the area S is.So in the application, I assume S is the bounding_box where Pj belong to.
And my formula about spatial degree centrality:

$$C_{i,b} = \sum_{j \in Ni} |p_j \cap b|$$

b is the bounding_box which user i belong to.

2. Spatial Closeness Centrality

Antonio Lima defines the spatial closeness centrality , what he defines it as[10]:
Given a target point p* on Earrth, the spatial closeness centrality for a user i towards this point as its average geographic distance from all the significant place of his/her

connection

$$C^{c}_{i\ P^*} = \frac{1}{\sum_{j \in Ni} n_i} \sum_{j \in Ni} d_G(P_j, P^*)$$

.

In the application, I modify the formula.The new formula is:

$$C_i = \frac{1}{\sum_{j \in Ni} N_i} \sum_{j \in Ni} d_G(p_j, s_j)$$

Sj is the user location set and each user in the set Sj has relationship with Pj. So the formula means: user i spatial closeness centrality is the average geographic distance from user i to all user locations who user i have relationship with.

4.6 The tool/app should be able to analyze the user's movemensts and visualization on the map.

In the application, there are three kinds of means to analyze user' movements.

1. Where the users were in a specific time fragment ?

We use tweet2r() function to retrieve tweets in a particulartime fragment, so we could also track where the users were during this particular time fragment.

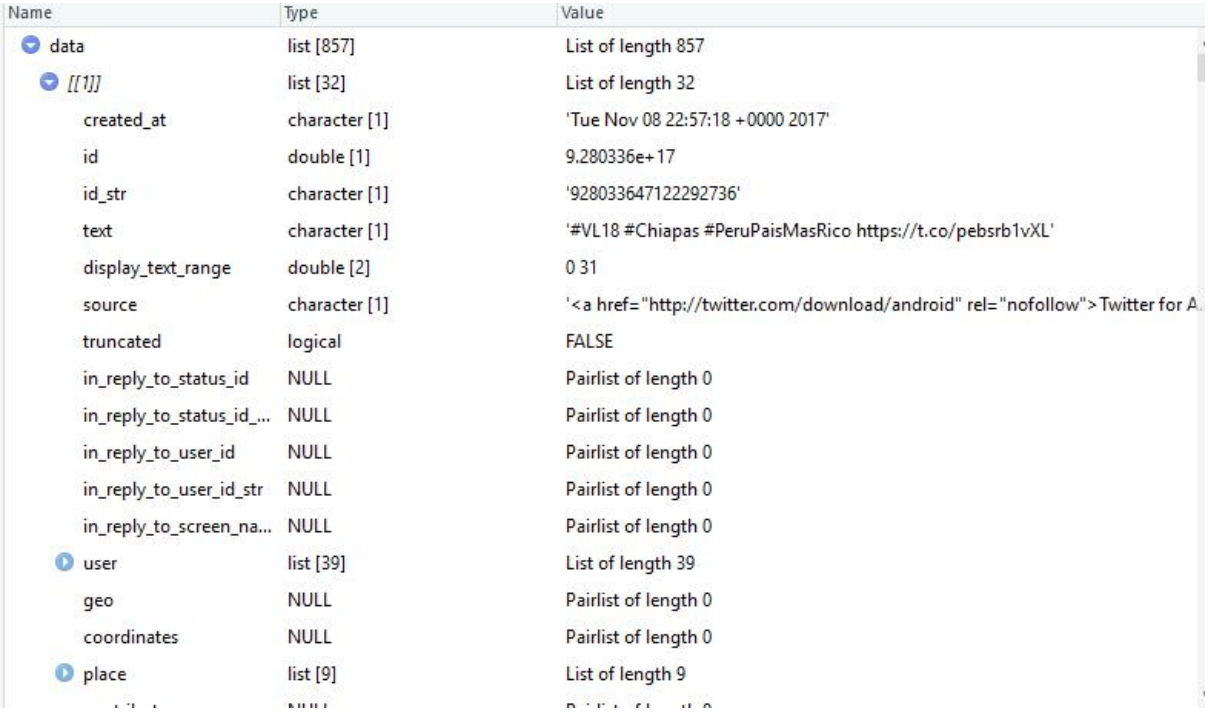
2. Where the users have been recenly ?

We also wish to know where the users positions are currently.

3.Where the users were within a specific area and specific time fragment?

The researcher can specify the study area and time fragment, and the application draws the corresponding area on the map and the user's status during this time fragment based on the input of the application.

Preprocess data



Name	Type	Value
data	list [857]	List of length 857
[[1]]	list [32]	List of length 32
created_at	character [1]	'Tue Nov 08 22:57:18 +0000 2017'
id	double [1]	9.280336e+17
id_str	character [1]	'928033647122292736'
text	character [1]	'#VL18 #Chiapas #PeruPaisMasRico https://t.co/pebsrb1vXL'
display_text_range	double [2]	0 31
source	character [1]	'<a href="http://twitter.com/download/android" rel="nofollow">Twitter for A
truncated	logical	FALSE
in_reply_to_status_id	NULL	Pairlist of length 0
in_reply_to_status_id_...	NULL	Pairlist of length 0
in_reply_to_user_id	NULL	Pairlist of length 0
in_reply_to_user_id_str	NULL	Pairlist of length 0
in_reply_to_screen_na...	NULL	Pairlist of length 0
user	list [39]	List of length 39
geo	NULL	Pairlist of length 0
coordinates	NULL	Pairlist of length 0
place	list [9]	List of length 9

Figure 4.6.1 use tweet2r funtion get tweets data

Fig 4.6.1 is the tweet data we use the function tweet2r() to get, it has an item named created_at. it means when the user post a tweet.

I use function as POSIXct() and unclass() eg:(unclass(as.POSIXct("created_at"))) converting the time eg:(Tue Nov 08 22:57:18) to a number value. This value means how many seconds from 1970.01.01 till Tue Nov 08 22:57:18.

I convert all tweets created_at item to the number value and create a data.frame, like follow Table 4.6.1 and basetime columns is the number value.

	V1	lon	lat	basetime	V5
1	CarlosArambula1	-99.2084	19.2555	1510182000	Tlalpan
2	Javo77tres	-99.2040	19.4046	1510182000	Miguel Hidalgo
3	turner_erika	-99.2446	19.4940	1510182000	Naucalpan de Juárez
4	Pocas_Chichis	-99.0505	19.3718	1510182000	Iztapalapa
5	EdmeeAguirre	-99.1534	19.4493	1510182000	Cuauhtémoc
6	AraaWL	-99.0580	19.1834	1510182000	Milpa Alta
7	DafneRivas_	-99.2337	19.6942	1510182000	Cuautitlán Izcalli
8	Pocas_Chichis	-99.0505	19.3718	1510182000	Iztapalapa
9	Ithaliasex	-99.1534	19.4493	1510182000	Cuauhtémoc
10	Figuraaaa	-99.1534	19.4493	1510182000	Cuauhtémoc
11	jeipir	-99.1694	19.4125	1510182000	Cuauhtémoc
12	PaoolaMaria	-99.1615	19.3924	1510182000	Benito Juárez
13	newemeg	-99.0878	19.4459	1510182000	Venustiano Carranza
14	MonchirrisEsp	-99.0170	19.4657	1510182000	Nezahualcóyotl
15	BetzaMPaz	-99.1534	19.4493	1510182000	Cuauhtémoc
16	gabbytoscano	-99.2115	19.4517	1510182000	Miguel Hidalgo
17	vicentegtz	-99.1615	19.3924	1510182000	Benito Juárez
18	vii_rivas	-99.1615	19.3924	1510182000	Benito Juárez

Table 4.6.1 Dftime data.frame

Select tweets in a specific time fragments.First you should know what a sliderbar does.

Figure 4.6.2 two kinds of sliderbar[21].

There are two kinds of slider bar; we use the second one mainly, because it selects a time fragment rather than a time point.

It can return two values according to the user's selection, between the value is the specific time fragment.

1. Where the users were in a specific time fragment ?

Use the merge() function; merging users location data.frame and dftime data.frame timeset<-merge(dftime,ssdfdf,by="V1",all.y=T) in order to select sample users's basetime.

Then I use subset() function to select the data.frame in the specific time fragment.

readyset<-subset(timeset,timeset >sliderbar left value)

readyset<-subset(readyset,readysetsliderbar right value))

Lastly, plot the subset points on the map.

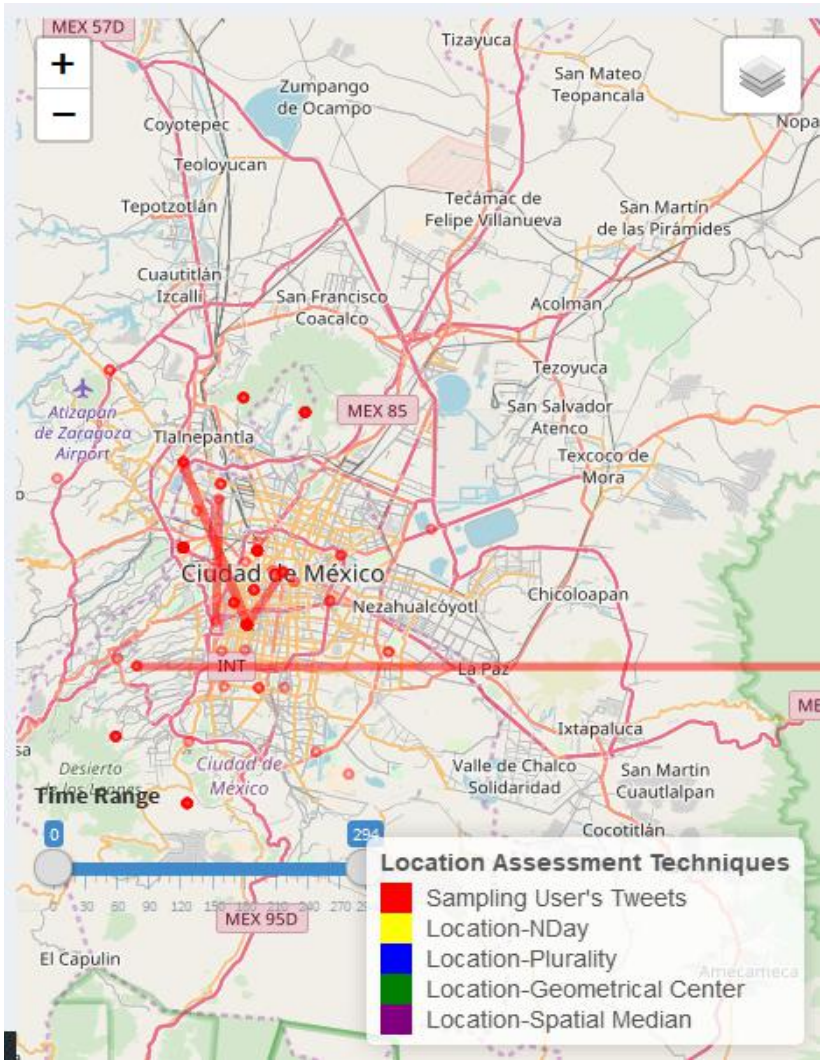


Figure 4.6.3

We can see from Fig 4.6.3 user select time fragment is from 0 to 282. Each step results in one hour .The Fig 4.6.3 shows that from 0 hour to 282 hours where the users post tweets. The red point means that user at that position post a tweet. The

red line is the movement of the user.

2. Where the users have recently been?

I use function get_timeline(): obtaining one or more user timelines (tweets posted by target user(s))) to get the tweets recent position.

It can according to tweet user names to retrieve timelines of multiple users.For example if I write

tmls <- get_timeline(c("KFC", "ConanOBrien", "NateSilver538"), n = 5) in R console

tmls will be like follow

	status_id	created_at	user_id	screen_name	text	source	reply_to_status_id	reply_to_user_id	reply_to_screen_name	is_quo
KFC.1	1001070591175249923	2018-05-28 12:00:02	15392391	kfc	When you're lost in a sea of hunger, only the Extra Crisp...	Media Studio	NA	NA	NA	FALSE
KFC.2	1000946602088026113	2018-05-28 03:47:21	15392391	kfc	@IamMarcTerrell Sorry to see this. Please DM me so we c...	Conversocial	1000942106809073665	1259779759	IamMarcTerrell	FALSE
KFC.3	1000945172253958149	2018-05-28 03:41:40	15392391	kfc	@WelchsGrape92 Sorry to see this. Please DM me so we ...	Conversocial	1000936129414115330	847263849350746112	WelchsGrape92	FALSE
KFC.4	1000944669369556993	2018-05-28 03:39:40	15392391	kfc	@TheVanderburg Can you DM with more information? I ...	Conversocial	1000929853124808707	1054712148	TheVanderburg	FALSE
KFC.5	1000925458169565184	2018-05-28 02:23:20	15392391	kfc	@Jima330 Can you DM with more information? I would l...	Conversocial	1000919756898058240	24214746	Jima330	FALSE
ConanOBrien.1	1000816434333745152	2018-05-27 19:10:06	115485051	ConanOBrien	I can't find anyone to see Solo with me this weekend, so...	Hootsuite	NA	NA	NA	FALSE
ConanOBrien.2	1000431401169276928	2018-05-26 17:40:07	115485051	ConanOBrien	I can finally rest easy knowing that https://t.co/eTwLq6ck...	Hootsuite	NA	NA	NA	FALSE
ConanOBrien.3	1000090437779099649	2018-05-25 19:05:15	115485051	ConanOBrien	My wife's new nickname for me is "Prince Harry if I squint."	Hootsuite	NA	NA	NA	FALSE
ConanOBrien.4	999723022532726784	2018-05-24 18:45:17	115485051	ConanOBrien	I never get more than halfway through my sci-fi movie pi...	Hootsuite	NA	NA	NA	FALSE
ConanOBrien.5	999458876687663104	2018-05-24 01:15:39	115485051	ConanOBrien	I played #TheShow18 with fellow great athlete Aaron Ju...	Twitter Web Client	NA	NA	NA	FALSE
NateSilver538.1	1000770751706927104	2018-05-27 16:08:35	16017475	NateSilver538	RT @Herring_NBA: My latest for @FiveThirtyEight: How t...	Twitter for Android	NA	NA	NA	FALSE
NateSilver538.2	1000425223919779840	2018-05-26 17:15:35	16017475	NateSilver538	RT @mWilstory: My view: If we agree not to name a Whi...	Twitter Web Client	NA	NA	NA	FALSE
NateSilver538.3	1000416043658104832	2018-05-26 16:39:06	16017475	NateSilver538	I think White House coverage would be better if reporte...	Twitter for Android	NA	NA	NA	TRUE
NateSilver538.4	1000388073287503872	2018-05-26 14:47:57	16017475	NateSilver538	@mattyglesias The coverage of de Blasio is often petty a...	Twitter for Android	1000361026179751937	15446531	mattyglesias	FALSE
NateSilver538.5	1000150835471507462	2018-05-25 23:05:15	16017475	NateSilver538	@ForecasterEnten Harry, as much as you and I might co...	Twitter Web Client	1000149641650999297	138141495	ForecasterEnten	FALSE

Table 4.6.2

In Table 4.6.2 It returns each user most recent five(n) tweet information, each tweet may include geotag, and if I change n to one, it will return the user most recent tweet information,and the geotag of the tweet is the user's recent position.

