Collecting Twitter Data

Alex Leslie

9/27/2018

This Workshop

Welcome! This workshop will introduce the general process for obtaining data from the social media platform Twitter, including a hands-on tutorial of the TAGS tool for Google Sheets. In the process, we will also address what kinds of information we can learn from social media data as well as several important factors to consider when formulating and carrying out research.

For the hands-on portion, you'll need to have a Google account and a Twitter account. If you don't have either of these but want to follow along, sign up now; it takes just a minute.

Preliminaries

What Is Our Archive

To begin with: what does Twitter data represent in the first place?

It's essential to know what our archive actually consists of in order to know what any conclusions we draw from it actually mean.

The seeming ubiquity of social media and the popular conception of social media platforms as democraticizing "public squares" belie skewed demographics of use - which can change in significant ways over time. The Pew Research Center's biennial social media update indicates several of Twitter's demographic imbalances.

What Is Our Archive?

Use of different online platforms by demographic groups

% of U.S. adults who say they use ...

	Facebook	YouTube	Pinterest	Instagram	Snapchat	LinkedIn	Twitter	WhatsApp
Total	68%	73%	29%	35%	27%	25%	24%	22%
Men	62	75	16	30	23	25	23	20
Women	74	72	41	39	31	25	24	24
White	67	71	32	32	24	26	24	14
Black	70	76	23	43	36	28	26	21
Hispanic	73	78	23	38	31	13	20	49
Ages 18-29	81	91	34	64	68	29	40	27
18-24	80	94	31	71	78	25	45	25
25-29	82	88	39	54	54	34	33	31
30-49	78	85	34	40	26	33	27	32
50-64	65	68	26	21	10	24	19	17
65+	41	40	16	10	3	9	8	6
<$30,000	66	68	20	30	23	13	20	20
$30,000-$49,999	74	78	32	42	33	20	21	19
$50,000-$74,999	70	77	34	32	26	24	26	21
$75,000+	75	84	39	42	30	45	32	25
High school or less	60	65	18	29	24	9	18	20
Some college	71	74	32	36	31	22	25	18
College+	77	85	40	42	26	50	32	29
Urban	75	80	29	42	32	30	29	28
Suburban	67	74	31	34	26	27	23	19
Rural	58	59	28	25	18	13	17	9

Note: Whites and blacks include only non-Hispanics. Hispanics are of any race.
Source: Survey conducted Jan. 3-10. 2018
"Social Media Use in 2018"

PEW RESEARCH CENTER

Figure 1: U.S. Twitter Users Demographics

What Is Our Archive?

