AUSTRALIAN JOURNAL OF BASIC AND APPLIED SCIENCES

ISSN:1991-8178 EISSN: 2309-8414
Journal home page: www.ajbasweb.com

Real Time Prediction of Twitter users location on Google map using Python

M. Vadivukarassi, P. Aruna and N. Puviarasan

Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu, India.

Address For Correspondence:
M. Vadivukarassi, Department of Computer Science and Engineering, Annamalai University, Chidambaram -608002, Tamil Nadu, India.
E-mail: vadivume28@gmail.com

ARTICLE INFO

Article history:
Received 26 April 2016
Accepted 21 July 2016
Published 30 July 2016

Keywords:
Geolocation; Python; Social media; Twitter; Time zone; Visualization.

ABSTRACT

Social media has become a major platform for information sharing. Twitter is a prime example of social media in which researchers can verify their hypotheses, and practitioners can mine interesting patterns and build real world applications. Twitter is a fabulous source of information for many–to–many crisis communication. Python is used for implementation. These Twitter tweets are analyzed based on the keyword search and the list of items is displayed. Then, the collected twitter items are stored in the database. Finally, the top 10 twitter data are detected using attributes and the top real world events are visually placed on the online Google map. This visualization enables us to discover and understand the events easily. It is the best way to know what people are doing at every moment. So, at any instances in our life, we can gather lot of information about the world.

INTRODUCTION

Data mining plays a crucial role in extracting useful information from social media. The reason is, it contains personal trivial data which is not very enlightening or useful to a large group of people. It is used in many areas for analysis. Companies and organizations can perform sentiment analysis for their products and services (Agarwal, A., *et al*., 2011; Kouloumpis, E., *et al*., 2011). It can also help in detecting and predicting disasters (Sakaki, T., *et al*., 2010) and events such as influenza (Achrekar, H., *et al*., 2011). Social media has become a very important tool to stay in touch with friends, to market products and services offered by companies and even to make announcements by government agencies and news channels. One of the social networking sites which has gained vast popularity is Twitter. This research work deals with the data obtained from Twitter which is mined for getting useful information for a real-world scenario.

Twitter and its Importance:

Twitter as a successful technology platform that has grown virally and become "all the rage," given its ability to satisfy some fundamental human desires relating to communication, curiosity, and the self-organizing behavior that has emerged from its chaotic network dynamics. Twitter is an online social network (OSN) used by millions of people all over the world. It enables people to stay connected with their friends, family and colleagues. With advancement of technology, it has become easier to access Twitter using mobile devices like iPhones and iPads. It enables its users to post messages which are 140 characters or less which are called tweets. Users can also retweet messages, which is posting the message posted by other users. This can be thought of as email forwarding. These tweets can be displayed to all users or only to the people following the user. A user can follow other users but it is not necessary for the user who is being followed to follow back. This makes the links

To Cite This Article: M. Vadivukarassi, P. Aruna and N. Puviarasan., Real Time Prediction of Twitter users location on Google map using Python. ***Aust. J. Basic & Appl. Sci.,*** *10(12): 91-97, 2016*

in Twitter directed. Currently, Twitter has 288 million monthly active users with an average of 500 million tweets being sent per day ("About,"18January2016).

Twitter has become an important resource for the field of data mining because of its many features. It has a varied variety of users which can represent a sample of the entire population. The revolution of information and communication technology (ICT) has made it possible for billions of people to access social networking sites ensuring that they have a wide reach of people. They can post messages on the go which ensures that the real-time nature of the messages. Compared to emails, this "push" of information is almost instantaneous. Users also have the freedom to follow or join groups that they like. It also caters for security for its users, where they can decide to post tweets publicly or privately. If they decide their tweets to be public, then they can be viewed by anybody in the Twitter network. However, if they are private, then only the people in the user's network can view them. Mostly, people post about their trivial personal experiences but sometimes they post messages which can contain information which will be valuable on mining. This information can be about events like politics, traffic jams, riots, fires, earthquakes, storms, etc.,

Therefore, Twitter can also act as a non-traditional medium to obtain news as people can tweet information which is newsworthy. They can even create messages as news which can be used in early warning detection systems. However, the most important feature for this study is the real time nature of the information dissipation in the Twitter network. It further becomes useful when 80 per cent of the users are mobile users ("About,"18January2016) which can provide us with exact geo-location and more up-to-date information. These users may post several times a day contrary to bloggers who post once every few days. In 2011, when tsunami hit Japan, Twitter was used as the means of communication when traditional modes of communication went down (Taylor, C., 2011).

