How to Collect, Clean, and Use

Twitter Data with Python

By Christian Howard-Sukhil June 2019

Download the Python scripts

- 1. Go to the GitHub page for the Twitter project, "TwitLit" (<u>https://github.com/TwitLit/TwitLitSource</u>).
- 2. Download both the Python files (TwitterScraper.py and count_jsonl_rows.py).
- 3. Put both of these files in the folder/directory in which you want to collect all of your Twitter data.

Scrape Twitter using the Python Twitter Scraper

- 4. Open Terminal and navigate to the folder/directory that contains the python Twitter Scraper (TwitterScraper.py).
 - a. To change directory in Terminal, enter the following into the command line: cd [name of file]
- 5. In this directory, run the python Twitter Scraping script by entering the following into the command line: python3 TwitterScraper.py
- 6. This will then prompt you for the search term
 - a. The script can search for # specific terms; simply enter the hashtag followed by the term when prompted (e.g., #twitfiction)
 - b. To search for multiple terms simultaneously, put OR between terms



7. Once you enter the search term, you will be prompted to enter a date range. You'll be asked for the start date (YYYY-MM-DD); after inputting this and pressing enter, you'll be asked for the end date (YYYY-MM-DD).



- 8. After you input the date range, the script will run. This could take a few seconds or several hours, depending on how many search results there are.
- 9. Once the script finishes running, you'll see both a txt file and a csv file in the folder containing your python Twitter Scraper. These will automatically be named in the following format: search-term_start-date_end-date. (NB: If you used a hashtag in your search term, the hashtag will not appear in the document name since doing so would cause problems with the file naming system.)
 - a. The txt file will include a full list of Tweet IDs (unhydrated).
 - b. The csv file will contain the Tweet IDs plus selected information about each tweet, including the full text of the tweet, the timestamp, and the user screen name.

Cleaning up your csv files

- 10. The csv file may use commas to separate columns instead of semicolons, which can make your data messy. To fix this, do the following:
 - a. Open a NEW Excel spreadsheet.
 - b. Under the data tab, click the "From text" button:



- c. You should receive a pop-up from which to select the file in question. (NB: If you don't, check to see if you already have the file open in a different excel sheet. If so, close this file and repeat this step.)
- d. Select the file and click "Get Data."

 count_jsonl_rows.py cultural-ap6-01_ids.txt cultural-ap9-06-01.csv DACA_2012-01_ids.txt DACA_2016-02-01.csv twarc.log TwitterScraper.py 	
	cultural-appropriation_2019-01-01_2019-06-01.csv Comma Separated Spreadsheet (.csv) - 23.9 MB Tags Add Tags Created Monday, June 17, 2019 at 10:52 AM Modified Today, 9:53 AM Last opened Today, 9:41 AM

e. You'll get another pop-up menu called "Text Import Wizard." Select the "Delimited" button and then click "Next."

Text Import Wizard - Step 1 of 3
The Text Wizard has determined that your data is Fixed Width.
f this is correct, choose Next, or choose the Data Type that best describes your data.
Delimited - Characters such as commas or tabs separate each field. Fixed width - Fields are aligned in columns with spaces between each field.
Start import at row:
Preview of selected data: Preview of file /Users/cfh00/cultural-appropriation_2019-01-01_2019-06-01.csv.
Preview of selected data: Preview of file /Users/cfh00/cultural-appropriation_2019-01-01_2019-06-01.csv. 1 id;timestamp;user_screen_name;user_name;retweets;favorites;text,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

f. Select the value you'd like to use as the column delimiter in the csv file. For Twitter data, you will most likely want to select "Semicolon." Make sure no other boxes are selected and press "Next."

Text Import Wizard - Step 2 of 3						
This screen lets you set the delimiters your data contains.						
Delimiters	-		Ĩ			
Tab	Treat consecutive delimiters as one					
🗸 Semicolon			Tex	aualifier:	0	
Comma						
Comma						
Space						
Space						
Space Other:						
Space Other:	ted data:					
Space Other:	ted data:					
Space Other:	ted data:		1			1
Space Other:	ted data: timestamp 2019-01-05	18:58:49	user_screen_name 1515202808	user_name Dark Diamond.		retweets favor β β,
Space Other: Preview of select id 1081701583992295424 1081701033871622144 1081700163412680801	ted data: timestamp 2019-01-05 2019-01-05 2019-01-05	18:58:49 18:56:38 18:53:06	user_screen_name 1515202808 925038576269778945 82088984676380171	µser_name Dark Diamond. Anna Rainey NiniaFraa6		retweets favor 0 0, 0 1,
Space Other: Preview of select id 1081701583992295424 1081701033871622144 1081700146541260801 1081693937769936896	ted data: 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05	18:58:49 18:56:38 18:53:06 18:50:08	user_screen_name 1515202808 925038576269778945 820898984763891712 2924171436	user_name Dark Diamond. Anna Rainey NinjaFrog6 Slacker Life		retweets favor 0, 0, 0, 1, 0, 0,
Space Other: Dreview of select 081701583992295424 1081701033871622144 10816993057709936896 10816993957730935264 1081270127320275	ted data: 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05	18:58:49 18:56:38 18:53:06 18:50:08 18:50:01	user_screen_name 1515202808 925038576269778945 8208389384763891712 2924171436 857659344 93900	user_name Dark Diamond. Anna Rainey Ninjafrog6 Slacker Life Opal		retweets favor 0 0, 0 1, 0 1, 0 0, 0 1,

g. On the final page of the Text Import Wizard, make sure your "column data format" is checked appropriately. Usually, the "General" button should be fine. Click "Finish."