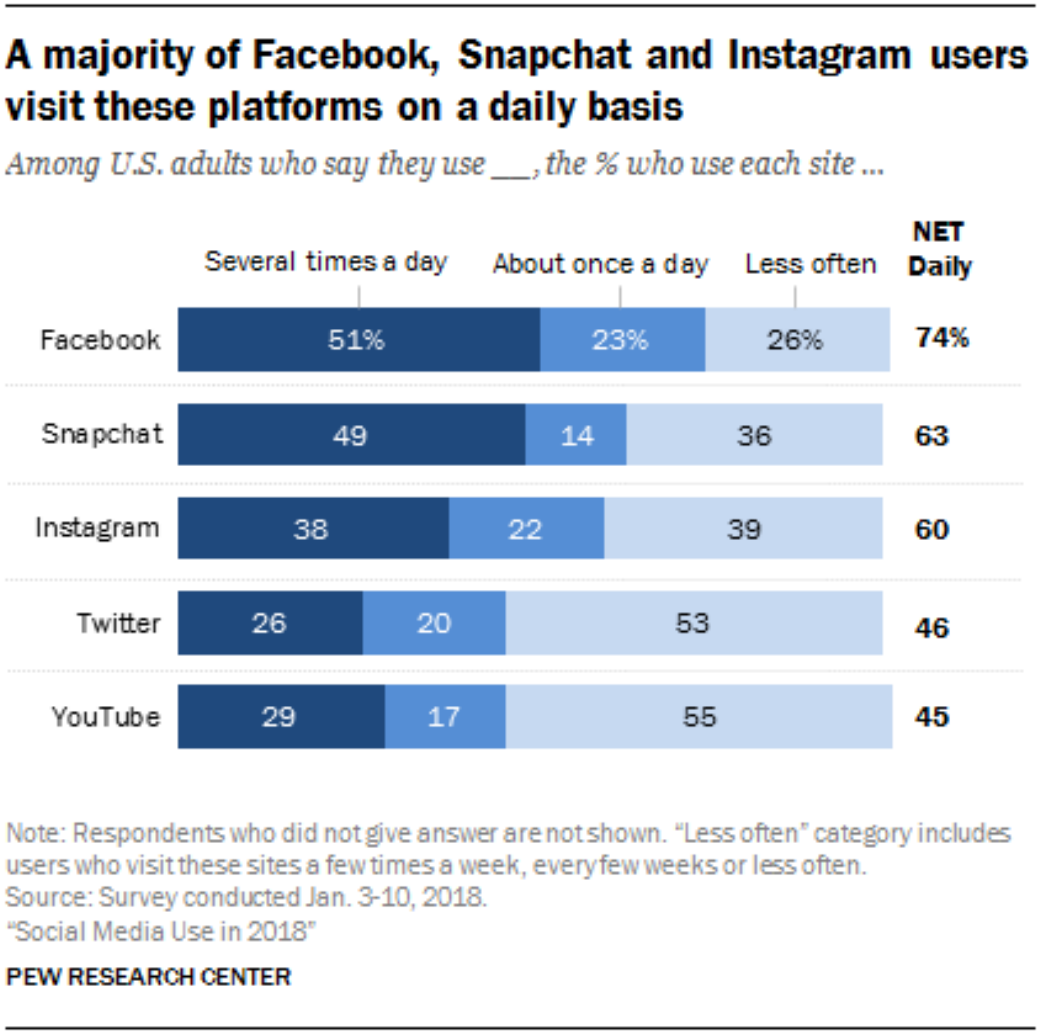

Figure 2: U.S. Social Media Usage Time

What Is Our Archive?