Literature Survey:

Sakaki *et al*. have mined Twitter data for real-time earthquake detection (Sakaki, T., *et al*., 2010). They created an application for earthquake reporting system in Japan. This system consists of two parts – event detection and the probabilistic spatiotemporal model of the event. The detection is performed by making a classifier using a Support Vector Machine. Avvenuti *et al* proposed a novel architecture for an early warning system and validated it with an implementation in (Avvenuti, M., *et al*., 2014). They made use of social sensing where a group of people or a community provides similar information that might be obtained from a single sensor. They used Streaming API of Twitter for up-to-date tweets. Events are detected by temporal and spatial analysis. For temporal analysis, they created a novel burst detection method which observes a peak of the number of messages in a time window. They extracted location from the content of the tweet for spatial analysis using TagMe (Ferragina, P. and U. Scaiella, 2012).

Damage assessment was done by using a bigger set of general keywords, images and videos. The results obtained from the experiment were checked with official data to show that earthquakes with a magnitude equal or greater than 3.5 on richter scale can be timely detected with 10 per cent false positives. One of the methods was suggested by Nguyen *et al*. (Nguyen, T.-M., *et al*., 2011). They built an earthquake semantic network using human activity on Twitter based on Web Ontology Language. A Twitter activity was defined by five attributes, namely – action, object, location, time and actor. The network is connected by the relationships – Next and Because Of. They also created automatic data for the network. This network was further used to recommend suitable actions in the face of a disaster. One of the problems with the works of Banerjee *et al*. (Banerjee, N., *et al*., 2009) was that the list of actions and objects had to be prepared before extraction.

Jodi Blomberg *et al*. have outlined two ways in which law enforcement clients can harness the information contained in social media sites such as Facebook and Twitter. These relatively simple examples are intended to be the basis for thinking about how the enormous amounts of social media data can be collected and analyzed to turn tweets and posts into useful information (Jodi Blomberg, S.A.S., C.O. Denver, 2012). Vinay Kumar Jain *et al*. had gave rapid information in various situations like symptoms corresponding swine flu, prevention techniques used by the user and awareness about the medicine and labs, but mostly can give timely information to government and health agencies (Vinay Kumar Jain, Shishir Kumar, 2015).

proposed System:

The objective of the work is to analyze and visualize the top 10 twitter user locations based on the keyword using Google map. In the existing system, the twitter data are analyzed and extracted the relevant attributes based on the keyword. But in this proposed work, the visualization process is implemented using Google map. This visualization enables us to discover and understand the events easily. It is the best way to know what people are doing at every moment. So, at any instances in our life, we can gather lot of information about the world. In order to predict the relevant attributes, the following steps are used to analyzed.

1. Collection of data.
   Collection of historical tweets based on the keywords for specific times and events of interest.
2. Data analyze and extraction.

Extraction and analyzing the tweets for patterns and anomalies to track in the future.

3. Visualization of data.

Visualizing the real time tweet stream for items of interest in Google map.

This approach represents the details of the experiment to trace the trajectory of a keyword and use it for issuing the forewarning to the susceptible people. The details of the research work are explained in the figure 1 below.

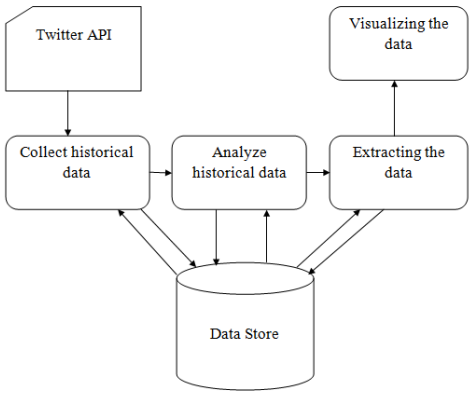


Fig. 1: System Overview

Collection of Historical Data:

Twitter has provided developers with two APIs – Twitter Search API and Twitter Streaming API. The Search API allows users to query against the indices of recent or popular tweets. However the search API does not g*et al*l the tweets but captures the only the relevant ones. Generally the tweets from the past week are extracted and the rate limits of tweets are 180 requests query per 15 minutes. In contrast to the Search API, the Streaming API can provide the user with all tweets for maintaining the completeness of the dataset. It needs a persistent connection open for streaming. The main benefit of this API is getting the real-time stream of tweets. However, this API cannot obtain the tweets that were published before opening the connection. It requires authentication which has been mandated by Twitter for its entire APIs. However, since it uses the Search API, there are some limitations. It can over-represent the more influential users which might lead to some bias in the data.

