

# Final Project - Twitter Adapter Report

**Team Member: Deep Patal, Hongye Gong, Yukun Liang**  
github link: <https://github.com/hongyegong/CS6083>

---

## Introduction:

This is a twitter adapter project with both server side in php and mysql database and front-side user interface by html/css/JS. Basically, we retrieve bunch of data from twitter server in real-time down to our local database and then parse them into our local normalized database. Then we render all data onto the webpage through html/css/js and then analyze sentimental and location information of every single tweet by the php\_nlp bases classifier and google geo coding API.

---

## How to implement?

First, I will give a short workflow of how whole project works. Then I will clarify that which part of coding are responsible for in the this project.

There are Three parts of this project.

The first part is for retrieve data from twitter server to local database. The database is populated in two steps: getting the tweets, and parsing them into multiple tables. It is important to separate these operations, because tweets may be sent by the Twitter API at a very fast rate. If each tweet is parsed and inserted into multiple tables as it is received, the code and database may not be able to keep up with the data flow, and tweets will be lost. My solution is to store the tweets as they are received in a simple cache table without doing any parsing. A separate process does the parsing and storage into separate tables.

The first step of collecting tweets is done by `get_tweets.php`, which is run as a continuous background process. When a new tweet is received the Twitter streaming API, `get_tweets.php` uses `db_lib.php` to insert it into the `json_cache` table. The connection with the Twitter streaming API is maintained by the Phirehose library.

The Twitter streaming API returns data in JSON format. To make the collection process as fast as possible, the entire JSON payload for a single tweet is saved to the database as a single string without any parsing.

A separate background process is run for `parse_tweets.php`, which gets the JSON data for each tweet from the `json_cache` table, parses it into it's component parts, and inserts them into separate table for use by the other modules in the Twitter adapter. Once again, `db_lib.php` is used to manage the MySQL code.

The rest of the modules of this project then are able to rely on this database of tweets and supporting data without having any direct contact with the Twitter API. Once you install the

Twitter database server code, then can build my own Twitter apps that are assured of a real-time source of Twitter data.

The second part is for displaying which is a webpage the users directly interact with. Let's follow the flow of tweet data from web server to browser using the sample Web page included with this code, index.php:

When index.php is opened by a web browser, it uses `require_once()` to call `twitter_display.php` in the plugin directory.

`Twitter_display.php` loads `tweet_list_template.txt`, which defines the HTML structure of the list of tweets. This template has macros that show where to include all the tweet data.

When the macro for the list of recent tweets is reached, `get_tweet_list.php` is called with `require_once()`.

`Get_tweet_list.php` extracts the most recent tweets from the tweets table in the MySQL database, and loops through them. For each set of tweet data it uses `tweet_template.txt` to assemble an HTML version of the tweet.

The text of each tweet is passed to the `linkify()` function. The `linkify()` function in `display_lib.php` converts each entity (@mentions, tags, and URLs) into formatted hyperlinks.

When `get_tweet_list.php` has a complete set of tweets in HTML format, it returns it to `twitter_display.php`, and the list appears in the Web page.

After the page loads, `site.js` starts and sets up parts of the interface, including a More Tweets button at the bottom of the tweet list and a count of new tweets link at the top of the list. It also establishes a refresh interval for calling the Web server with Ajax to get the count of new tweets. Every time the refresh interval is reached, `site.js` calls `get_new_tweet_count.php` with Ajax, gets the count of new tweets, and displays them in the new tweets count.

If the user clicks the More Tweets button, `site.js` requests a set of older tweets from `get_tweet_list.php`. These are returned as HTML, and appended to the end of the tweet list.

All of the interactions with the Twitter database server are done through the `db_lib.php` script in the db directory.

After we retrieve the data from the twitter server with tweeter streaming API and then render the content onto our webpage. The next is to do the data analysis in both sentiment analysis and the geo-location analysis.

We use the Naive Bayes trainer and classifier to gather all the words in positive, negative or neutral and then classify the single sentence based on the probability of being positive or negative. This process including language preprocess and classifying is done by invoke the API which is provided by `php_nlp_tools`.