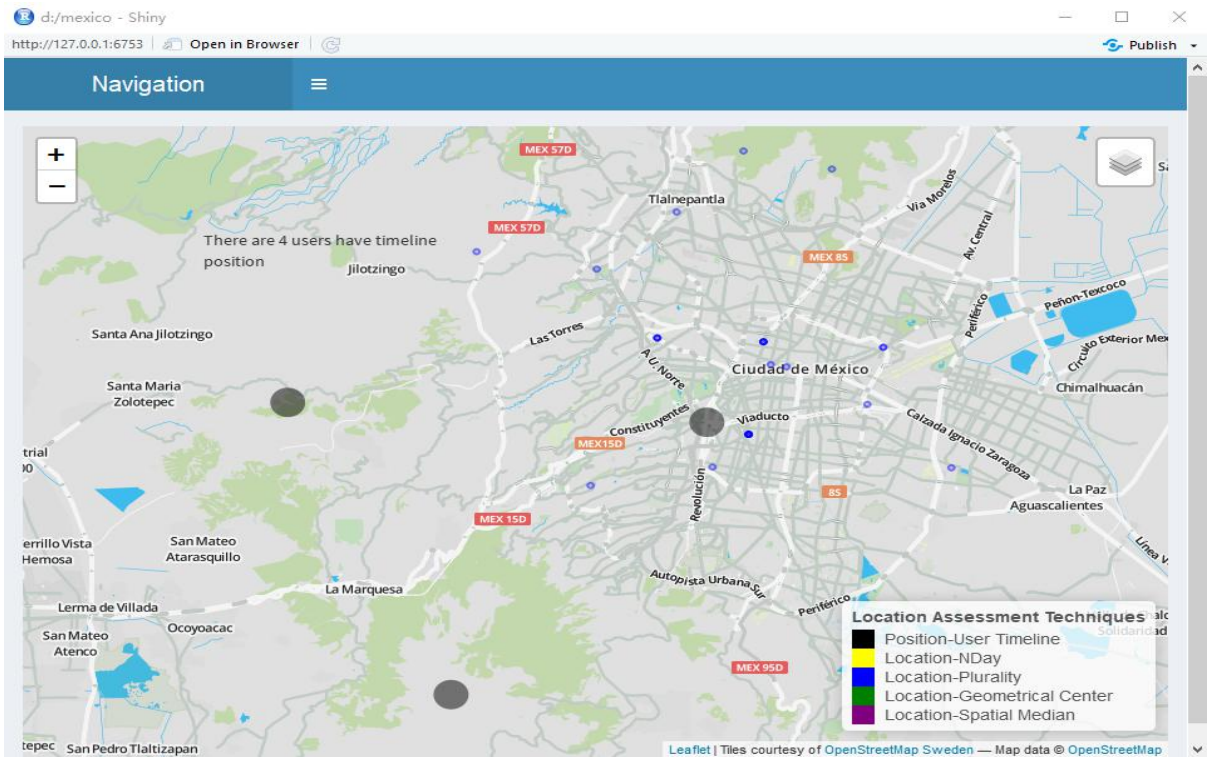


Figure 4.6.4

We can see from the figure 4.6.4 There are 30 users, but only 4 users have a recent location. Not all users will have a recent geographic location, even if they have had a specific geographic location.

3. Where the users were in specific area and specific time fragment?

I predefine 4 kinds of area: they are Circle, Boundingbox, Rectangle, and Hexagon.

- **Circle**

According a point with coordinate and radius to draw a circle on map

- **Boundingbox**

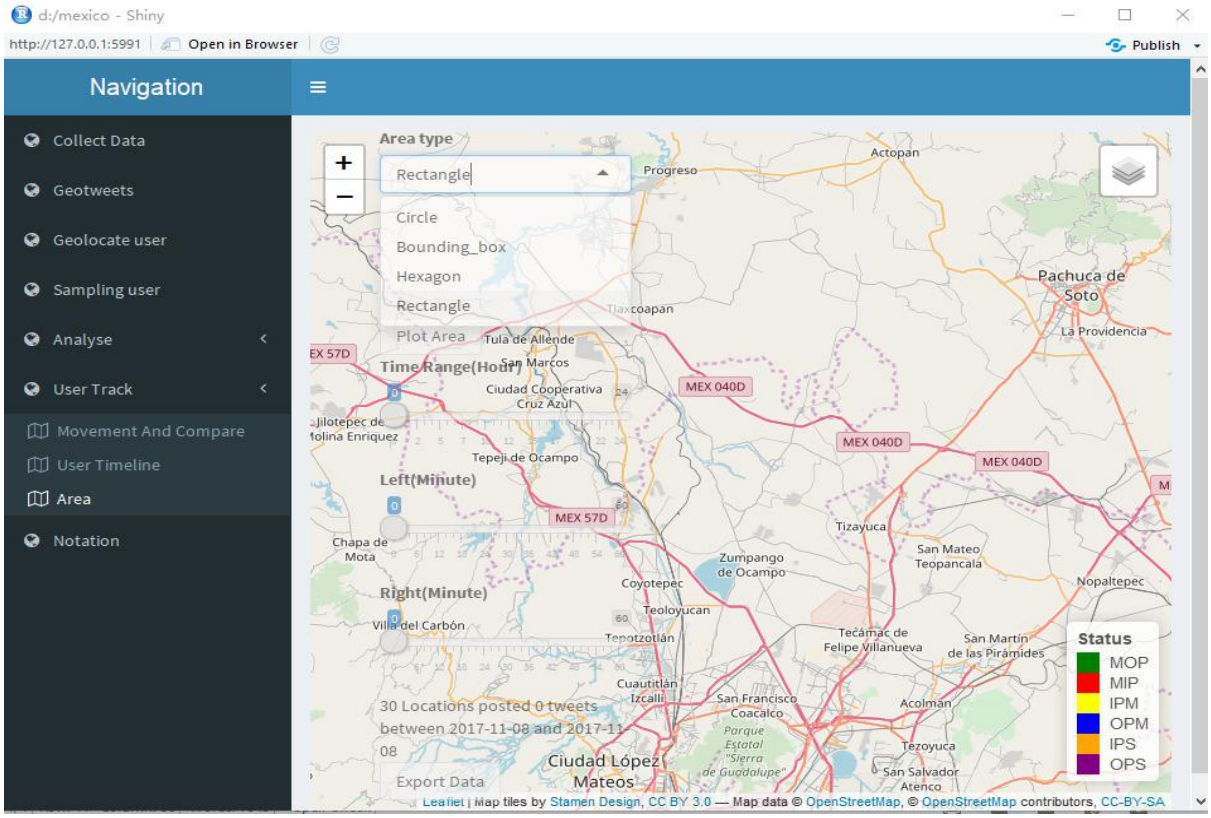
According to the name of place of the user's input to draw a bounding box on map, twitter predefines the bounding box and name of place.

- **Rectangle**

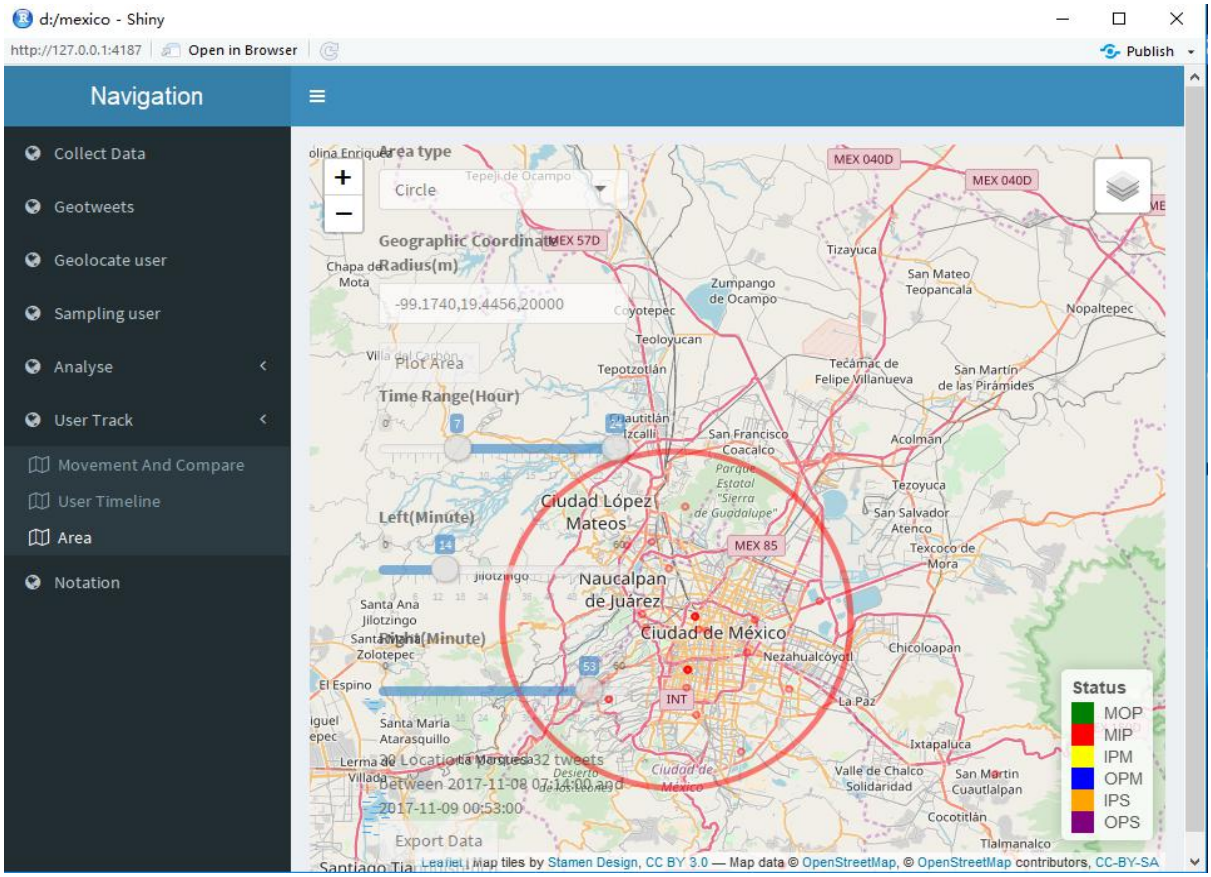
According to points with coordinates to draw a rectangle on map.

- **Hexagon**

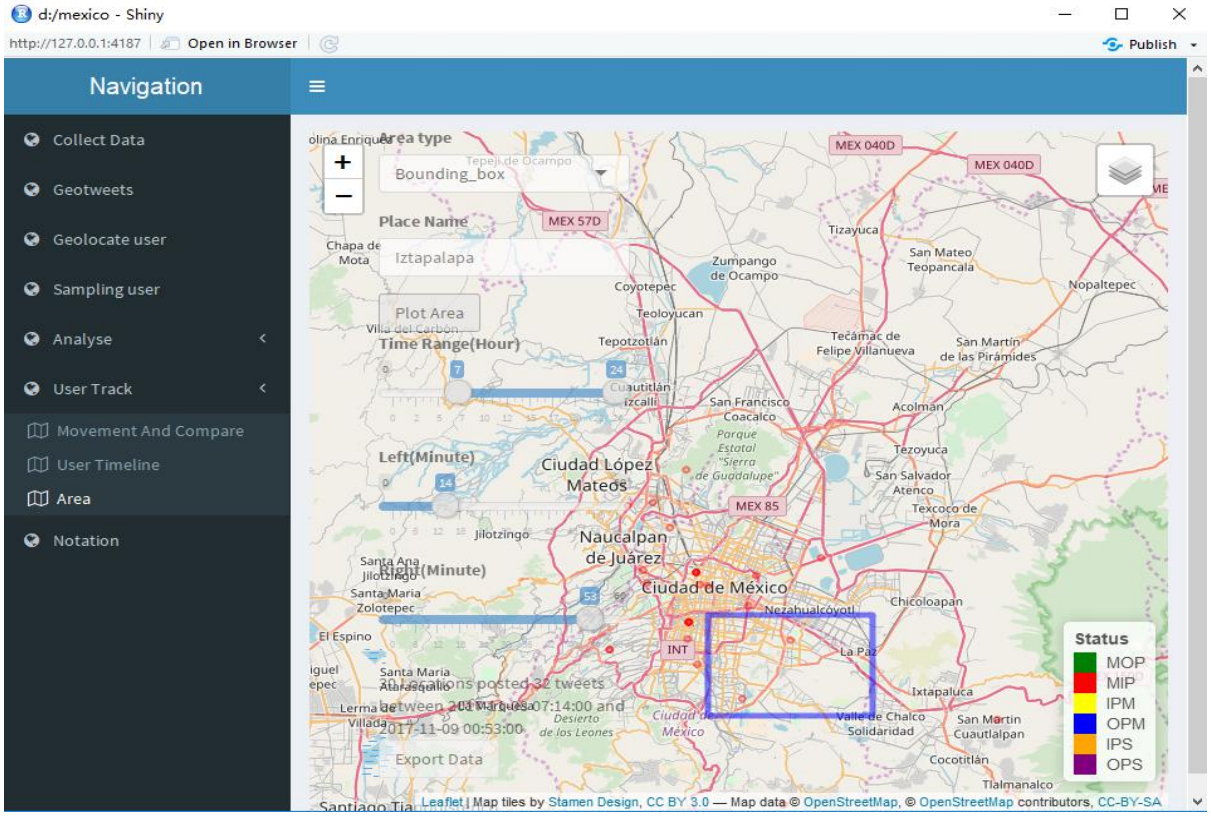
According to six points with coordinates to draw a hexagon on map, the order of the six points must be clockwise.