Table 1: Tweet Features Of The Database

FIELD NAME	MEANING
Tweet_id	A unique number for the tweet
Lso_language_code	The language of the tweet English is represented by'en'
User_Id	A unique number for the user who posted the tweet
Text	The context of the tweet
Created_at	UTC time when the tweet was created
Time	Local time when the time was created
Geo-coordinates	Latitude and Longitude from where the tweets was published
Entities	Contains details about hashtags,URLs and symbols

Also, the API can access only a subset of all the tweets but obtained a large number for performing the experiments. Streaming API would have given a more complete dataset which would have given a more accurate result because the tweets obtained would be in real-time In the table 1, the field name and its meaning are presented. These field name are collected by monitoring the twitter API. The tweets based on the particular keyword are extracted and stored in the database.

Historical Data Analysis:

For predicting the keyword, the extraction of name and location (Time Zone) from the tweets are collected and stored in the database. These two are the most important fields for the experiment of predicting the relevant keywords.. Thus, it is important to do temporal and spatial analysis which mainly deals with getting the time and the most accurate locations from a tweet.Data are analysed incrementally by the system: tweets that have already been analysed are marked, while new tweets will be processed for analytical purposes. This helps us deal with scalability issues in terms of computing power, as tweets that has already been analysed will not be

processed again; the results from each analysis will be aggregated to the results. This also helps us update results in a much predictable and stable manner.

Visualization of top results:

Twitter has a separate API for real time streaming. Twitter monitoring that allows developers to pull public statuses from all users, filtered in various ways such as userid, keyword, and geographic location. However, implementation of a system that monitors Twitter is quite a bit more complicated. Data volumes can quickly increase and Twitter encourages developers to plan for traffic to double every few months. In addition to storage issues, developers should consider how often to "catch" the stream; this decision will vary based on the nature of the request. Geolocation is the process of identifying the geographic location of an object such as a mobile phone or a computer. Twitter allows its users to provide their location when they publish a tweet, in the form of latitude and longitude coordinates. With this information, we are ready to create some nice visualization for our data, in the form of interactive maps.

Experimental Results:

All experiments are conducted on PC with an Intel Core i3- 2350M 2.3 GHz and 4 GB of RAM with Windows 7 as the operating system. To collect data from twitter, the first step is to create a Twitter user account in the website https://apps.twitter.com/ . To Get Twitter API keys, first click "Create New App" and Fill out the form, agree to the terms, and click "Create your Twitter application". Then Keys and Access Tokens are created. They are "API key", "API secret", "Access token" and "Access token secret". The sub-sections discuss the results obtained at each step of the experiment.

Python is used for implementation. Python is a great tool for grabbing data from the web. The scripts used to extract data from twitter are written in Python. The basic steps of using Python are to access the Twitter API, to read and manipulate the data returned and to save the output. Python library is used called as Python Twitter Tools to connect to Twitter API and downloading the data from Twitter. These Python Twitter Tools is simple to use yet fully supports the Twitter API. There are many other libraries in various programming languages such as Twitter API.

Data Collection:

Twitter API is used for collecting the data from twitter. API stands for Application Programming Interface. It is a tool that makes the interaction with computer programs and web services easy. Many web services provide APIs to developers to interact with their services and to access data in programmatic way. Twitter API is used to download tweets related to keyword. The keywords are selected based on the twitter user. If the twitter user would like to know about the some event for every moment, they can analysis in the twitter API and can gather the lot of information about the world. These tweets are stored in CSV format. This format makes it easy to humans to read the data, and for machines to parse it.