As far as the geolocation part, we first get the dataset of all cities over the world which contains about 2 or 3 millions records and store them into our database for later detecting if the particular word in a tweet is a location information or not. For the query efficiency we need to create the index for the cityname column in that table we just created since there is too much data there and it's necessary to create such an index for that column. Then we just look up every word in one sentence in database. Suppose we have 100 words in one tweet (which is not always the case) and we have 20 tweets showing on the page. Then we most have 2000 words in one page and the time for each select query is approximately 0.0001 seconds so it won't take above 1 seconds to execute these operations which process all words on the page at the same time.

After we get the location words then simply merge them together and then input the location information into google geo API. Then we will get the latitude and longitude and invoke the google API again, then we can finally show the tweet location on the google map.

That's basically the workflow of the whole project.

Following is what every file specifically dose:

- Twitter Database Server Source Code Files:

Phirehose Library - Phirehose library for capturing tweets from the Twitter streaming API.

config.php - General configuration options for the this Twitter adapter.

db\_config.php - Database configuration options for the Twitter database server.

db\_lib.php - Database library used by the entire Twitter adapter.

db\_test.php - Simple test script for the Twitter database server.

get\_tweets.php - Gather tweets in real-time using the Twitter streaming API.

monitor\_tweets.php - Run as cron job that reports errors by email if the tweet collection fails.

parse\_tweets.php - Parse tweets into a normalized database schema.

- Twitter Display Source Code Files:

lib - php\_Insight library for sentiment analysis using Naive Bayes classifier algorithm.

NlpTools - Another tools for natural language processing library used for tokenizing sentence.

default.css - CSS style sheet for the Twitter display plugin.

get\_new\_tweet\_count.php - Return the count of new tweets as text/HTML.

get\_tweet\_list.php - Return the most recent tweets as HTML.

index.html - Example of a Web page using the Twitter display plugin.

twitter\_display.php - Main module of the Twitter display plugin.

display\_lib.php - Functions used to display tweets: linkify() and twitter\_time().

site.js - Use Ajax to refresh the new tweet count and load more tweets.

tweet\_list\_template.txt - Text template for formatting a complete list of tweets on a Web page.

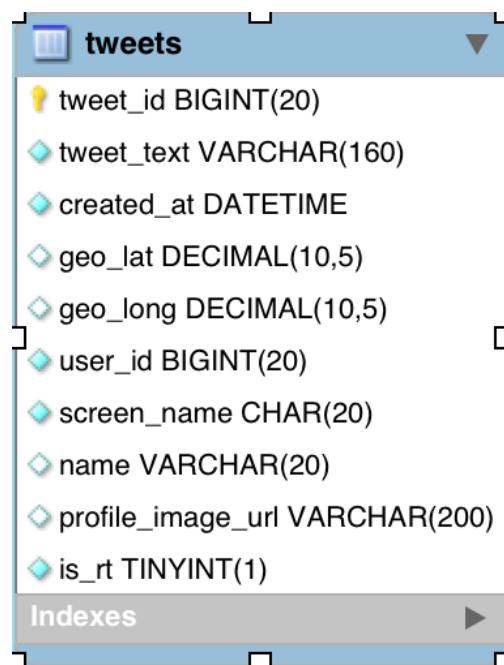
tweet\_template.txt - Text template for formatting a single tweet.

twitter\_display\_config.php - Configuration options for the Twitter display plugin.

---

Database tables:

**Tweets table:**



The image shows a screenshot of a database table structure for a table named 'tweets'. The table has the following columns and data types:

Column Name	Data Type
tweet_id	BIGINT(20)
tweet_text	VARCHAR(160)
created_at	DATETIME
geo_lat	DECIMAL(10,5)
geo_long	DECIMAL(10,5)
user_id	BIGINT(20)
screen_name	CHAR(20)
name	VARCHAR(20)
profile_image_url	VARCHAR(200)
is_rt	TINYINT(1)

Below the columns, there is a section labeled 'Indexes' with a right-pointing arrow, indicating that there are indexes defined for this table.