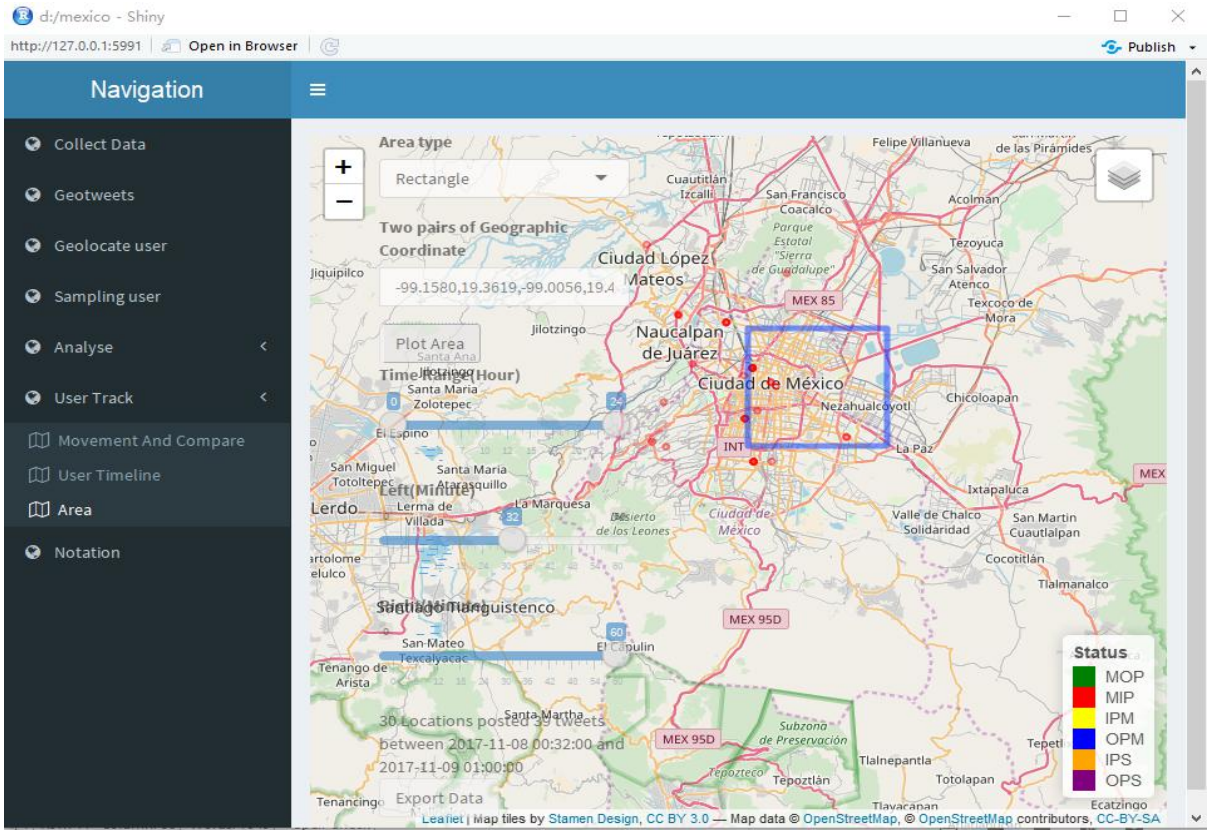
(a)


d:/mexico - Shiny
http://127.0.0.1:4187
Open in Browser
Publish
Navigation
Collect Data
Geotweets
Geolocate user
Sampling user
Analyse
User Track
Movement And Compare
User Timeline
Area
Notation
Area type
Circle
Geographic Coordinate
Radius(m)
-99.1740,19.4456,20000
Plot Area
Time Range(Hour)
Left(Minute)
Right(Minute)
Ciudad López Mateos
Naucalpan de Juárez
Ciudad de México
Status
MOP
MIP
IPM
OPM
IPS
OPS
Export Data
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap, © OpenStreetMap contributors, CC-BY-SA

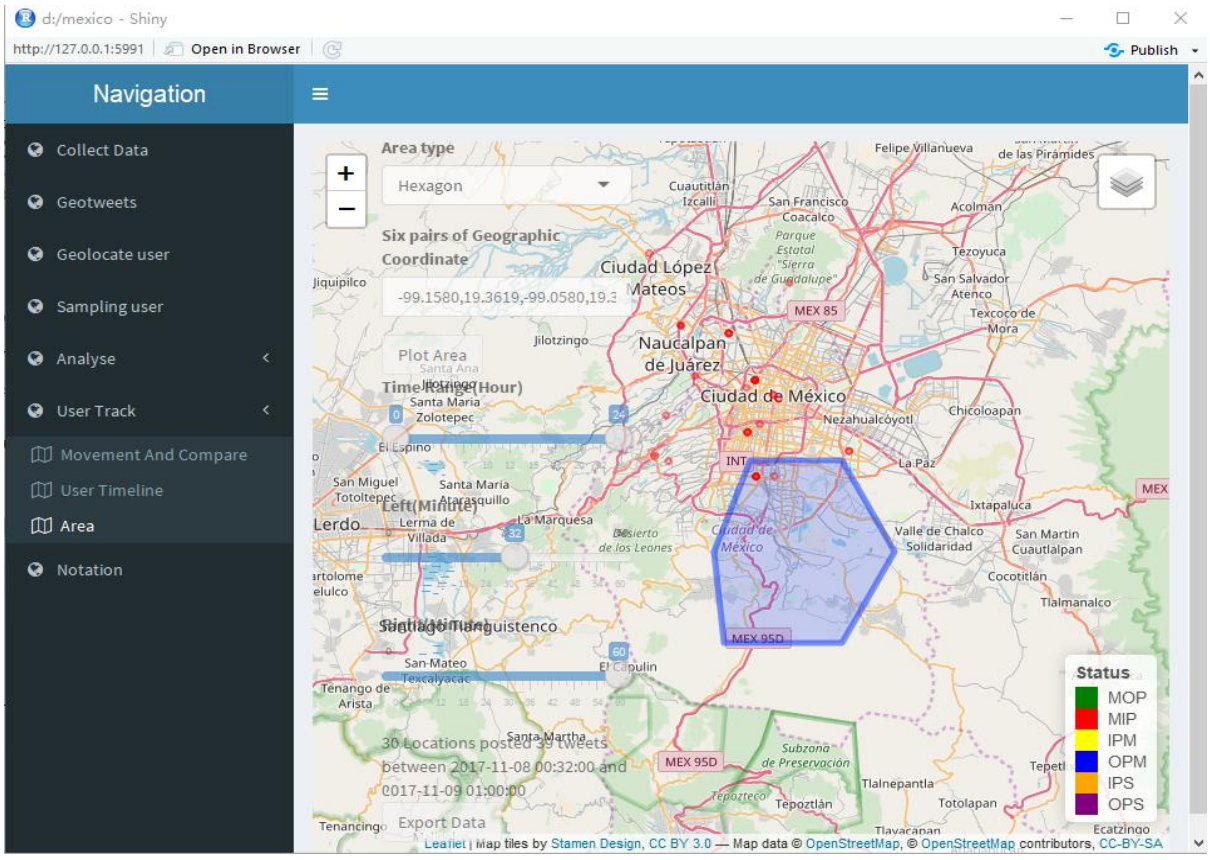
(b)


d:/mexico - Shiny
http://127.0.0.1:4187
Open in Browser
Publish
Navigation
Collect Data
Geotweets
Geolocate user
Sampling user
Analyse
User Track
Movement And Compare
User Timeline
Area
Notation
Area type
Bounding_box
Place Name
Iztapalapa
Plot Area
Time Range(Hour)
Left(Minute)
Right(Minute)
Ciudad López Mateos
Naucalpan de Juárez
Ciudad de México
Status
MOP
MIP
IPM
OPM
IPS
OPS
Export Data
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap, © OpenStreetMap contributors, CC-BY-SA

(c)



(d)



(e)

Figure 4 6.5 four kinds of predefined area

Fig 4.6.5 (a) is what I predefined as the four areas. Sliderbar ‘Time Range(Hour)’ unit is hour, ‘left(Minute)’ unit is minute,‘right(Minute)’ unit is minute too.The three sliderbar mean a time fragment,which from Time Range left point plus Left(Minute) point to Time Range right point plus Right(Minute) point.Fig(b) time fragment is from 2017-11-08 07:14:00 to 2017-11-09 00:53:00.
Fig 4.6.5 (b) is the circle area, the Center coordinate is -99.1740,19.4456 and radius is 2000 meters.
Fig 4.6.5 (c) is the bounding box area: it is the bounding box of Iztapalapa in Mexico.
Fig 4.6.5 (d) is the rectangle area, I input -99.1580,19.3619,-99.0056,19.4940 two pairs of coordinates.and the application, according to the coordinates, draws a rectangle on map.
Fig4.6.5 (e) is the hexagon area, I input six pairs of coordinates, and the application according to the coordinates, draws a hexagon on map.

1. I predefine six user’s status. They are: move out, move in, in move, out move, in stayer, and out stayer.

- **Move out**

In the specific time fragment the user moves out from specific an area(one of the four predefines areas).

- **Move in**

In the specific time fragment the user moves into a specific area(one of the four predefines areas).

- **In move**

In the specific time fragment the user will only in the specific area make some movements.

- **Out move**

In the specific time fragment the user will only out of the specific area make some movements

- **In stayer**

In the specific time fragment the user only stays in the specific area and does not make any movements.

- **Out stayer**

In the specific time fragment user only stays out of the specific area and does not make any movements.

2. How to determine the status of the user in a specific time fragment and specific area?

- **Move out**

If the user posted more than one tweet, and during the specific time fragment,the geotag of the first one of the tweets is in the specific area,and the geotag of the last one of the tweets is out of the specific area.

- **Move in**

If the user posted more than one tweet, and during the specific time fragment,the geotag of the first one of the tweets is out of the specific area,and the geotag of the last one of the tweets is in the specific area.

- **Out move**

If the user posted more than one tweet,and during the specific time fragment, all geotag of the tweets are out of the specific area.

- **In move**

If the user posted more than one tweet,and during the specific time fragment, all geotag of the tweets are in the specific area.

- **In stayer**

During the specific time fragment, if the user only posted one tweet, and the geotag of the tweet is in the specific area.

- **Out stayer**

During the specific time fragment, if the user only posted one tweet, and the geotag of the tweet is out of the specific area.

3 How to determine if a point is in a specific area?

- **Circle**

If the distance from tweet to center circle is greater than the radius of the circle, the tweet is outside the circle, otherwise it is inside the circle

- **Bounding_box**

If the geotag of the tweet is less than the first coordinate of bounding box,and greater than second coordinate of bounding box.

- **Rectangle**

If the geotag of the tweet is less than first coordinate of rectangle,and greater than the second coordinate of the rectangle.

- **Hexagon**

There is a theorem: if there is a point in a hexagon, then each edge of the hexagon can form a triangle with this point. We have six triangles, and the sum of the area of the six triangles must be equal to the hexagon’s area.

4.7 The tool/app should be able to compare user's movements in different time ranges and visualization on the map.

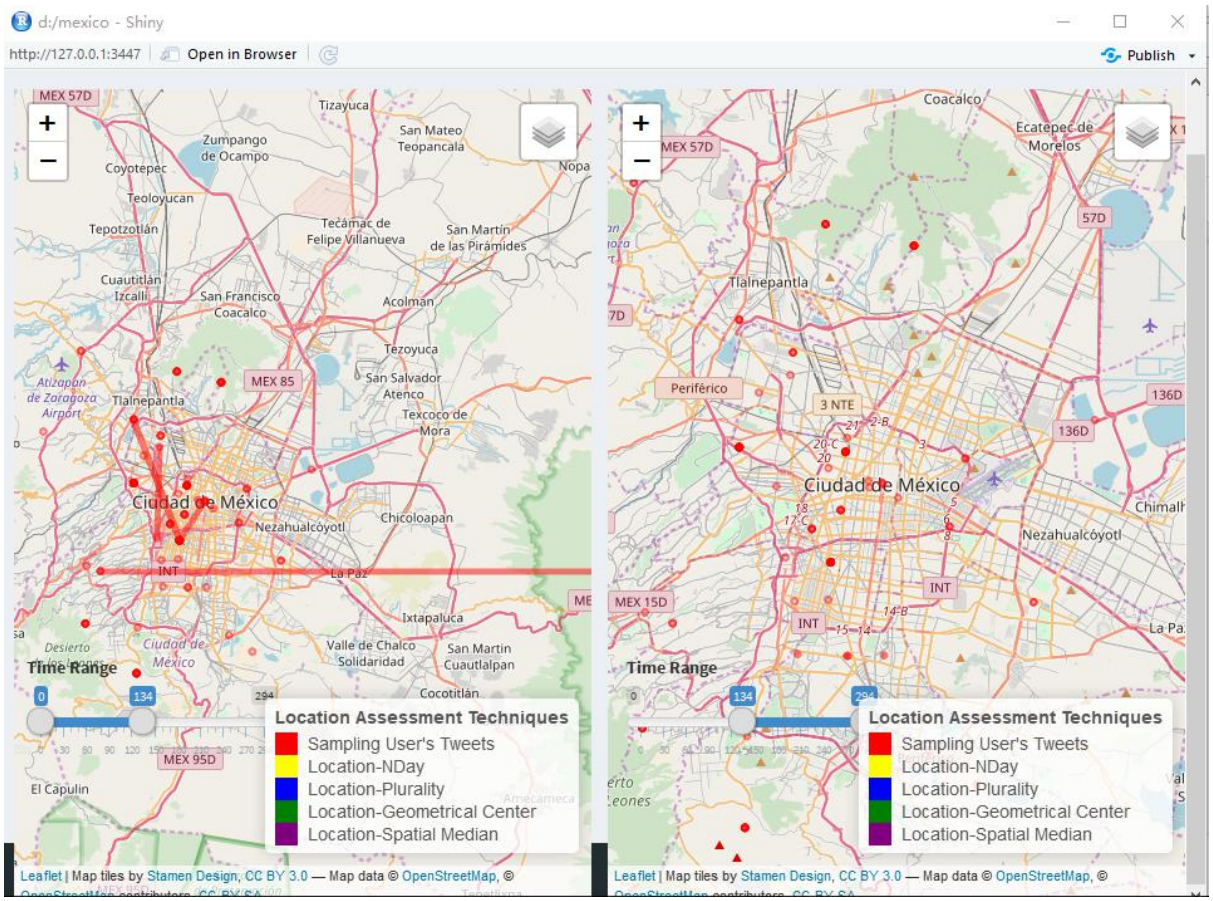


Figure 4.7.1 compare movement

I added two maps using the same function in a same place as Fig 4.7.1showed, so that we can compare two different time fragment of user's movements.