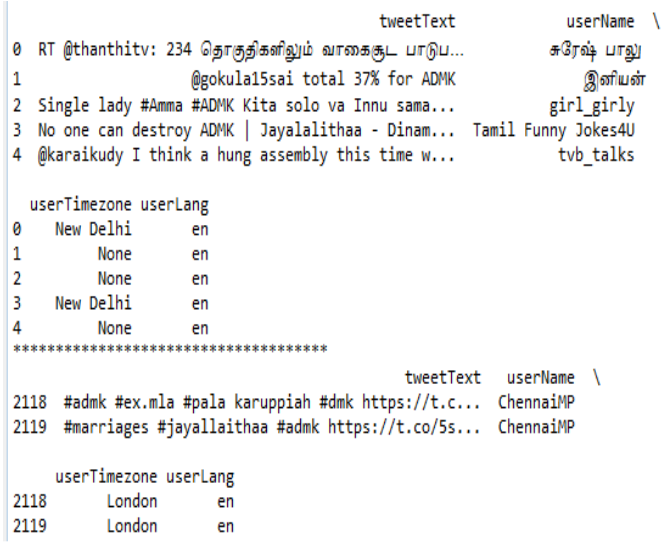
The tweets are collected based on the keyword as in the figure 2 above and it can store in the databse. The 5000 tweets are collected in 15 minutes for a single keyword. The fields that were extracted are – id_str, from_user, text,created_at, time, geo-coordinates, iso_languagecode, to_user_id_str, from_user_ id , source, profile_image_url, status_url and entities_str. These fields are stored in the database. The Twitter users in different metropolitan areas are collected by checking the self-reported locations in their profiles and the Twitter API designed to return the recent or popular tweets in a specified geo-circle defined by latitude, longitude, and radius.

{"created_at":"Thu Dec 17 13:28:05 +0000
2015","id":677480387132559360,"id_str":"677480387132559360","tex
t":"Ruby AdBlocker - Surf the web without ads, Block ads now!...
https:\/\/t.co\/KVt2Wc3i6N","source":"\u003ca href=\"http:\/
\/dlvr.it\" rel=\"nofollow\"\u003edlvr.it\u003c\/a
\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply
_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_
user_id_str":null,"in_reply_to_screen_name":null,"user":
{"id":3033266649,"id_str":"3033266649","name":"T.
H.","screen_name":"th8rt","location":null,"url":null,"descriptio
n":"music
4ever","protected":false,"verified":false,"followers_count":9113
,"friends_count":10015,"listed_count":362,"favourites_count":288
54,"statuses_count":105737,"created_at":"Thu Feb 12 19:53:52 +
0000 2015","utc_offset":-25200,"time_zone":"Mountain Time (US &
Canada)","geo_enabled":false,"lang":"en","contributors_enabled":
false,"is_translator":false,"profile_background_color":"C0DEED",
"profile_background_image_url":"http:\/\/abs.twimg.com\/images
\/themes\/theme1
\/bg.png","profile_background_image_url_https":"https:\/
\/abs.twimg.com\/images\/themes\/theme1
\/bg.png","profile_background_tile":false,"profile_link_color":"
0084B4","profile_sidebar_border_color":"C0DEED","profile_sidebar
_fill_color":"DDEEF6","profile_text_color":"333333","profile_use
_background_image":true,"profile_image_url":"http:\/

Fig. 2: Collection of tweets based on keyword

Extraction and Visualization:

The data are collected using Twitter API and twitter user name, text, time zone and language are extracted. The 5000 tweets are collected based on the keyword (e.g.: ADMK). The twitter users who are posted the tweets about the keyword currently are collected and extracted and stored in the database. Among 5000 tweets, 2120 data are extracted based on the above keyword. The attributes of first 4 and last 2 tweets are displayed are shown below in the figure 3. The extracted tweets are analyzed and the top 10 time zone are predicted based on the twitter user keyword. These top 10 tweets are stored in the database and visualized using Google map (The Google Map Tools", Available: https://batchgeo.com/).