<

## Twitter user table:

users	
🔑	user_id BIGINT(20)
💎	screen_name VARCHAR(20)
💎	name VARCHAR(20)
💎	profile_image_url VARCHAR(200)
💎	location VARCHAR(30)
💎	url VARCHAR(200)
💎	description VARCHAR(200)
💎	created_at DATETIME
💎	followers_count INT(10)
💎	friends_count INT(10)
💎	statuses_count INT(10)
💎	time_zone VARCHAR(40)
💎	last_update TIMESTAMP
Indexes	

1

SELECT \* FROM twitter.users;

100%

1:1

Result Grid

Filter Rows:

Q Search

Edit:

Export/Import:

user_id	screen_name	name	profile_image_url	location	url	description
3140889318	AndrewBedsole	Andrew Bedsole	http://pbs.twimg.com/profile_images/585434914...	NYC		M.A. candidate in
192014399	B_Hernandez23	Brian Hernandez	http://pbs.twimg.com/profile_images/551982492...	Pennsylvania		Tech Junkie; At ho
5797332	halans	Jean-Jacques Halans	http://pbs.twimg.com/profile_images/459102742...	Sydney, Australia	http://about.me/halans	Belgian Waffle Co
2851760021	ooyuztechnology	OYUZ Technology	http://pbs.twimg.com/profile_images/527884782...		http://www.ooyuz.com/newsarticles?term=techn...	Technoogy news f
522052547	elbo913	Eli Bobo	http://pbs.twimg.com/profile_images/268218407...	New York		
23558652	jonathankoren	jonathankoren	http://pbs.twimg.com/profile_images/519718905...	Via Lactea	http://www.jonathankoren.com	The Real Jonathan
19190989	TripCase	TripCase	http://pbs.twimg.com/profile_images/575056494...	Wherever you travel	http://travel.tripcase.com	The place where r
54033121	adamsherwin	Adam Sherwin	http://pbs.twimg.com/profile_images/372409801...	Brooklyn, NY	http://about.me/sherwinadam	Photo-Video Guy,
316010030	undavorojo	King Of Excuses@	http://pbs.twimg.com/profile_images/592101187...	'Dile Pene a la vida'	http://undavorojo.blogspot.com.es/	'Arise and be... all
64489851	alexrothstein	ruph loren	http://pbs.twimg.com/profile_images/591061626...	alief		finesse the world
3165252706	alvaroo_navarro	Álvaro Navarro	http://pbs.twimg.com/profile_images/587698295...	Madrid		FOLLOW BACK #
128817127	Microspecialist	Adnan Hendricks	http://pbs.twimg.com/profile_images/458972035...	The Netherlands,...	http://microspecialist.wordpress.com	Microsoft Speciali
258810247	TheTranshuman	Frank Rummel	http://pbs.twimg.com/profile_images/594540852...	The #Singularity (#...	http://google.com/+frankrummel	WE ARE #TRANS
63636163	colincornaby	Colin Cornaby	http://pbs.twimg.com/profile_images/431955231...	Portland, OR		I do application de
2296020547	TechGiantNews	Tech Giant News	http://pbs.twimg.com/profile_images/424213779...		http://www.techgiantnews.com	Get latest technol
36486544	MichaelATerry	Michael Terry	http://pbs.twimg.com/profile_images/425479131...	Centerton, AR	http://www.typesandshadows.net	Predestined, calle
3105433974	AmilluxLLC	Amillux, LLC	http://pbs.twimg.com/profile_images/593166448...	Texas	http://www.amillux.com	Picking ourselves
1961671051	patrickmac888	patrick macalalad	http://pbs.twimg.com/profile_images/378800000...			I AM THE B E S T
2812076916	gabbydrew24	Follow me JB :( 3	http://pbs.twimg.com/profile_images/571960184...			
2455434540	Luz_Espinal54	Luz Espinal	http://pbs.twimg.com/profile_images/458644539...	Puerto Plata		
2460091578	1PatriciaOrtega	Patricia Ortega	http://pbs.twimg.com/profile_images/459952370...	San Juan		Crossfitter, Fan, T
741168883	teamtivity	Tivit Me!	http://pbs.twimg.com/profile_images/450984245...	Silicon Valley	http://www.tivit.me	The simplest way
337680665	d9yshox	Antoine Garrison	http://pbs.twimg.com/profile_images/144845904...	Oakland		Professional writer
600202020	QuibMacK	QuibMacK	http://pbs.twimg.com/profile_images/500000000...	Durango	http://fb.com/quibmac	Just how serious
users 1						
						Apply Revert