Text Import Wizard - Step 3 of 3						
This screen lets you select each column and set the Data Format.						
Column data for	mat					
General						
Text						
Date: MDY	0					
Do not impor	t column	(Skip)				
Do not impor	t column	(Skip)				Advanced
Oo not impor	t column	(Skip)				Advanced
Do not impor	t column	(Skip)				Advanced
Do not impor	t column ted data:	(Skip)				Advanced
Do not impor Preview of selec	t column ted data: <u>General</u>	(Skip)	General	General		Advanced
Do not impor Preview of select	ted data:	(Skip)	General user_screen_name t=t=za?ea	General Jser, name Deale, Diamond		Advanced
Do not impor Preview of selec General id 1081701583992295424	ted data: <u>General</u> timestamp 2019-01-05 2019-01-05	(Skip) 18:58:49 18:56:38	General user_screen_name 1515202808 D25038576269778945	<u>General</u> user_name Dark Diamond. Anna Rainey		Advanced
Do not impor Preview of select General id 1081701033871622144 1081701045541260880	ted data: ted data: timestamp 2019-01-05 2019-01-05 2019-01-05	(Skip) 18:58:49 18:56:38 18:53:06	General Lser_screen_name 1515202808 925038576269778945 820898984763891712	<u>General</u> user_name Dark Diamond. Anna Rainey VinjaFrog6		Advanced
Do not impor Preview of select General 1081701583992295424 1081701033871622144 108170016541260801 1081699397769936896	t column ted data: timestamp 2019-01-05 2019-01-05 2019-01-05 2019-01-05	(Skip) 18:58:49 18:56:38 18:53:06 18:50:08	General Lser_screen_name 1515202808 9250038576269778945 20898984763891712 2924121436	General user_name Dark Diamond. Anna Rainey VinjaFrogó Slacker Life		Advanced
Do not impor Preview of select General 1081701583992295424 1081700145541260801 1081699365713035264 1081699365713035264	ted data: <u>General</u> timestamp 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05 2019-01-05	(Skip) 18:58:49 18:56:38 18:53:06 18:50:08 18:50:00	General Lser_screen_name 1515202808 925038576269778945 820898984763891712 2924171436 857650344 738301	General Liser_name Dark Diamond. Anna Rainey VinjaFrog6 Slacker Life Dpal Re VoicasOf		Advanced

h. You'll be asked where you want to import your data; the first box of the first column of the existing spreadsheet should be fine. Select this box, then press "OK."

Import Data						
Where do you war	nt to put the data?					
Existing sheet:	=Sheet1!\$A\$1					
New sheet						
O PivotTable						
Properties	Parameters	Cancel	ОК			

i. Your data should import into the new Excel spreadsheet. Save the file as a csv using an appropriate file name.

Analyzing your data using Twarc

(For more detailed instructions about using DocNow/Twarc, see Christian Howard's blog post, "Twitterature: Mining Twitter Data" at https://scholarslab.lib.virginia.edu/blog/twitterature-mining-twitter-data/)

- If you want more information than that provided by the automaticallydownloaded csv file, you'll need to download Twarc by Documenting the Now (DocNow - <u>https://www.docnow.io/</u>).
- 12. To hydrate the txt files (gathered when you scraped Twitter; see Step 9 above) using Tware, you'll enter the following into your command line: Tware hydrate [tweet]_ids.txt > [tweet].jsonl
 - a. Replace [tweet] with the name of your file
- 13. Once tweets are hydrated, you should use the "Dedupe" tool in order to ensure that you don't have any duplicates in your data. To do this, enter the following in your command line: utils/deduplicate.py [tweet].jsonl > [tweet]_deduped.jsonl
 - a. Replace [tweet] with the name of your file
- 14. After hydrating and deduping your dataset, you can play with different Twarc possibilities, such as creating word clouds from your tweets and identifying user locations using geojson. See DocNow/Twarc on GitHub for detailed instructions: <u>https://github.com/DocNow/twarc</u>.

Counting your Tweets

15. If you want to find out how many tweets you have in a given search query, you'll need your data in a jsonl file. You'll also want to ensure that you aren't counting

any replicated tweets, so you should deduplicate your file first. To do this, follow instructions 11-13 above.

- 16. Once you have your deduped jsonl file, navigate to the folder/directory that contains your Python rows counting script (count_jsonl_rows.py). It is important that the jsonl file you want to count is in the same folder/directory as this counting script.
- 17. Enter the following into your command line: python count_jsonl_rows.py

```
[Macintosh-6:~ christianhoward$ cd Twitter-Search-API-Python/
[Macintosh-6:Twitter-Search-API-Python christianhoward$ python count_jsonl_rows.py
```

18. This will prompt you to enter the name of the file.

[Macintosh-6:~ christianhoward\$ cd Twitter-Search-API-Python/ [Macintosh-6:Twitter-Search-API-Python christianhoward\$ python count_jsonl_rows.py enter name of file: deduped_community_combined_2011.jsonl

19. After you enter the name of the file, the script will run and then output the "number of rows in the file." Since each row corresponds to a unique tweet, you now know how many tweets are in the file in question.

Macintosh-6:~ christianhoward\$ cd Twitter-Search-API-Python/ Macintosh-6:Twitter-Search-API-Python christianhoward\$ python count_jsonl_rows.py enter name of file: deduped_community_combined_2011.jsonl number of rows in file: 1472