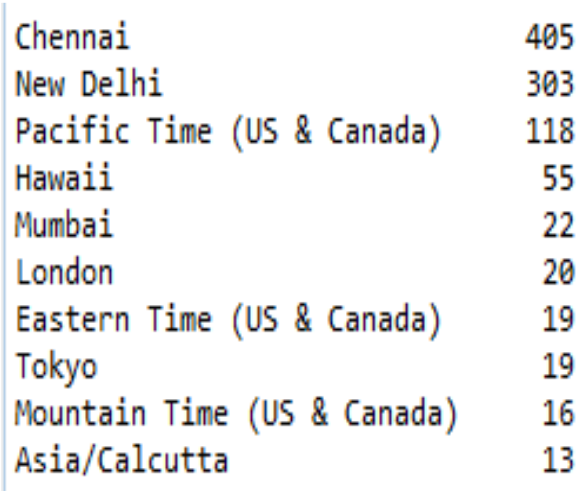
```
                                                tweetText            userName  \
0  RT @thanthitv: 234 தொகுதிகளிலும் வாகைசூட பாடுப...          சுரேஷ் பாலு
1                  @gokula15sai total 37% for ADMK              இனியன்
2  Single lady #Amma #ADMK Kita solo va Innu sama...          girl_girly
3  No one can destroy ADMK | Jayalalithaa - Dinam...  Tamil Funny Jokes4U
4  @karaikudy I think a hung assembly this time w...           tvb_talks

  userTimezone userLang
0    New Delhi       en
1         None       en
2         None       en
3    New Delhi       en
4         None       en
************************************
                                                tweetText   userName  \
2118  #admk #ex.mla #pala karuppiah #dmk https://t.c...  ChennaiMP
2119  #marriages #jayallaithaa #admk https://t.co/5s...  ChennaiMP

     userTimezone userLang
2118       London       en
2119       London       en
```

Fig. 3: Extraction of tweets

Finally, the top 10 results are analyzed and displayed as shown in figure 4 below. The time zones of the twitter user are predicted based on the keyword so that the people can gather the information of that particular event in twitter for every moment easily.



```
Chennai                        405
New Delhi                      303
Pacific Time (US & Canada)     118
Hawaii                          55
Mumbai                          22
London                          20
Eastern Time (US & Canada)      19
Tokyo                           19
Mountain Time (US & Canada)     16
Asia/Calcutta                   13
```

Fig. 4: Top 10 timezone of relevant keyword

The graphical analysis of the top real world time zone is analyzed using the bar chart as given below in the figure 5. It selects tweets based on the location of the keyword vs. the number of the tweets given by the largest blue rectangle on the left side of the histogram.

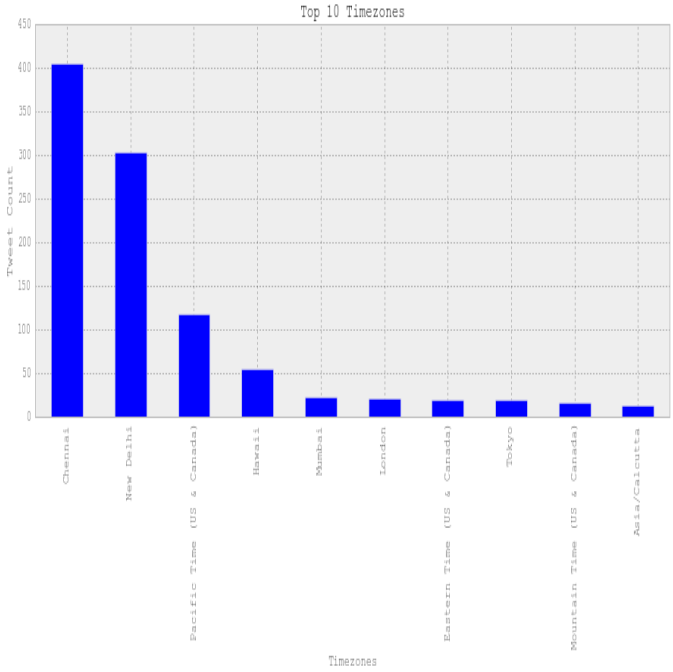


Fig. 5: Graphical analysis of timezone

A more sophisticated approach for spam detection can be used for a real-time system. One of the most important aspects of this experiment is the time and location extraction from the tweets obtained in the data collection stage. This research is implemented a fresh approach of getting the location in three ways – 1) The place where the tweet was posted from, 2) the content of the tweet and 3) the location of the Twitter account. The data from Twitter based on events are collected and analyzed the tweets that contained Time zone. The results from these experiments proved that the insight on the Time zone shared based on events or keywords, and provide the profile of top 10 Time zone results. The top 10 Time zone are displayed in the map as given below in figure 5.