Result Grid

Form Editor

Field Types

Query Stats

Execution Plan

Action Output

	Time	Action	Response	Duration / Fetch Time
✓ 1	14:42:01	SELECT * FROM twitter.tweets LIMIT 0, 50000	49523 row(s) returned	0.0033 sec / 0.127 sec
✓ 2	14:42:54	SELECT * FROM twitter.users LIMIT 0, 50000	41333 row(s) returned	0.0033 sec / 0.100 sec

## Tweet Hashtag table:

tweet_tags	
tweet_id	BIGINT(20)
tag	VARCHAR(100)
Indexes	

1

SELECT \* FROM twitter.tweet\_tags;

100%

1:1

Result Grid

Filter Rows:

Q Search

Edit:

tweet_id	tag	
594635036726550531	apple	
594635036726550531	tech	
594635054992596992	nwo	
594635074777313280	TECNOLOGÍA	
594635121535393792	outofprocessinghell	
594635121535393792	fb	
594635137070956544	unifydock	
594635137070956544	AppleWatch	
594635177587965952	FreeAppleWatch	
594635227785502721	products	
594635251084873731	apps	
594635251084873731	mobile	
594635251084873731	technology	
594635324547993603	luxury	
594635324547993603	AppleWatch	
594635324547993603	unifydock	
594635498934546432	updates	
594635500838907905	Tech	
594635534930157568	KCA	
594635534930157568	VoteJKT48ID	
594635540890132481	unifydock	
594635540890132481	international	
594635540890132481	kickstarter	
594635540890132481	FreeAppleWatch	

tweet\_tags 1

Action Output

	Time	Action
6	14:45:03	SELECT * FROM twitter.tweets LIMIT 0, 50000
7	14:45:06	SELECT * FROM twitter.tweet_tags LIMIT 0, 50000

Tweet URLs table:

tweet_urls
tweet_id BIGINT(20)
url VARCHAR(140)
Indexes

1 • `SELECT * FROM twitter.tweet_urls;`

100% 1:1

Result Grid Filter Rows: Search Edit:

tweet_id	url
594634975036592128	https://medium.com/@noralev/what-happens-aft...
594634996297519104	http://mattgimmell.com/distractions/
594634998730366976	http://bit.ly/1OPcCG3
594635010382155776	http://tcrn.ch/1JGRKNL
594635014454673408	http://bit.ly/1EMPuAT
594635036726550531	http://bit.ly/1EYtmG9
594635046356709376	http://9gag.com/gag/aVW4RYy?ref=t
594635074777313280	http://informatyucatan.com/?p=118023
594635104397303810	http://techcrunch.com/2015/04/25/the-apple-wat...
594635115369631744	http://wp.me/p4g8He-1SYi
594635121535393792	http://i0.kym-cdn.com/photos/images/facebook/...
594635137070956544	https://www.kickstarter.com/projects/167259172...
594635177587965952	http://www.mind-mory.com/giveaways
594635201331888128	http://bit.ly/1yQw5z0
594635227412041728	http://bit.ly/1yQw5z0
594635227785502721	http://buff.ly/1brCwxE
594635249264398336	http://apple.co/1cyFmlg
594635249264398336	http://youtu.be/Ao8cGLIMtvg
594635251084873731	http://wp.me/p4dfnD-2EHg
594635324547993603	https://www.kickstarter.com/projects/167259172...
594635345863421952	https://itunes.apple.com/gb/app/live-media-play...
594635345863421952	http://tl.gd/n_1sm1k0h
594635449366421504	http://ift.tt/1DSH5bN
594635450155740700	http://youtu.be/7DEFWk4MwQ

tweet\_urls 1

Action Output

User mentions table:

tweet_mentions
tweet_id BIGINT(20)
source_user_id BIGINT(20)
target_user_id BIGINT(20)
Indexes