The left map displays the time fragment from 0 hour to 134 hours and right map time fragment displays 134 hours to 294 hours.

5 Evaluation

5.1 Evaluation of program outcome based on specific data

1 Heat map

- **What is heat map?**

"A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors. "

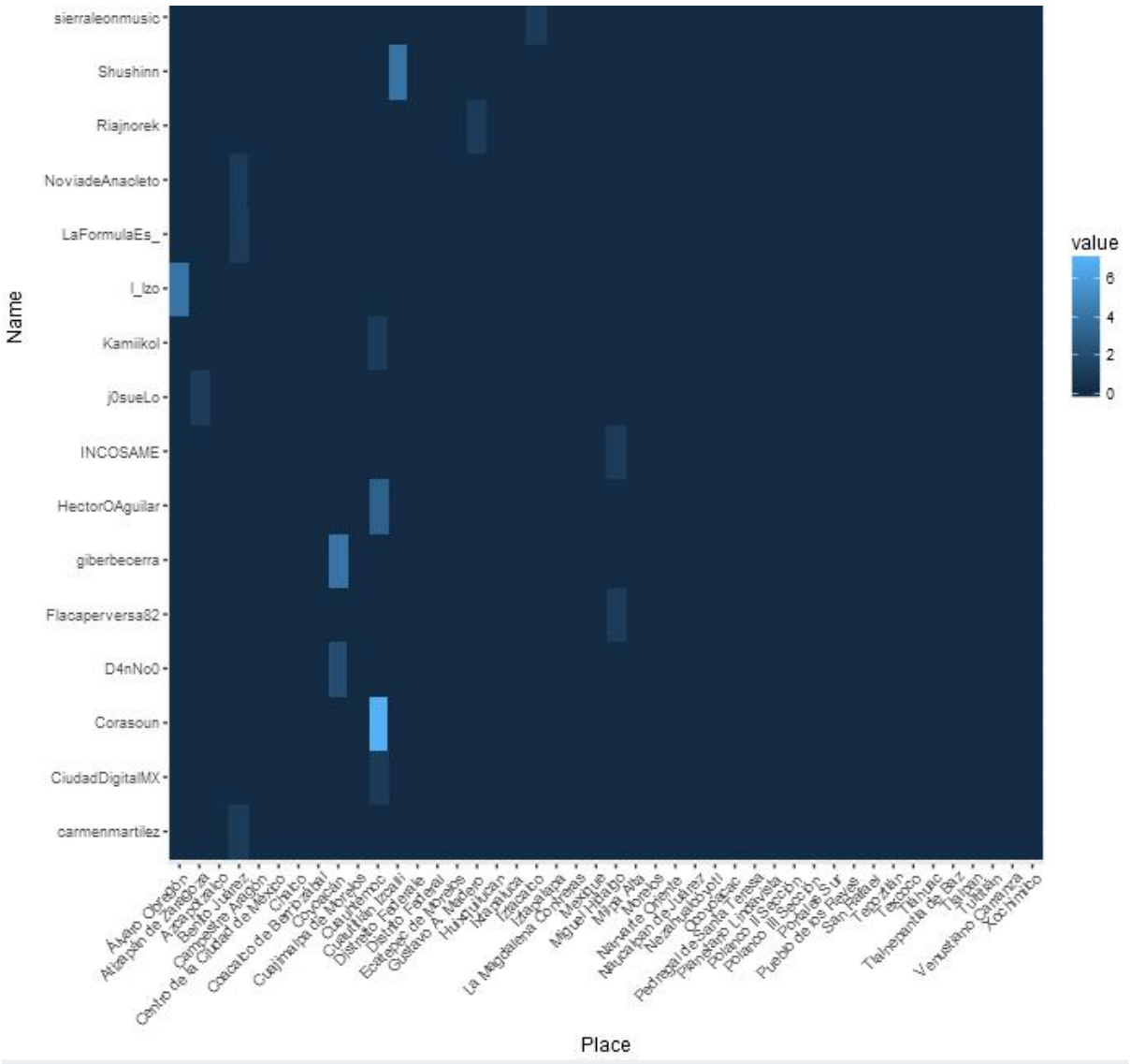


Figure 5.1.1 sample tweets heap map

I use the heat map function and sample location data.frame plot a heap Fig 5.1.1 Left is the sample user's name, the bottom is the name of place and the Fig 5.1.1 means whoever went to what place posted how many tweets.

From figure 5.1.1 we can decipher that the location named Cuauhtémoc is more popular than other places.

2 spatial degree centrality

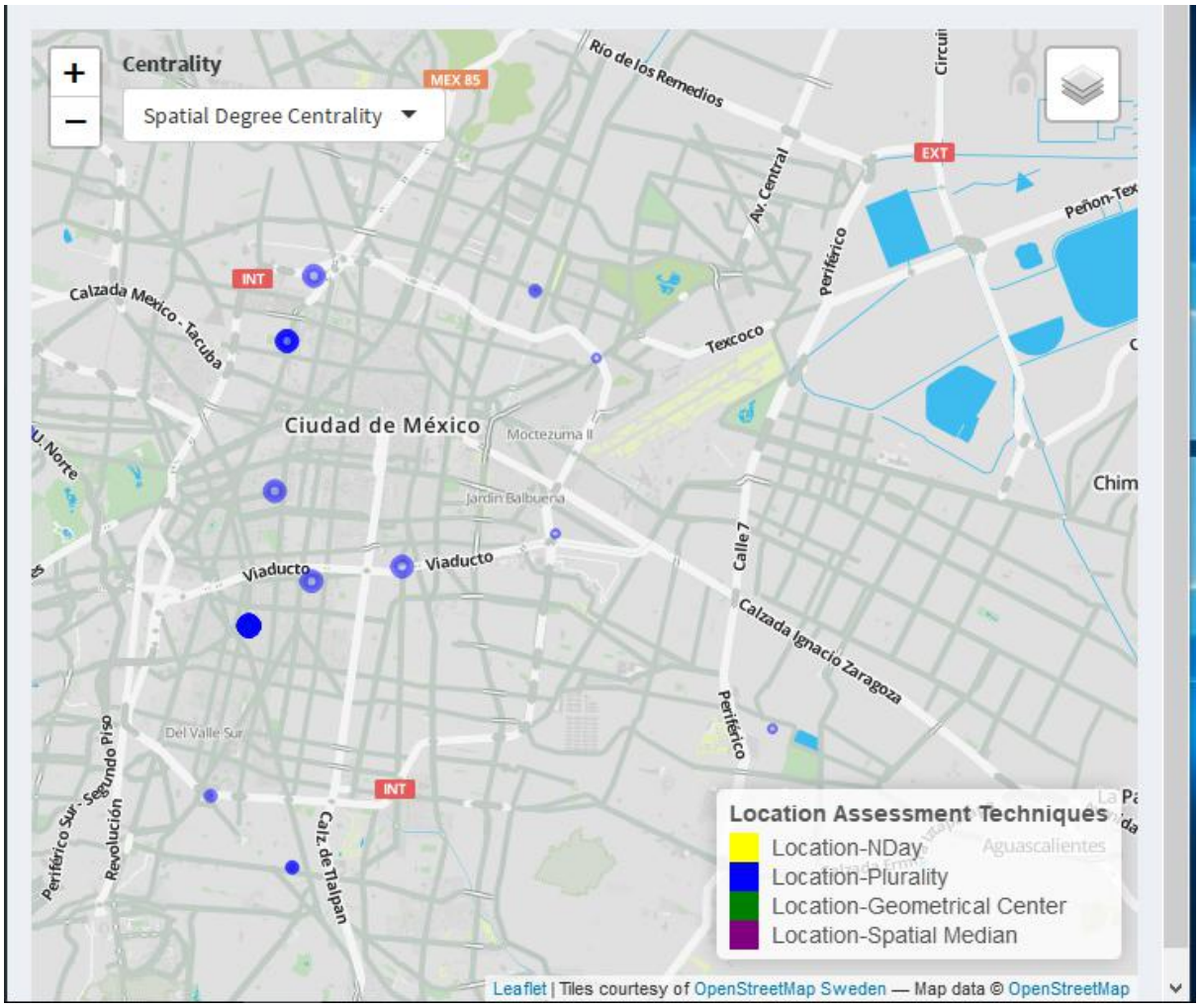


Figure 5.1.2

After sampling users, we can take some measurements, Fig 5.1.2 portrays that after sampling plurality user's location, and taking spatial degree centrality measures, the bigger the points are, and the higher centrality of the the users. From figure we can find that people who are in the center of the city have a higher spatial degree centrality

3 spatial closeness centrality

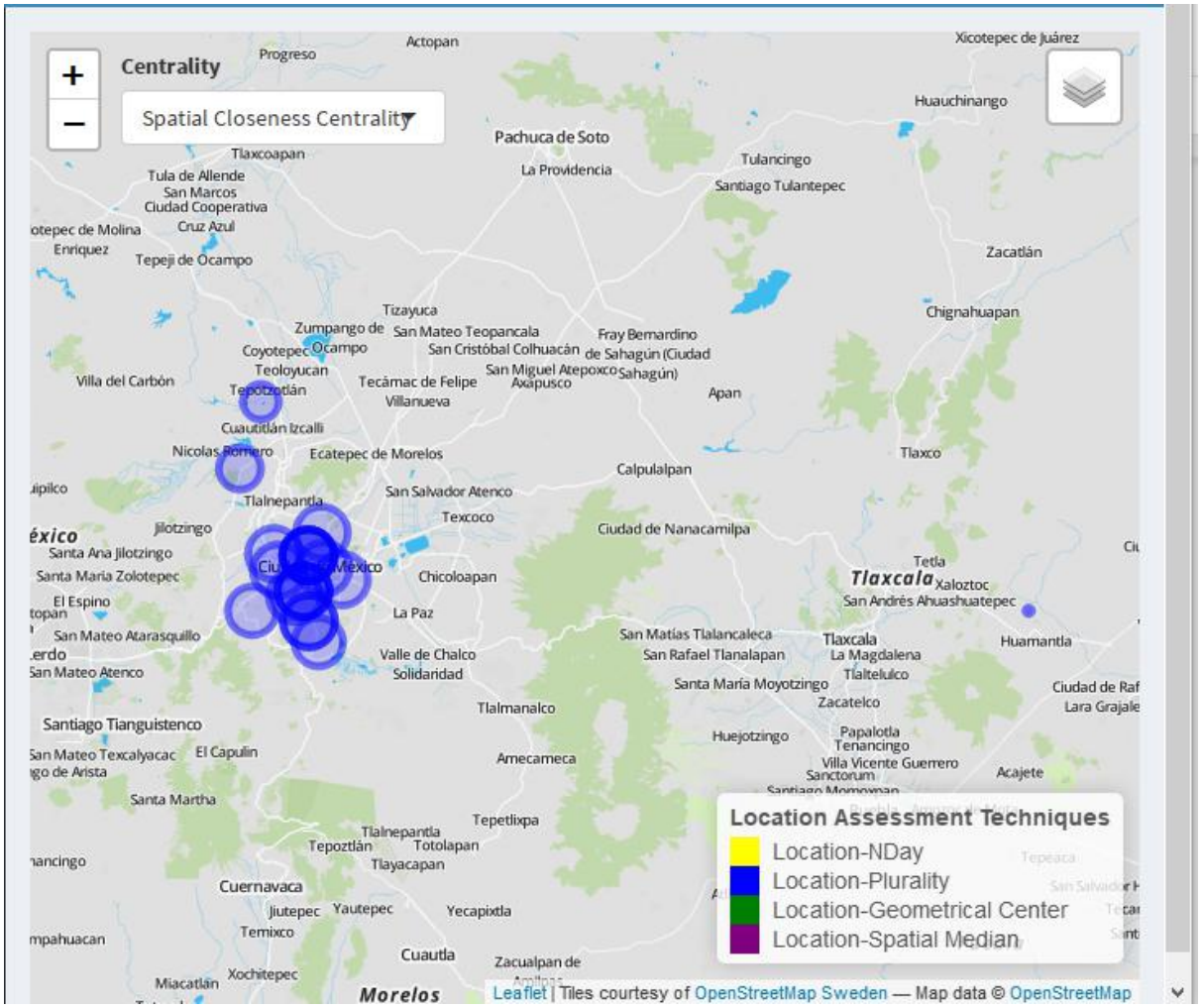


Figure 5.1.3

After sampling users,we can take some measurements,figure 5.1.3 means after sampling plurality user's location,and take spatial closeness centrality measure.the more bigger the points are,the higher centrality the users are. From figure 5.1.3 we can find that people who are in the center of the city have a higher spatial closeness centrality

6 Summary and Outlook

6.1 Advantage of the application

1. This application is easy to operate, the UI is simple and easy to understand.

2. The application can geolocate users, sample users, trackusers, and analyze the outcome based on user's selection; every step is visualization.

