#Charlottesville on Twitter

Vincent La

August 21, 2017

Data Description

Source

This dataset is a collection of tweets taken from the Twitter Streaming API. This is a continuous collection of tweets mentioning "Charlottesville" or using the hashtag #Charlottesville. However, a small amount of tweets were lost due to programming bugs and intermittent connection failures.

Format

The data is available as either five CSV files or a single SQLite database.

SQLite Caveat Because Twitter IDs can be very massive, thereby causing integer overflows, some of them may have been stored as strings instead. Remember that SQLite supports dynamic typing, and column types are treated merely as a suggestion.

License

Tweet Samples

Each file or table named 'aug**_sample.csv' contains a random sample of 50,000 tweets (in accordance with the Twitter Developer Agreement) from each day. It should be noted that due to programming bugs and intermittent connection failures, a small number of tweets were not collected. Therefore, these samples may potentially be less than truly random. Furthermore, because I started collecting data on August 15, that day's sample only includes tweets after 9PM Eastern Time.

Attributes

Since the vast majority of attributes are unmodified and self-explanatory, I'm only going to describe the less obvious ones and the attributes I either created or cleaned (there's only two). For the rest, I will describe what attributes from the Twitter API they came from. An overview of tweet attributes can be found here on Twitter's website.

Attribute	Source	Description
id	Unmodified	Integer corresponding to the Tweet ID
user_id	user -> 'id'	
user_name	user -> 'name'	Twitter user name
screen_name	user -> 'screen_name'	
user_statuses_count	user -> 'statuses_count'	
user_favorites_count	user -> 'favorites_count'	
friends_count	user -> 'friends_count'	
followers_count	user -> 'followers_count'	
user_description	user -> 'description'	How the user chooses to describe them self
user_location	user -> 'location'	**Note:** Twitter places no restrictions on what users can enter as their location
user_time_zone	user -> 'time_zone'	
user_profile_text_color	user -> 'profile_text_color'	
user_profile_background_color	user -> 'profile_background_color'	
full_text	Either text or extended_tweet -> 'full_text'	
created_at	Unmodified	UTC timestamp of when Tweet was posted. For reference, Eastern Standard Time is 4 hours behind UTC.
is_retweet		A binary variable I created to help me subset the data. Not very useful outside of that.
retweeted_status_text	retweeted_status -> 'text'	
retweeted_status_id	retweeted_status -> 'id'	
quoted_status_text	quoted_status -> 'text'	The text of the tweet that this status referenced (if applicable)
quoted_status_id	quoted_status -> 'id'	
in_reply_to_screen_name	Unmodified	
in_reply_to_status_id	Unmodified	
in_reply_to_user_id	Unmodified	
hashtags	entities -> 'hashtags'	I used a Postgres function to flatten out the JSON array which contained the list of hashtags.

Summary Statistics

tweet_count_time_series

This table contains the number of tweets created at every time stamp. It was computed using the **entire Postgres** database, not just the sample posted on Kaggle. As you may have guessed, summing over tweet_count will give you the total number of tweets contained in the original database.

Note Large time gaps between sequential timestamps should be taken as a sign of my internet or stream cutting out. This is only really a problem with the first two days of the dataset before I modified my script.

Attribute	Description
created_at	UTC timestamp with datetime information down to the second
created_at_day	Date of the timestamp (parsed from created_at)
created_at_hour	Hour of the timestamp (parsed from created_at)
tweet_count	The number of tweets with created_at as their timestamp

Image Credit

The banner used above was made personally by combining and modifying images from:

- Evan Nesterak [Source] [License]
- Wikipedia user Cville Dog [Source]
- Associated Press [Source]