1	SELECT * FROM twitter.tweet_mentions;		
100%	1:1		
Result Grid	Filter Rows:	Search	Edit:
tweet_id	source_user_id	target_user_id	
594634975036592128	3140889318	2020351	
594635046356709376	316010030	316010030	
594635046356709376	316010030	16548023	
594635104397303810	258810247	816653	
594635106087608320	63636163	35293	
594635177587965952	1961671051	2749313005	
594635186932948992	2812076916	64929124	
594635186932948992	2812076916	20158584	
594635186932948992	2812076916	140195719	
594635249264398336	33013643	74580436	
594635249264398336	33013643	10228272	
594635565884145666	2020351	2020351	
594635638185586688	57704726	2749313005	
594635673786691584	457206998	2897431	
594635673786691584	457206998	52247685	
594635758348173312	2812076916	140195719	
594635796931432449	3011301300	26411479	
594635908651024384	711593403	26411479	
594635931258204161	2646742338	139353101	
594635931258204161	2646742338	51123740	
594636012455755777	15925189	18818627	
594636064548986880	49308532	2749313005	
594636317771759616	790036	2457172872	
594636317771759616	790036	2457172872	
tweet_mentions 1			
Action Output			

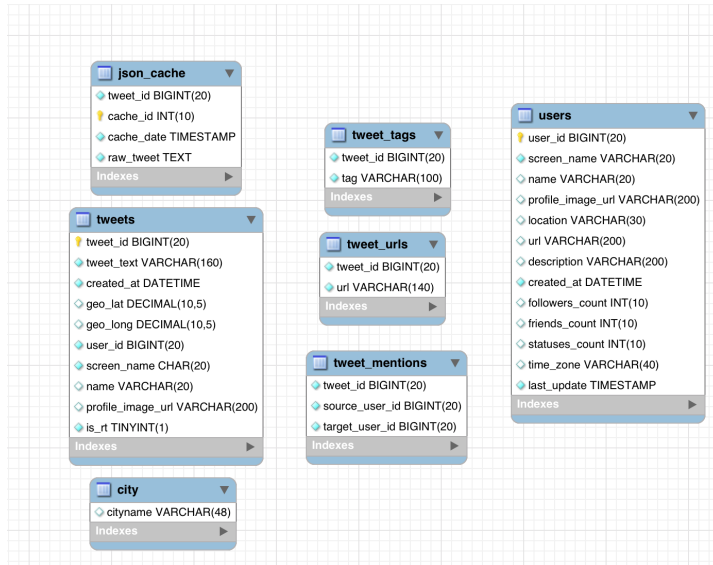
CITY\_NAME Table(for location analysis):

city
cityname VARCHAR(48)
Indexes



1	•	SELECT * FROM twitter.city;
100%	1:1	
Result Grid	Filter Rows:	Search
cityname		
▶ bab el ahmar		
chefe muzamane		
donji kraj		
el eneal		
el islero		
el quemado		
estancia jokho pekhe		
hayy al akrad		
las tres marias		
los higuerores		
madinat al habbanayah		
magharat `ubayd		
mala paukova		
mechta charef		
munhoque		
petar		
qaryat ad dawayah		
qaryat khatunyah		
qaryat `ajimi as sabbar		
san agustin		
santa feliicitas		
umm qusur		
vzemen orman		
city 1		

## EER Diagram:



## Project Sample input/out snapshot:

### Twitter display adapter

**2fantabulush** Mariizzzii♥  
SentimentAnalyze Geolocation  
I may or may not be excited for "Pitch Perfect 2". I also may or may not want to see "Hot Pursuit".  
2 days ago  
Map shows approximate location.

**NorthCapitolDC** North Capitol Main  
SentimentAnalyze Geolocation  
regram grsrtsgourmet Springy sugar cookies!!! Perfect for a springtime day and Mother's Day!!! @ ... <https://t.co/Jq7MbxGgzj>  
2 days ago  
Map shows approximate location.

**Writer\_HeidiG** Queen\_Bitch\_Carly  
SentimentAnalyze Geolocation  
RT @CanProvelt: Florida Man Dies In Police Custody After Cops

## Homepage index.php:

Get tweets of specific topics from tweet server in real-time:

870 new tweets available. [Refresh](#) to see them.