3. The application can reduce selection bias.

6.2 The use scenarios of application

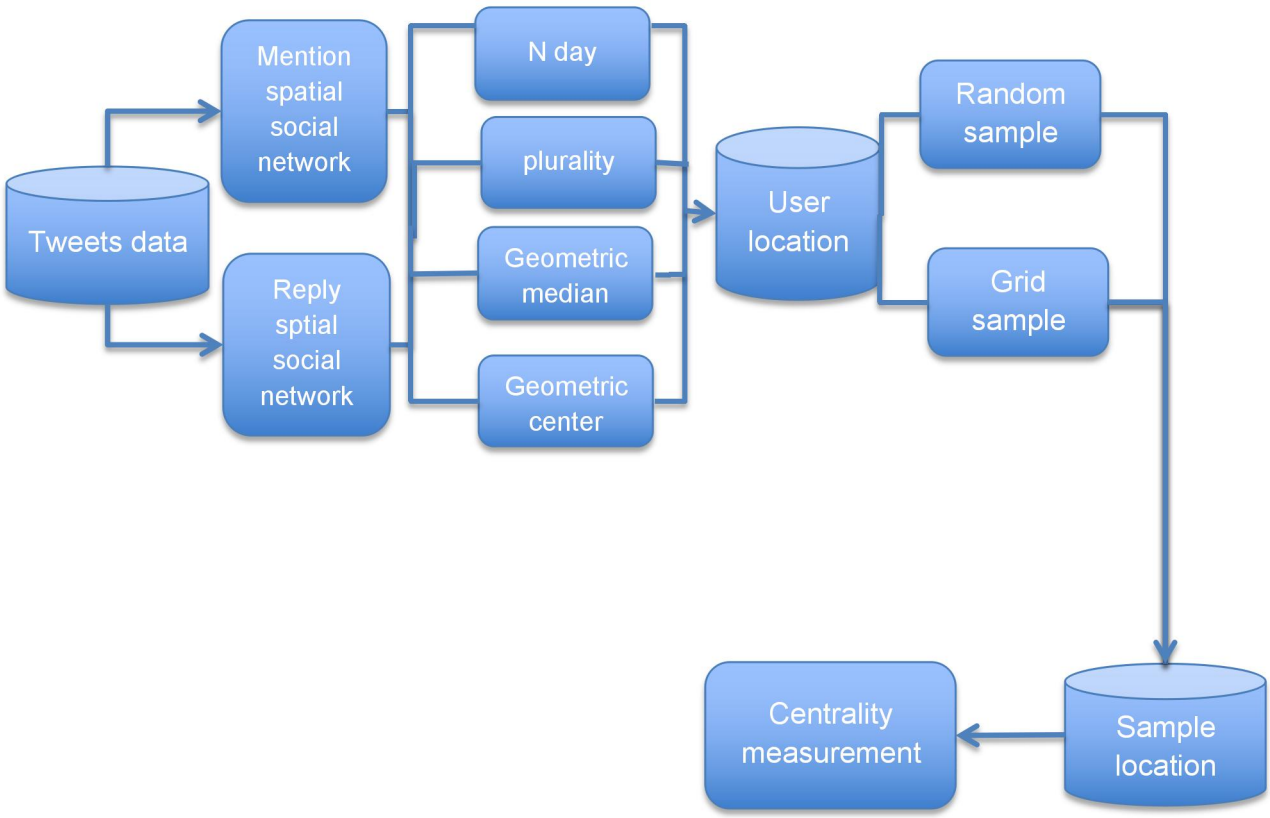


Figure 6.2.1 Application usage flow chart

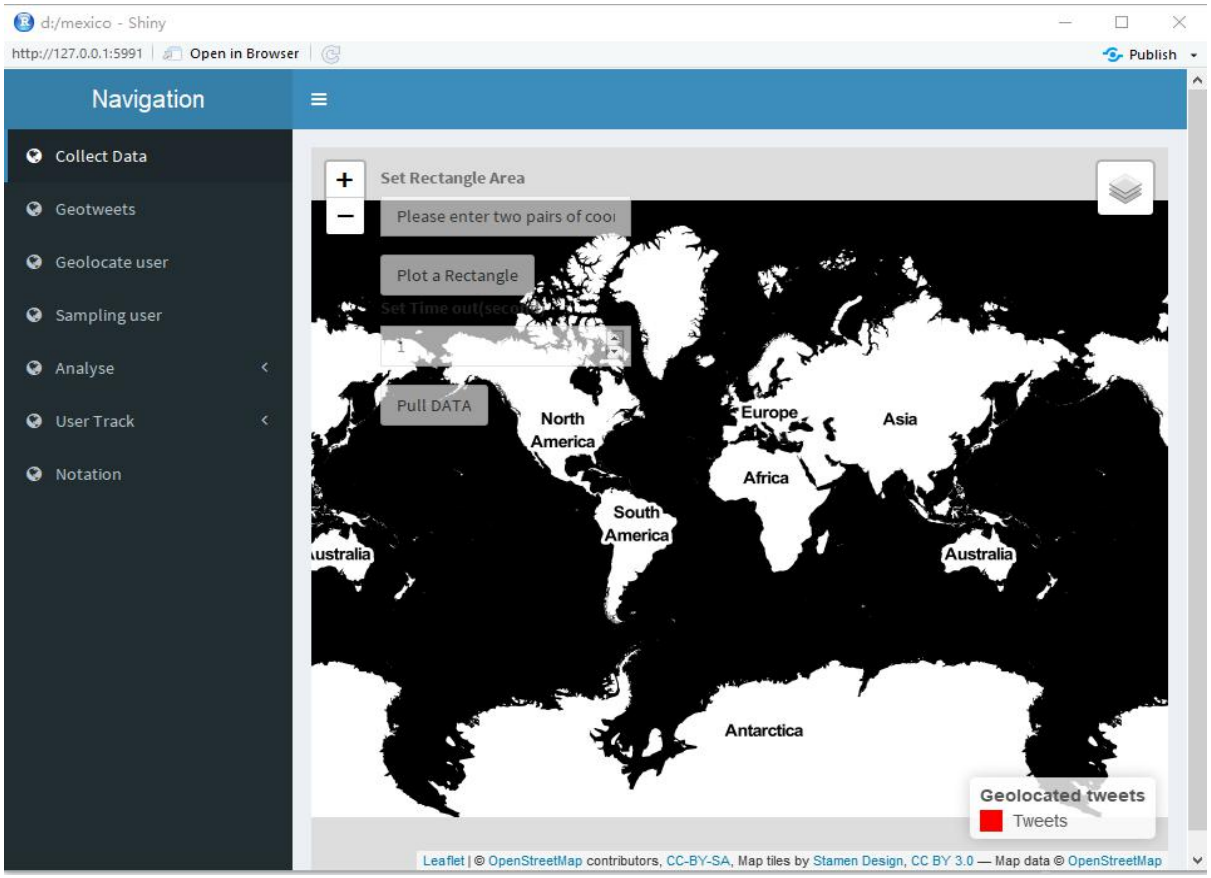


Figure 6.2.2

Fig 6.2.2 is what the application looks like. The left navigation has seven items: Collect Data, Geotweets, Geolocate user, Sampling user, Analyze, User track, Notation.

You can specify a special area and a special time fragment to collect tweet data from Twitter Streaming API using this application, you should just set two kinds of parameters.

1.Two pairs of coordinates of diagonal of a rectangle.

2.Timeout:a specific time fragment, unit is second. Of course, the user can manually pull data from Twitter Streaming API use other means.

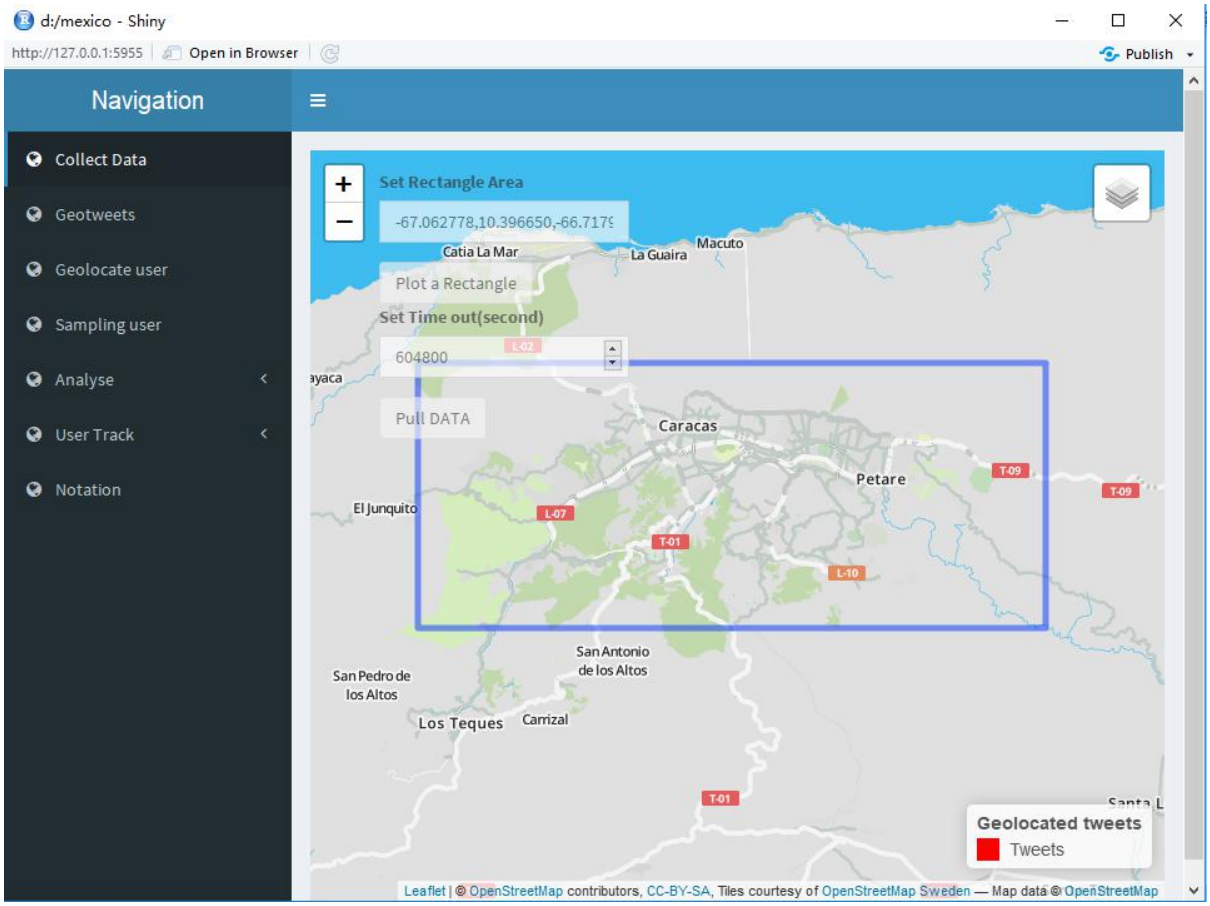


Figure 6.2.3

In Fig 6.2.3 I specific special area is:-67.062778,10.396650,-66.717953,10.540133

Time out is 604800s.

After pull data from Twitter Streaming API, can geolocated tweets and visualization spatial social network..

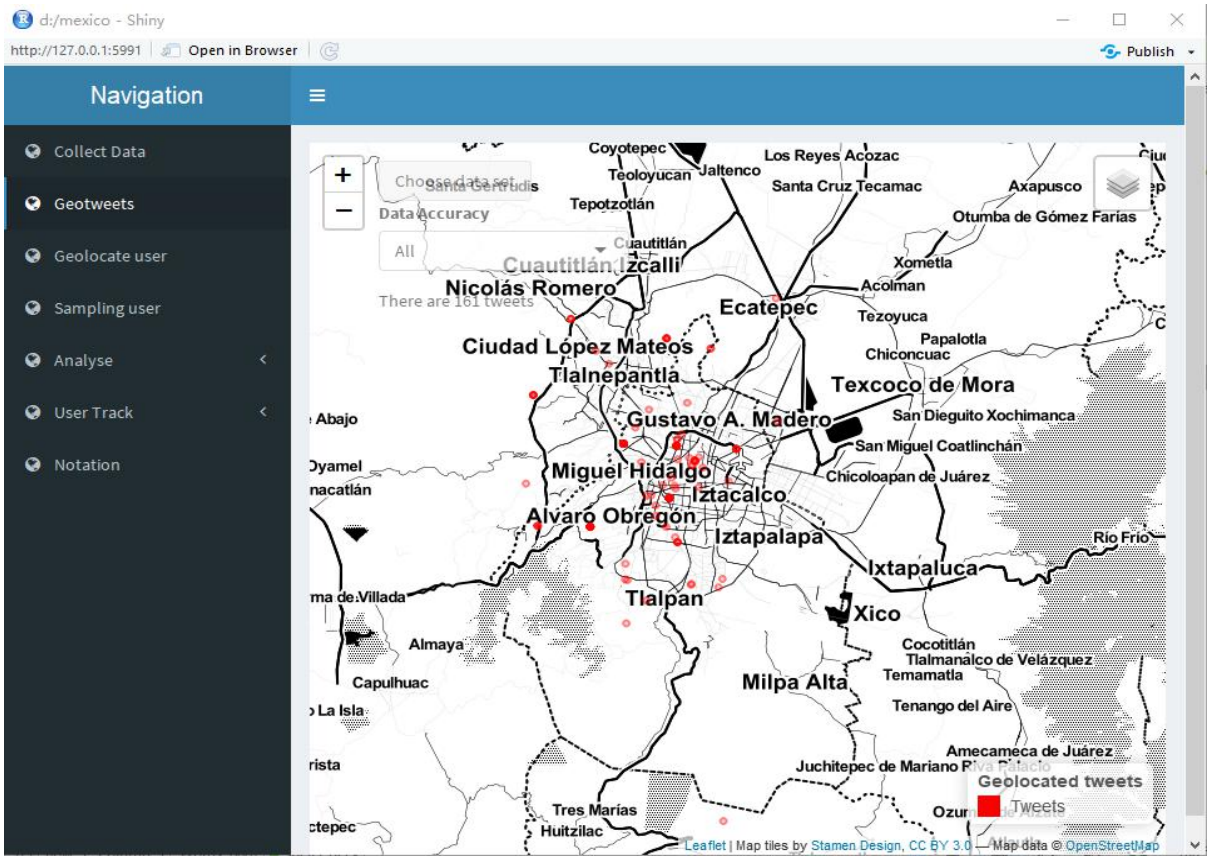


Figure 6.2.4

In Fig 6.2.4, I geotweets ‘All group’ on map.And there are 161 tweets.