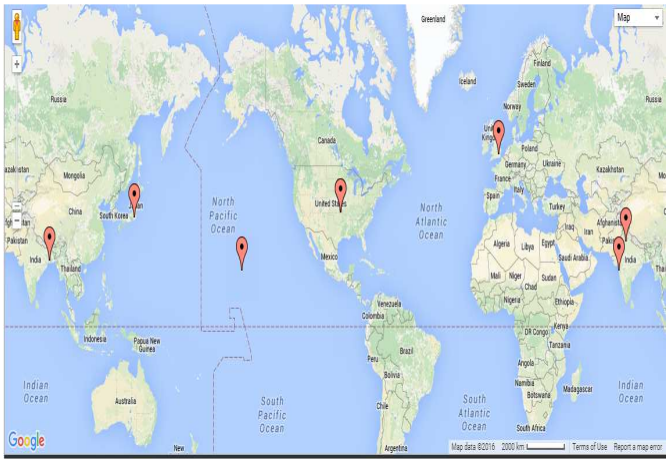


Fig. 5: Visualizing using Google map

Conclusion:

Twitter user keyword plays a major role in information dissipation in today's world of information and technology. It discusses about the running with Twitter's API using Python to interactively explore and analyze Twitter data, and provided some starting templates that you can use for mining tweets.This research work proposes a framework for finding out the attributes based on the keyword for real-time data extracted from Twitter. The steps that were involved are data collection, data analysis and visualization of the top results. The framework was validated by looking at the validation data and comparing the locations obtained by the

extrapolation of the curves. The keywords for data collection were given relevantly which could completely capture the informative points through Twitter API for capturing the tweets as they are posted. In conclusion, this research work was able to prove the validity of the keyword and the top 10 time zones and location are analyzed based on the twitter user keyword and visualizing the results in the Google map from the database.

REFERENCES

Agarwal, A., B. Xie, I. Vovsha, O. Rambow and R. Passonneau, 2011. "Sentiment analysis of Twitter data", LSM '11 Proceedings of the Workshop on Languages in Social Media.

Kouloumpis, E., T. Wilson and J. Moore, 2011. "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona.

Sakaki, T., M. Okazaki and Y. Matsuo, 2010. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors" ,WWW2010, Raleigh.

Achrekar, H., A. Gandhe, R. Lazarus, S.-H. Yu and B. Liu, 2011."Predicting Flu Trends using Twitter Data" ,The First International Workshop on Cyber-Physical Networking Systems.

"About,"18January2016.[Online]. https://about.twitter.com/company

Taylor, C., 2011 "Twitter Users React To Massive Quake, Tsunami In Japan,". [http://mashable.com/2011/03/10/japan-tsunami/.

Avvenuti, M., S. Cresci, M.N.L. Polla, A. Marchetti and M. Tesconi, 2014. "Earthquake Emergency Management by Social Sensing", Pervasive Computing and Communications Workshops (PERCOM Workshops), Budapest.

Ferragina, P. and U. Scaiella, 2012. "Fast and Accurate Annotation of Short Texts with Wikipedia Pages", IEEE Software, 29: 70-75.

Nguyen, T.-M., K. Koshikawa, T. Kawamura, Y. Tahara and A. Ohsuga, 2011. "Building Earthquake Semantic Network by Mining Human Activity from Twitter" , IEEE International Conference on Granular Computing.

Banerjee, N., D. Chakraborty, K. Dasgupta, A. Joshi, S. Mittal, S. Nagar, A. Rai and S. Madan, 2009. "User Interests in Social Media Sites: An Exploration with Micro-blogs", CIKM.

"The Search API," Twitter, https://dev.twitter.com/rest/public/search.

"The Streaming APIs," Twitter, [Online]. Available:https://dev.twitter.com/streaming/overview.

Jodi Blomberg, S.A.S., C.O. Denver, 2012. "Twitter and Facebook Analysis: It's Not Just for Marketing Anymore", Social Media and Networking, SAS Global Forum.

Vinay Kumar Jain, Shishir Kumar, 2015. "An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter", ICECCS, Procedia Computer Science, 70: 801-807.

Manuel Burghardt, 2015. "Visual Introduction to Tools and Methods for the Analysis of Twitter Data", 10plus1: Living Linguistics, Issue 1, Media Linguistics.

"The Google Map Tools", Available: https://batchgeo.com/