

Laboratory Tutorial 6: Working with a Twitter time line using SQL

In this laboratory tutorial you will:

1. Become familiar with the Twitter data schema
2. Run some basic and advanced SELECT queries on this dataset using SQL Interface

Preamble

Twitter is a micro-blogging service, which enables users to issue short messages of 140 characters or less. Each user can both follow and be followed by other users. Users are able view the messages of any users that they follow and, likewise, any users that follow them will be able to view their messages. Users can propagate other users' tweets using by retweeting (RT) a message. Users can also send messages to specific users by using the @<screenname> reference. As of 2013, Twitter has 200 million users world-wide and over 400 million tweets are sent per day.

In this tutorial you will manipulate and analyse Twitter data. The manipulation and analysis will involve SQL operations.

Tweetcatcher (TC) is a tool to collect data directly from Twitter, using its data API. TC saves a subset of the data fields returned by a Twitter search. The Tweets table schema, as shown on page 2, names and describes these fields.

Tweets table schema

Attribute name	Description	Domain
dateTime	the date and time of that the tweet was sent	Date/Time
screenName	the author's Twitter handle	String
tweet	the actual message content	String
<u>tweetId</u> (PK)	The unique id of the tweet	String
realName	the full provided name of the author	String
followers	the number of followers of the author	Integer
following	the number of users followed by the author	Integer
publicLists	the number of public lists the author belongs to	Integer
timeZone	the time zone in which the author has declared that they reside	String
geoCoords	the longitude and latitude of the author at the time the tweet was sent	String
userTweets	the number of status updates made by the author	Integer
retweets	the number of times that the status update has been retweeted	Integer
linkURL	any links embedded within the message	String
sentPos	a rating of positive sentiment	Integer, from 1 (not positive) to 5 (highly positive)
sentNeg	a rating of negative sentiment	Integer, from -1 (not negative) to -5 (highly negative)
sentNet	the sum of both Sent+ and Sent-	Integer from -4 (highly negative, not positive) to 4 (highly positive, not negative)

The TC tool collect tweets on a topic and convert the output into a database table like the one used here. The topic could be anything – 'UK and Brexit', 'Covid and vaccine', etc.

In this module we will not be using the TC tool. We will simply use the twitter data gathered by the tool. The aim of this tutorial is to familiarise you with the use of the data table that is returned by TC. You will formulate some queries to help you to answer certain questions about a data topic.

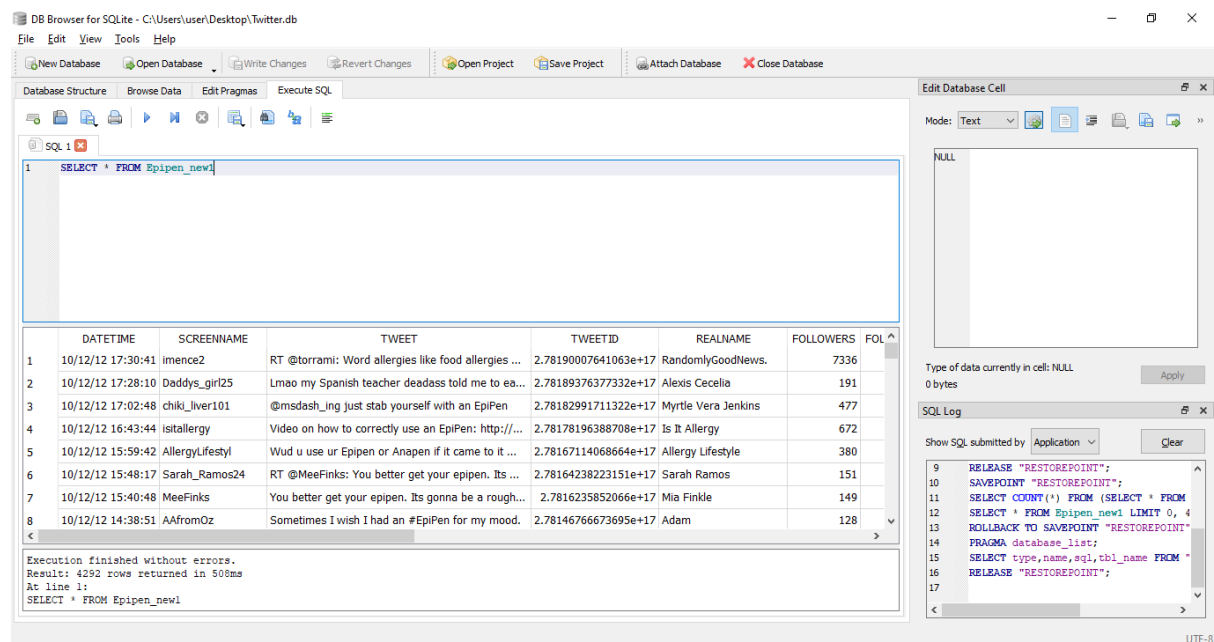
Exercise 6.1: Accessing the databases in SQL interface

We will continue to use the DB Browser for SQLite. The database that you will be using in this tutorial is called “Twitter” and can be downloaded from the resources folder on VLE (Go to VLE→DAT2→Practical→resources→Twitter).