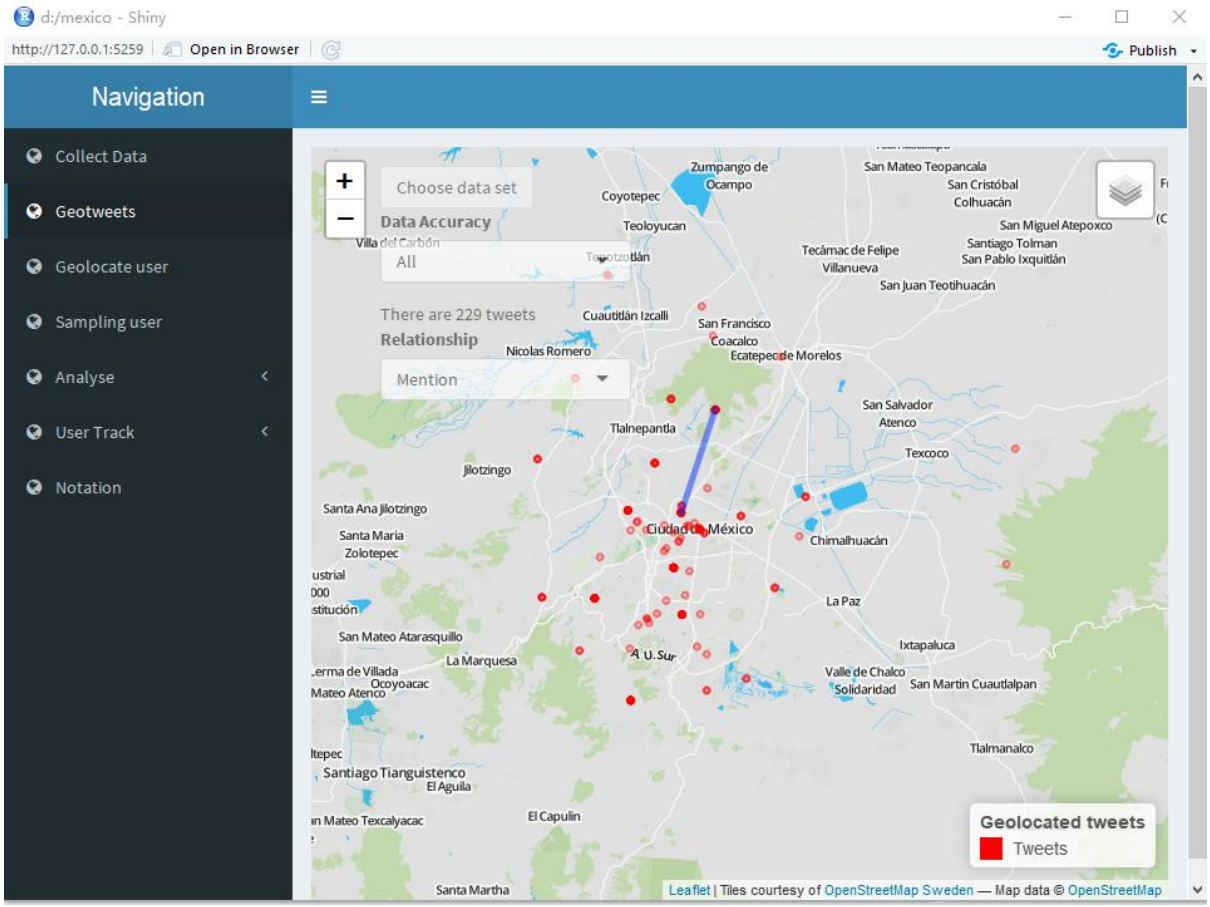


Figure 6.2.5 spatial social network

Fig 6.2.5 is a spatial social network, and it is based on mention relationship.

The next step is to chose the geolocated user, there is a select input like follow:

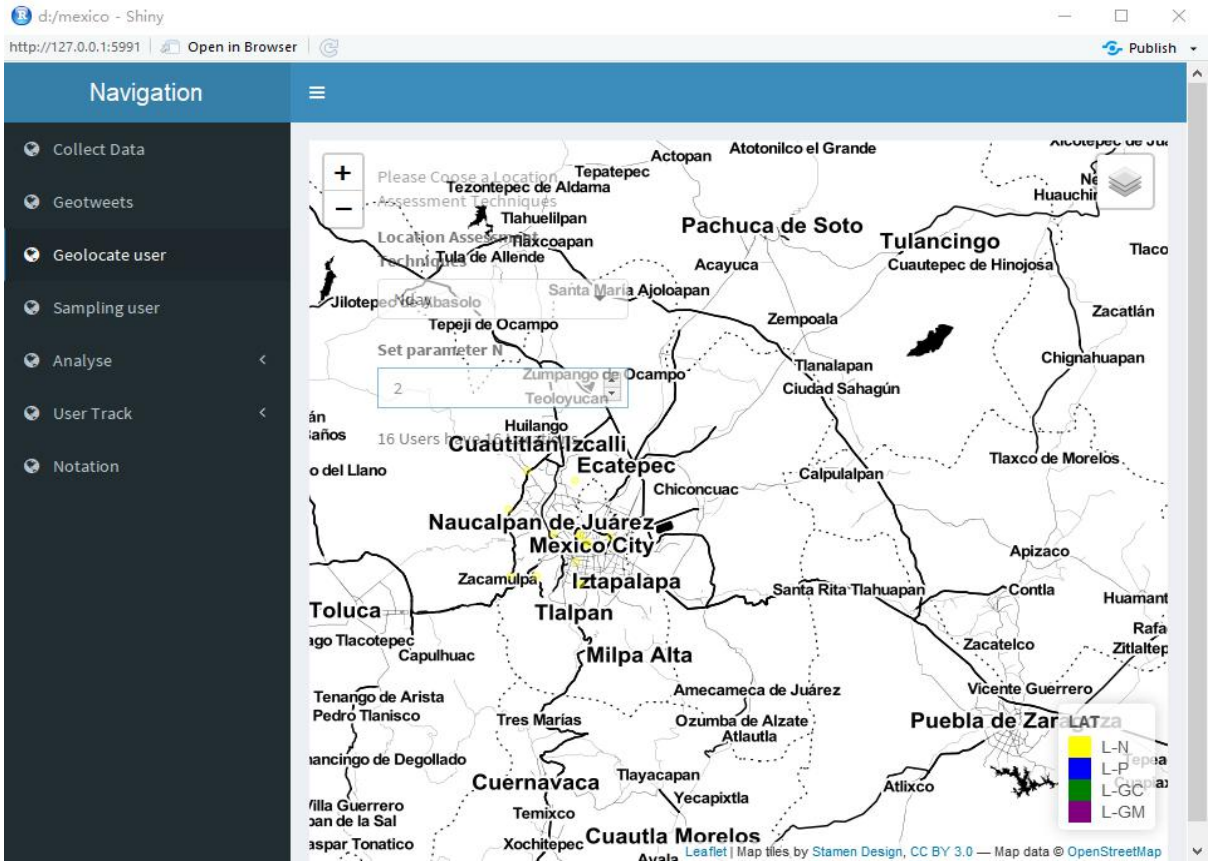


Figure 6.2.6

I choose N-day techniques to geolocate users in Fig 6.2.6. We can see some yellow points on the map and on the bottomright of the map you can see a “legend”, yellow map L-N(N-day), if you choose plurality techniques you will see blue points on the map. You can set value of parameter N.

Now we already geolocate users, the next step is to sample users.

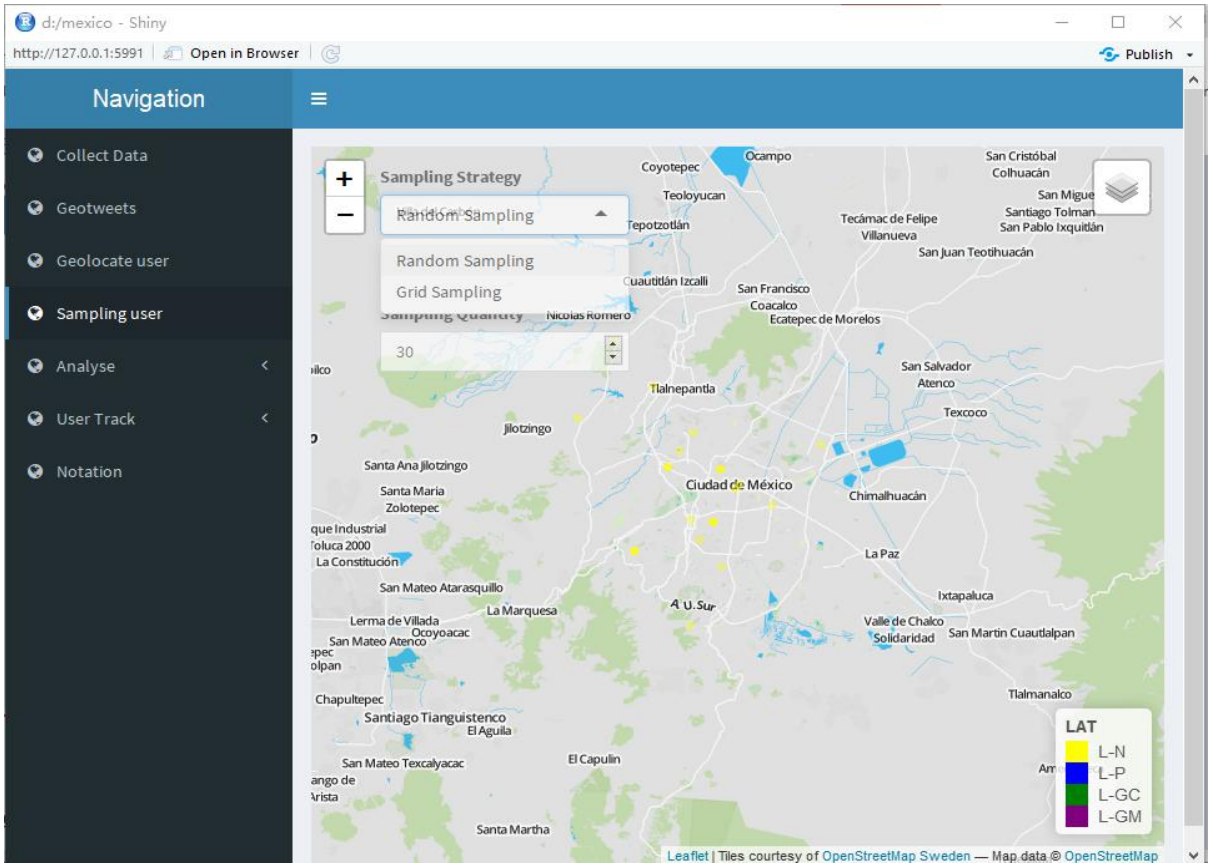


Figure 6.2.7

There are two strategies to sample users see Fig 6.2.7: random sampling and grid sampling. By imputing a number into sampling users, the number is the size of sampling space. I enter 30 and there are 30 yellow points on the map, because last step I choose n-day techniques to geolocate the user.

We have already discovered some user's locations, and can analyze based on these 30 samples of user's location.

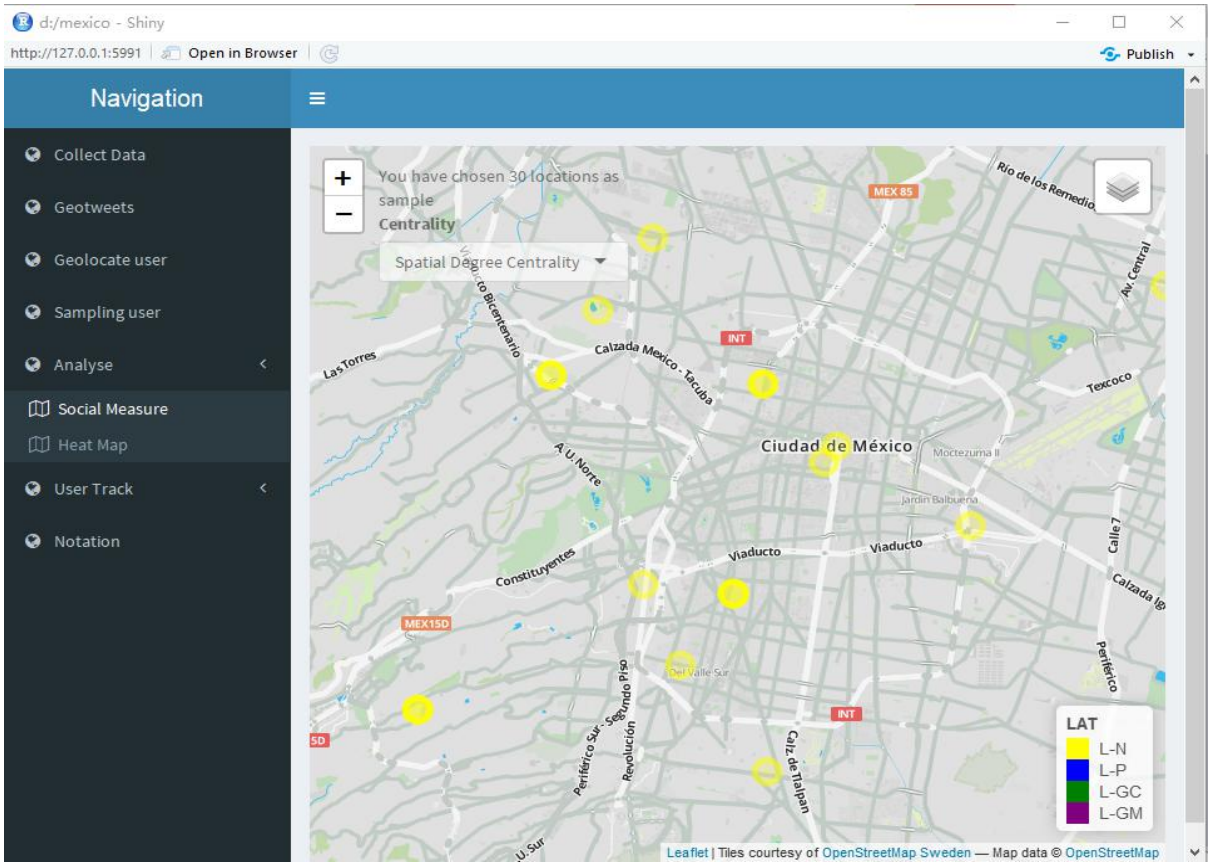


Figure 6.2.8

There are two metrics in Fig 6.2.8: spatial degree centrality and spatial closeness centrality. I chose Spatial Degree Centrality and the color is yellow.