**Jusara\_\_** Sarah

SentimentAnalyze

Geolocation

RT [@vinit\\_mittal](#): Sad but true.. <http://t.co/Pj3EjXrz4r>

less than a minute ago

Map shows approximate location.



**c\_alayna** Layna

SentimentAnalyze

Geolocation

I work myself up so much before a speech and then I literally do perfect

less than a minute ago

Map shows approximate location.



**cellabrations** hannah

## Asynchronously Get more tweets from database



**BraderzClayts** Bradley Clayton

SentimentAnalyze

Geolocation

RT @stillblazingtho: Overthinking is the biggest cause of our unhappiness. Keep yourself occupied. Keep your mind off things that don't hel...

about a minute ago

Map shows approximate location.



**Antiindustry** AntiIndustry.com™

SentimentAnalyze

Geolocation

Someone tell @S\_C\_ to start a distribution co. For positive hiphop its already been proven, the only rap music thats selling #JustFacts

about a minute ago

Map shows approximate location.

60 tweets displayed - View More

## Search whatever topics you want:

search



**DimitryJacobs** Dimitry Jacobs

SentimentAnalyze

Geolocation

RT @OmniFocus: OmniFocus 2.5.2 for iOS, featuring Apple Watch and crash fixes, has been approved and should be available shortly: <https://t...>

less than a minute ago

Map shows approximate location.



**LordJackTaylor** Jack Taylor

SentimentAnalyze

Geolocation

@SantanderCycles are you going to make an Apple Watch app? That would make it much easier...

less than a minute ago

Map shows approximate location.



**HelloHeartApp** Hello Heart

SentimentAnalyze

Geolocation

@BRGLiving Excellent! We also think that the #AppleWatch is going to be a game changer, here is why

## Geolocation analysis of a single tweet:



**Tigerbudy** Cassie Snyder

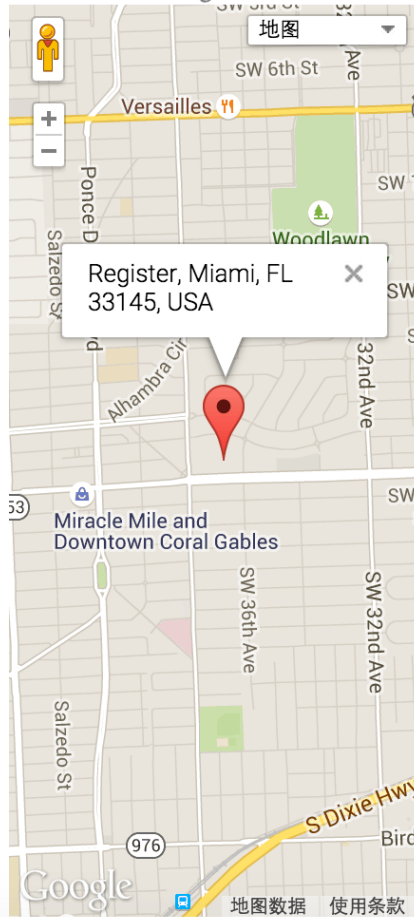
SentimentAnalyze

Geolocation

Please help me support Best Buddies Florida. Visit my page on the Bowling for Buddies website and register or...

<http://t.co/6Gu2grOYD8>

less than a minute ago



latitude: 25.7506651, longitude: -80.2529918

## Sentiment analysis of a single tweet:

---



**Tigerbudy** Cassie Snyder

SentimentAnalyze

Geolocation

Please help me support Best Buddies Florida. Visit my page on the Bowling for Buddies website and register or...

<http://t.co/6Gu2grOYD8>

about a minute ago

Map shows approximate location.



**kcin1122g** Nicholas S. Gonzalez

SentimentAnalyze

Geolocation

RT @ontarioeda: Top 10 Reasons to Visit Ontario, California at @ICSC\_RECon. 1) There is incredibly plush carpet at the booth.

☺ <http://t.co...>

about a minute ago

Map shows approximate location.



**DOLL\_BABIE90** China Charisse♥

## workflow