Download and save this database on your desktop. Next, in your DB Browser, click on ‘open Database’, then select the database (i.e., ‘Twitter’) that needs to be opened. In this database there is a table called ‘Epipen_new1’. Try the following query:

SELECT * FROM Epipen_new1

You should see a table appear as in the figure below, with 16 columns and some 4292 rows (tweets). The table schema presented earlier (see page 2) names and describes all of the columns. The table represents all tweets mentioning the word ‘Epipen’ over a 9 week period running from 4th October 2012 to the 12th of December 2012. Epipen is the trade name for a type of injection device used by people who are prone to severe allergic reactions, known as anaphylaxis, to certain agents like bee stings and peanuts. Epipen injects adrenaline into the blood stream, which can provide immediate relief from life threatening symptoms such as a constrained airway. Hence, for sufferers, an Epipen can be a life-saver and must be carried with them at all times.



DB Browser for SQLite - C:\Users\user\Desktop\Twitter.db

File Edit View Tools Help

New Database Open Database Write Changes Revert Changes Open Project Save Project Attach Database Close Database

Database Structure Browse Data Edit Pragma Execute SQL

SQL 1

```
1 SELECT * FROM Epipen_new1
```

	DATETIME	SCREENNAME	TWEET	TWEETID	REALNAME	FOLLOWERS	FOL
1	10/12/12 17:30:41	imence2	RT @torrami: Word allergies like food allergies ...	2.78190007641063e+17	RandomlyGoodNews.	7336	
2	10/12/12 17:28:10	Daddys_girl25	Lmao my Spanish teacher deadass told me to ea...	2.78189376377332e+17	Alexis Cecella	191	
3	10/12/12 17:02:48	chiki_liver101	@rnsdash_ing just stab yourself with an EpiPen	2.78182991711322e+17	Myrtle Vera Jenkins	477	
4	10/12/12 16:43:44	isitallergy	Video on how to correctly use an EpiPen: http://...	2.78178196388708e+17	Is It Allergy	672	
5	10/12/12 15:59:42	AllergyLifestyl	Wud u use ur Epipen or Anapen if it came to it ...	2.78167114068664e+17	Allergy Lifestyle	380	
6	10/12/12 15:48:17	Sarah_Ramos24	RT @MeeFinks: You better get your epipen. Its ...	2.78164238223151e+17	Sarah Ramos	151	
7	10/12/12 15:40:48	MeeFinks	You better get your epipen. Its gonna be a rough...	2.7816235852066e+17	Mia Finkle	149	
8	10/12/12 14:38:51	AAfromOz	Sometimes I wish I had an #EpiPen for my mood.	2.78146766673695e+17	Adam	128	

Execution finished without errors.
Result: 4292 rows returned in 308ms
At line 1:
SELECT * FROM Epipen_new1

Edit Database Cell

Mode: Text

Type of data currently in cell: NULL
0 bytes

SQL Log

Show SQL submitted by: Application

```
9 RELEASE "RESTOREPOINT";
10 SAVEPOINT "RESTOREPOINT";
11 SELECT COUNT(*) FROM (SELECT * FROM
12 SELECT * FROM Epipen_new1 LIMIT 0, 4
13 ROLLBACK TO SAVEPOINT "RESTOREPOINT";
14 PRAGMA database_list;
15 SELECT type,name,sql,tbl_name FROM *
16 RELEASE "RESTOREPOINT";
17
```

UTF-8

Exercise 6.2: What can we learn from Twitter about Epipen?

Retweeting (denoted by an RT prefix) is a common practice in Twitter. This is similar to forwarding an email message and is the primary means of propagating data within the Twitter network. Hence, the RT count of a message is an indicator of its popularity.

Q1. Write a query to list all details about Epipen with retweets of 12.

We can also analyse the users (authors) who post these tweets.

Q2. Write a query to find out all authors whose screenname is the same as their real name.

Q3. Write a query to find out all authors whose screenname is the same as their real name and with no URL link.

Hashtags (e.g., #DAT2) are widely used in Twitter as an informal way of classifying tweets into topics

Q4. Write a query to list all tweets that contain at least one hashtag?

Q5. From Q4 how many records did you result with?

TC uses an intelligent natural language processing algorithm to estimate the degree of positive and negative sentiment (feeling) expressed in each tweet. These scores are represented individually and in combination as the final three columns.

Q6. Write a query to show the frequency distribution of positive sentiment scores (i.e., how many tweets scored 1, how many scored 2 etc.)

Q7. From Q6 how many tweets received a positive rating of 4 or more?

Q8. Write a query to find out the number of tweets that had no measurable sentiment at all (meaning sentpos=1 AND sentneg=-1)?

Q9. From Q8 how many tweets did you get?

Summary and future work

In this tutorial you have experienced the data schema provided by the Tweetcatcher tool. You have worked with an instance of this schema, using DB Browser for SQLite. You have also gained experience working with a significantly larger dataset. This has enabled you to gain a better understanding of the utility of filtering and aggregating a large dataset.

Further Reading

Databases Illuminated (Chapter 1, Sections 1.3-1.5 | Chapter 5, Sections 5.1, 5.2, 5.4).