Now you can choose a selection heat map.

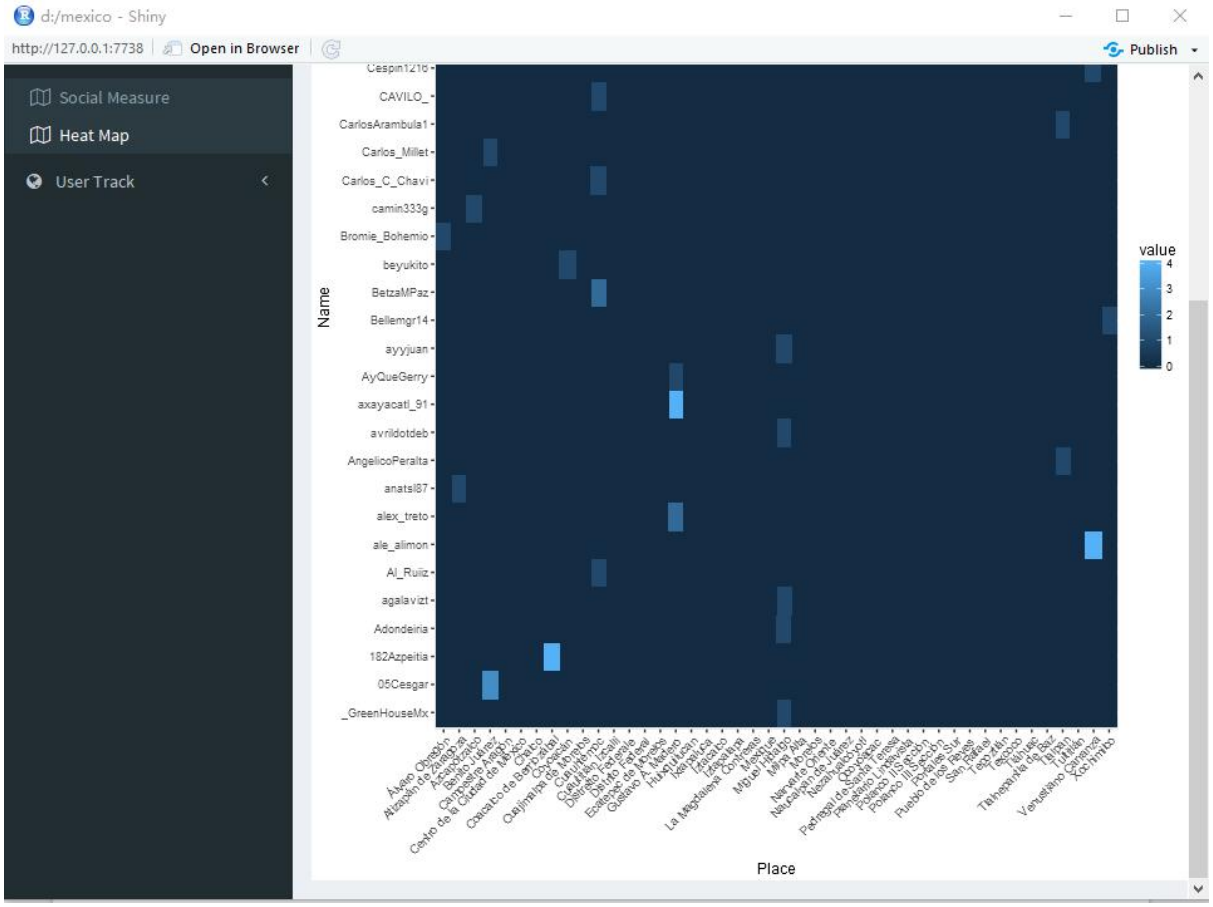


Figure 6.2.9

Fig 6.2.9 is a heat map. It is based on the 30 user's locations.

Futhermore, you will want to track users. To do this, choose track user movement and compare.

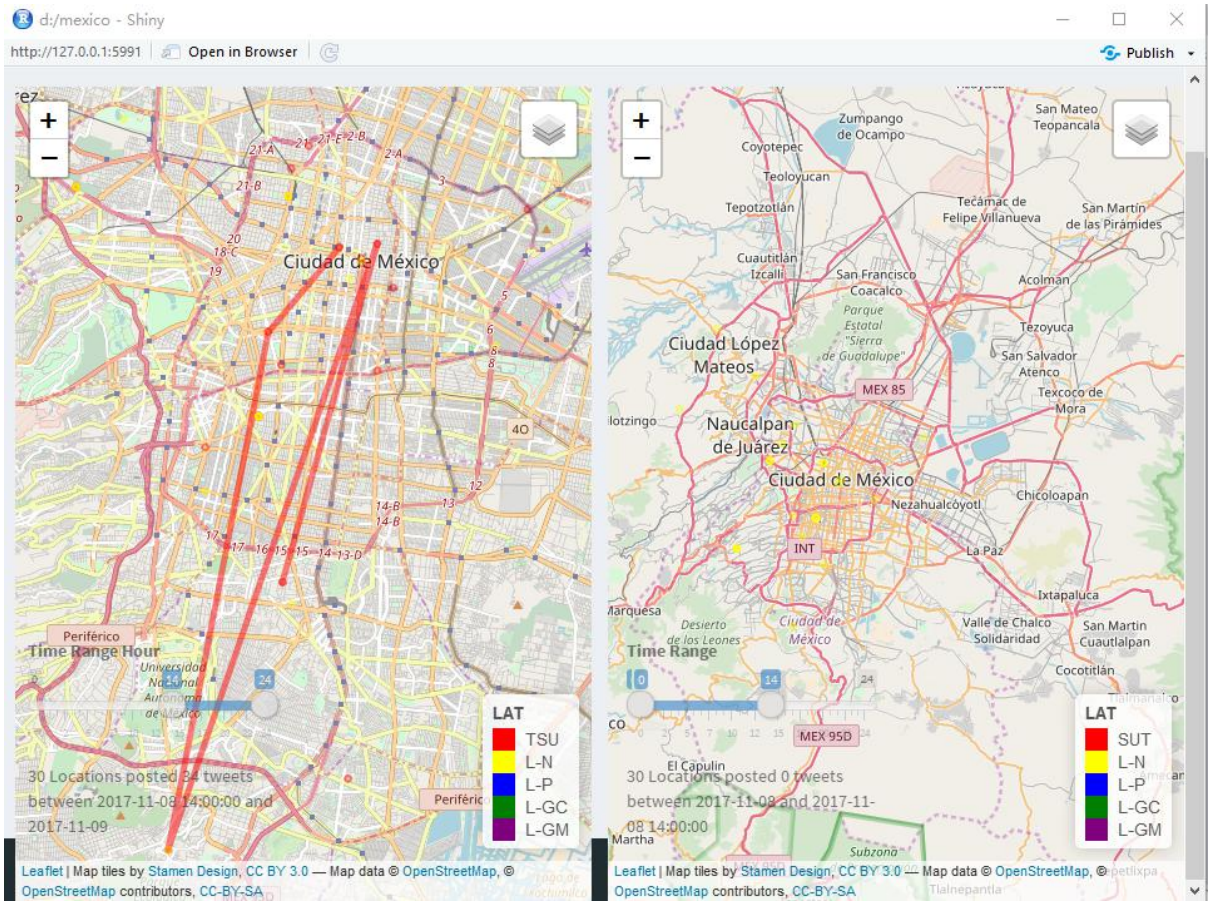


Figure 6.2.10

In this item, you can track users in the specific time fragment and compare them in different time fragments showed in Fig 6.2.10.

You will most likely also want to know the 30 user's recent locations, do this by choosing another item user timeline.

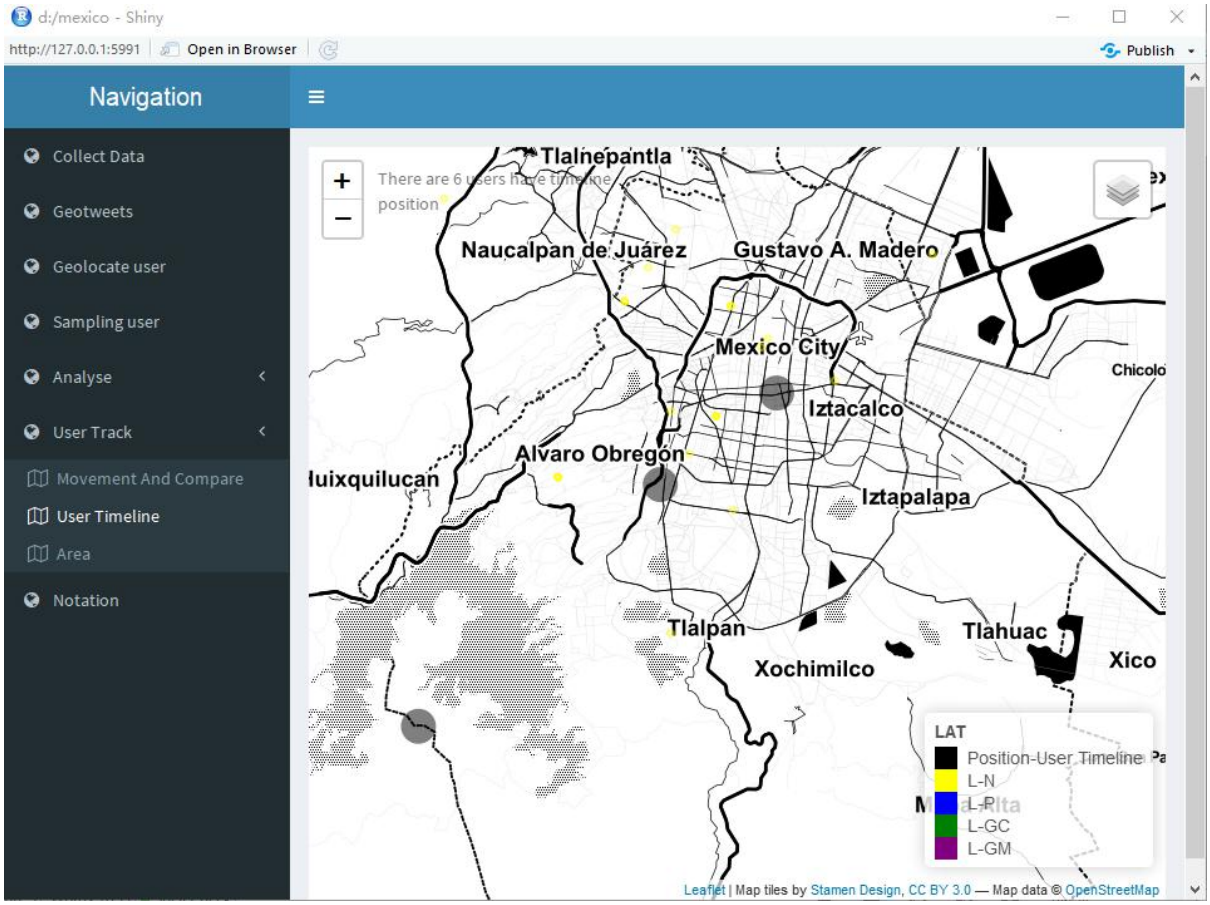


Figure6.2.11

Unfortunately, not all users have recently geolocated. The application reveals that only 6 users have recently geolocated, the black points are the recent user's location see in Fig 6.2.11

You can also specify an area and observe the user's movement pattern in this area in a specific time fragment.

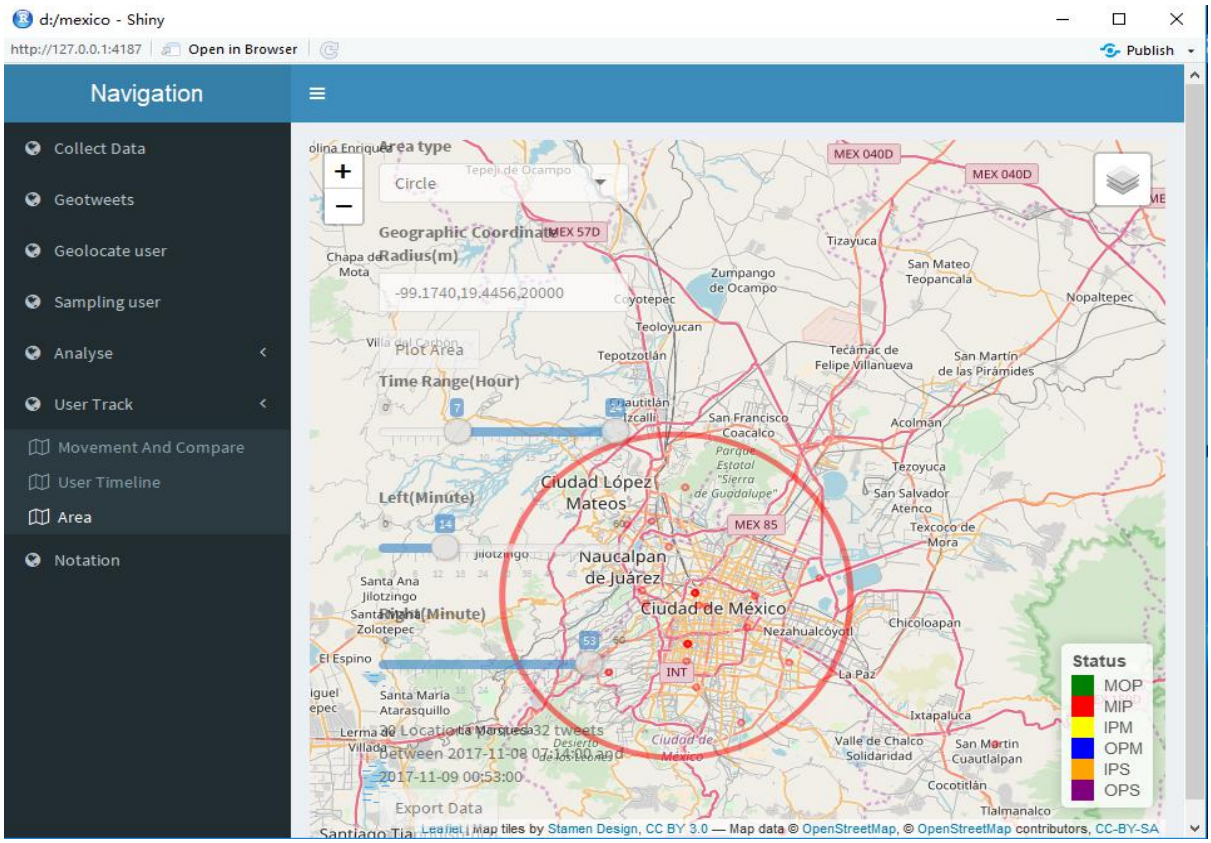
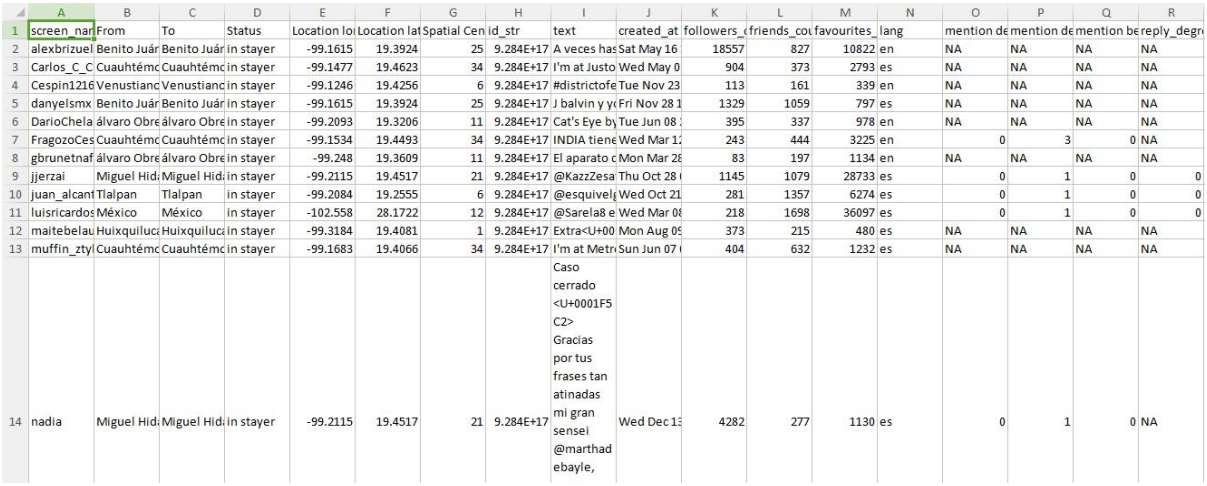


Figure 6.2.12 Circle area

In Fig 6.2.12 I specify a circle as a specific area.The center circle is -99.1740, 19.4456, radius is 2000 meters. And I specify a time fragment from 2017-11-08 07:14:00-2017-1109 00:53:00. You can see some points are in the circle and some points are out of the circle.If you want to know each point status, just tap Export Data.

After tap Export Data, you will find a csv file in your Hard disk root directory.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	screen_nar	From	To	Status	Location lo	Location lat	Spatial Cen	id_str	text	created_at	followers_c	friends_co	favourites_	lang	mention de	mention de	mention be	reply_degr
2	alexbrizuel	Benito Juár	Benito Juár	in stayer	-99.1615	19.3924	25	9.284E+17	A veces has	Sat May 16	18557	827	10822	en	NA	NA	NA	NA
3	Carlos_C_C	Cuauhtémc	Cuauhtémc	in stayer	-99.1477	19.4623	34	9.284E+17	I'm at Justo	Wed May 0	904	373	2793	es	NA	NA	NA	NA
4	Cespin1216	Venustianc	Venustianc	in stayer	-99.1246	19.4256	6	9.284E+17	#districtofe	Tue Nov 23	113	161	339	en	NA	NA	NA	NA
5	danyelsmx	Benito Juár	Benito Juár	in stayer	-99.1615	19.3924	25	9.284E+17	J balvin y y	Fri Nov 28 1	1329	1059	797	es	NA	NA	NA	NA
6	DarioChela	álvaro Obre	álvaro Obre	in stayer	-99.2093	19.3206	11	9.284E+17	Cat's Eye b	Tue Jun 08	395	337	978	en	NA	NA	NA	NA
7	FragozoCes	Cuauhtémc	Cuauhtémc	in stayer	-99.1534	19.4493	34	9.284E+17	INDIA tiene	Wed Mar 12	243	444	3225	en	0	3	0	NA
8	gbrunetnaf	álvaro Obre	álvaro Obre	in stayer	-99.248	19.3609	11	9.284E+17	El aparato c	Mon Mar 28	83	197	1134	en	NA	NA	NA	NA
9	jjerzai	Miguel Hid	Miguel Hid	in stayer	-99.2115	19.4517	21	9.284E+17	@KazzZesa	Thu Oct 28 (	1145	1079	28733	es	0	1	0	0
10	juan_alcant	Tlalpan	Tlalpan	in stayer	-99.2084	19.2555	6	9.284E+17	@esquivelg	Wed Oct 21	281	1357	6274	es	0	1	0	0
11	luisricardos	México	México	in stayer	-102.558	28.1722	12	9.284E+17	@Sarela8 e	Wed Mar 08	218	1698	36097	es	0	1	0	0
12	maitebelau	Huixquiluc	Huixquiluc	in stayer	-99.3184	19.4081	1	9.284E+17	Extra<U+00	Mon Aug 09	373	215	480	es	NA	NA	NA	NA
13	muffin_zty	Cuauhtémc	Cuauhtémc	in stayer	-99.1683	19.4066	34	9.284E+17	I'm at Metr	Sun Jun 07 (	404	632	1232	es	NA	NA	NA	NA
14	nadia	Miguel Hid	Miguel Hid	in stayer	-99.2115	19.4517	21	9.284E+17	Caso cerrado <U+0001F5 C2> Gracias por tus frases tan atinadas mi gran sensei @marthad ebayle,	Wed Dec 13	4282	277	1130	es	0	1	0	NA

Table 6.2.1 a part of Export Data

Figure 6.2.1 is a CSV file.It is about each point status and other information like the number of friends and centrality based on mention relationship and reply relationship.It is based on specific area and specific time fragment.

‘Columns Name’ is user’s name. ‘Columns From’ means: at the beginning of the specific time fragment,where the user was. ‘Columns To’ means: at the end of the specific time fragment, where the user was. ‘Columns Status’ means: which status the user belongs to

Last item is Notation,it has some explanations of abbreviated and some Precautions.

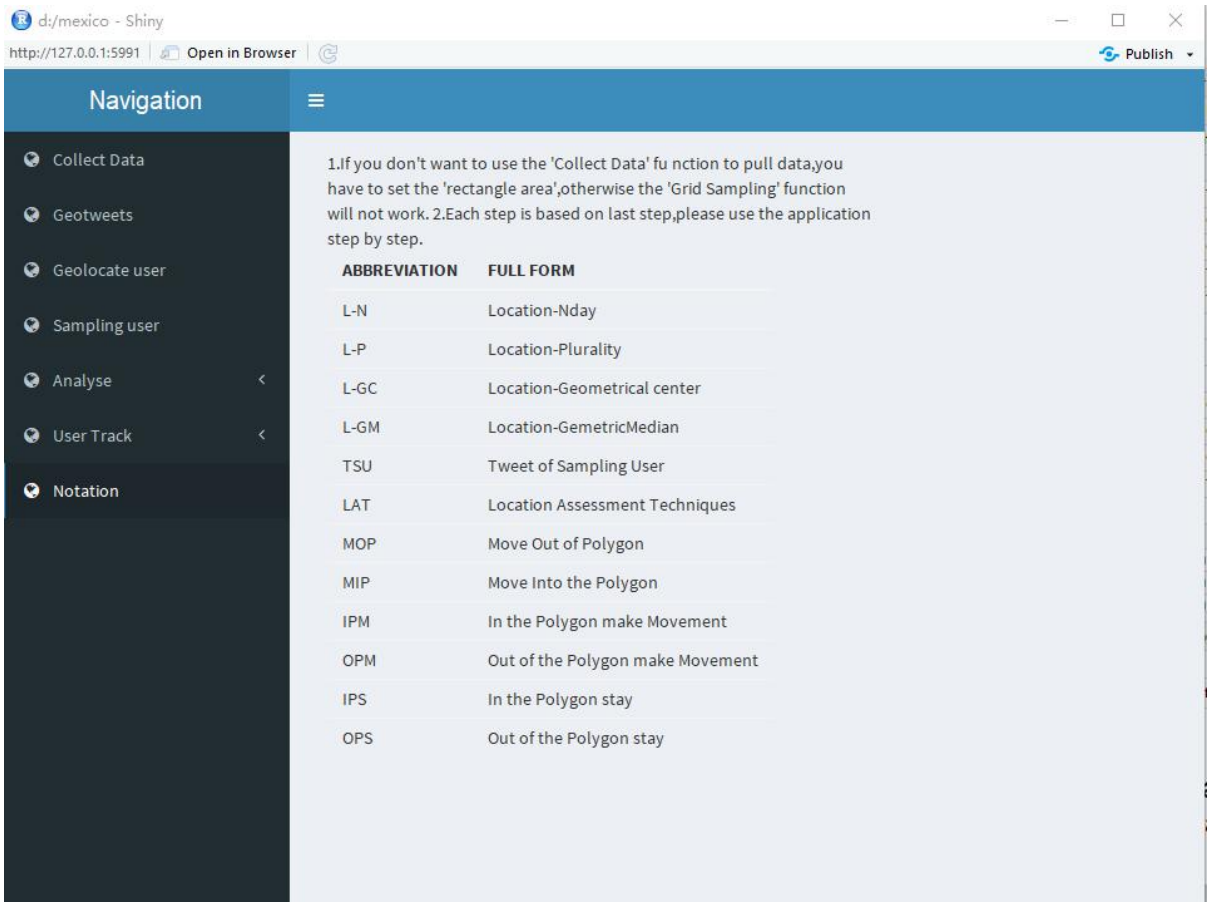


Figure 6.2.13 Notation

References

[1] Morstatter, F. and Liu, H. (2017). Discovering, assessing, and mitigating data bias in social media. Online Social Networks and Media, 1, pp.1-13.

[2]D. Jurgens, T. Finethy, J. McCorriston, Y. Xu and D. Ruths, "Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice", 2015.

[3]"Selection bias", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Selection_bias. [Accessed: 11- May- 2018].

[4] Dictionary of Cancer Terms → selection bias. Retrieved on September 23, 2009.

[5] M. Silk, "The next steps in the study of missing individuals in networks: a comment on Smith et al. (2017)", Social Networks, vol. 52, pp. 37-41, 2018.

[6]J. Smith, J. Moody and J. Morgan, "Network sampling coverage II: The effect of non-random missing data on network measurement", Social Networks, vol. 48, pp. 78-99, 2017.

[7] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. ICWSM

[8]Kariryaa, Ankit et al. “Defining and Predicting the Localness of Volunteered Geographic Information using Ground Truth Data.” CHI (2018).

[9]M. Manca, L. Boratto, V. Morell Roman, O. Martori i Gallissà and A. Kaltenbrunner, "Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study", Online Social Networks and Media, vol. 1, pp. 56-69, 2017.

[10]Lima, Antônio M. G. de and Mirco Musolesi. “Spatial dissemination metrics for location-based social networks.” UbiComp (2012).

[11]"Geo objects", Developer.twitter.com, 2018. [Online]. Available: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects. [Accessed: 29- May- 2018].

[12]"Requirements analysis", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Requirements_analysis. [Accessed: 29- May- 2018].

[13]"R: What is R?", R-project.org, 2018. [Online]. Available: https://www.r-project.org/about.html. [Accessed: 29- May- 2018].

[14]"Shiny", Shiny.rstudio.com, 2018. [Online]. Available: https://shiny.rstudio.com/. [Accessed: 29- May- 2018].

[15]"Leaflet — an open-source JavaScript library for interactive maps", Leafletjs.com, 2018. [Online]. Available: https://leafletjs.com/. [Accessed: 29- May- 2018].

[16]"R: Merge Two Data Frames", Stat.ethz.ch, 2018. [Online]. Available: https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html. [Accessed: 29- May- 2018].

[17]"subset function | R Documentation", Rdocumentation.org, 2018. [Online]. Available: https://www.rdocumentation.org/packages/base/versions/3.5.0/topics/subset. [Accessed: 29- May- 2018].
[18]"R: Data Frames", Stat.ethz.ch, 2018. [Online]. Available: https://stat.ethz.ch/R-manual/R-devel/library/base/html/data.frame.html. [Accessed: 29- May- 2018].
[19]"table function | R Documentation", Rdocumentation.org, 2018. [Online]. Available: https://www.rdocumentation.org/packages/base/versions/3.5.0/topics/table. [Accessed: 29- May- 2018].
[20]R. Plant, Spatial data analysis in ecology and agriculture using R. Boca Raton, Fla.: CRC Press, 2012.
[21]"Shiny - Slider Bar and Slider Range", Shiny.rstudio.com, 2018. [Online]. Available: http://shiny.rstudio.com/gallery/slider-bar-and-slider-range.html. [Accessed: 29- May- 2018].
[22]"Heat map", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Heat_map. [Accessed: 29- May- 2018].
[23]Cran.r-project.org, 2018. [Online]. Available: https://cran.r-project.org/web/packages/tweet2r/tweet2r.pdf. [Accessed: 29- May- 2018].
[24]"stream_tweets function | R Documentation", Rdocumentation.org, 2018. [Online]. Available: https://www.rdocumentation.org/packages/rtweet/versions/0.6.0/topics/stream_tweets. [Accessed: 05- Jun- 2018].
[25]Johnson, I.L., Sengupta, S., Schöning, J., & Hecht, B.J. (2016). The Geography and Importance of Localness in Geotagged Social Media. CHI.