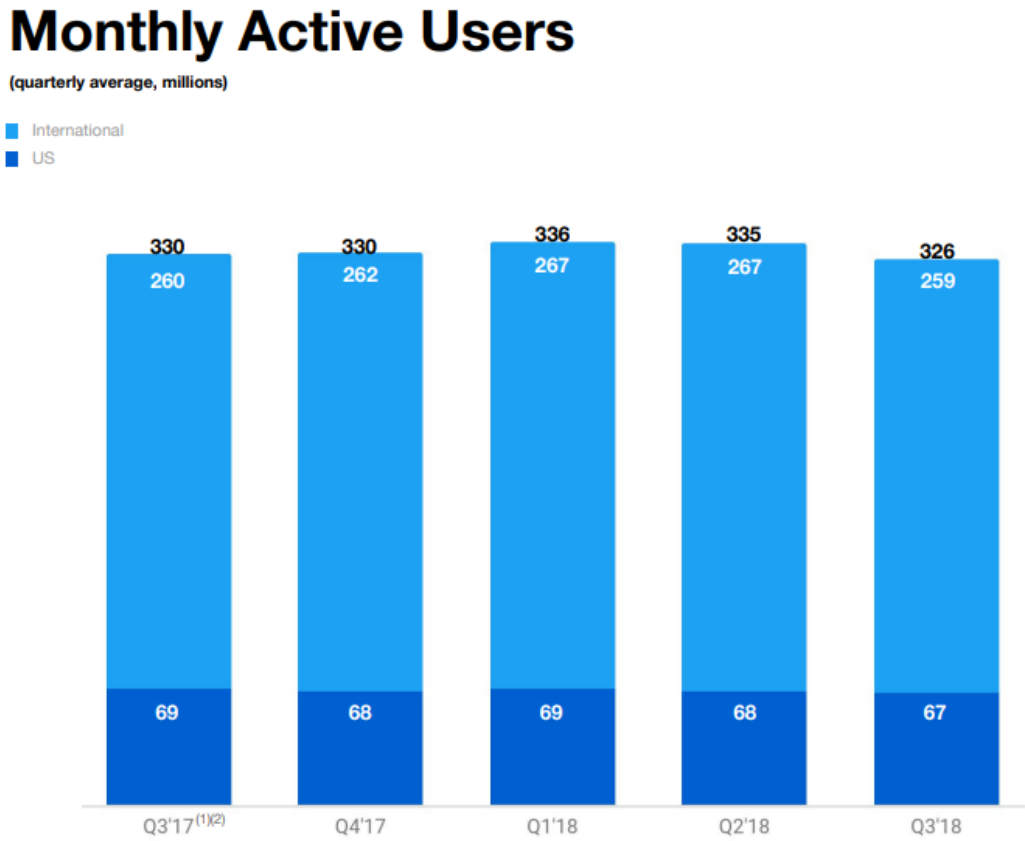

Figure 3: U.S. vs. International Active Users

What Is Our Archive?

What does this all tell us?

- One fourth of Americans use Twitter.
- American Twitter users are more often urban college graduates in higher income brackets.
- Usage also matters: we often think of Twitter's instaneity as its distinguishing feature, yet most users don't check it daily.
- American Twitter users, however, make up only about 20% of all Twitter users.

And don't forget: demographic data changes! Checking the Pew Research Center's previous biennial reports shows that these numbers shift.

What's the moral of the story? Claims about Twitter are always "just" claims about Twitter.

Technical Issues

Collecting isn't scraping. Web scraping is when you give a program a list of URLs and tell it to copy all of the HTML by which each page is generated. Scraping is great for stable, primarily-textual websites, but websites with user-specific, constantly-updating, infinite-scroll feeds make scraping impractical if not impossible for most purposes.

Most tools for obtaining social media data instead work through on an API, or Application Programming Interface, produced by the social media company itself. Like any API, the Twitter API is meant to make data housed in complex structures more accessible in a tidier format to developers and researchers.

Technical Issues

Reliance on an API, however, also means that Twitter controls the manner in which users access data. This means there are a couple important limitations:

- Access to tweets in a term- or hashtag-based search is limited to the past six to nine days.
- Access to user-based searches is limited to the most recent 3,200 tweets.
- Twitter caps the rate of return at 18,000 tweets per 15 minutes.
- The obtained tweets are not comprehensive: Twitter's algorithms actually return what it considers a representative sample of tweets, even though what it returns often appears complete. For this reason, even searches with relatively few results may not return all existent instances of the search term.

We'll need to keep these technical limitations in mind when we formulate our research question and draw any conclusions.

TAGS

TAGS

TAGS is a tool built for Google Sheets by Martin Hawksey to communicate with the Twitter API. It has the benefits of being easy to set up, use, and update, without any downloaded software or programming knowledge required.

Setting Up

- First, make sure you're logged in to your Google account.
- Go to https://tags.hawksey.info/get-tags/ and click the `TAGS 6.1` button. Click "Make a copy" when prompted.
- This will redirect to Google Docs and load a pre-built spreadsheet. Before doing anything with this, however, click on the `TAGS` tab in the top navigation bar and then "Setup Twitter access."
- Click "Continue," select the Google account you want to use (if you have more than one), and click "Allow" to grant permissions to TAGS. What are we allowing?
  - TAGS to manage all TAGS spreadsheets.
  - TAGS to connect to and execute the Twitter API.
  - TAGS to update searches, if you choose, automatically.

Setting Up

- There are two possible modes for using the Twitter API: either through your own Twitter account by obtaining developer permissions or by piggybacking on the TAGS Twitter developer account.
  - It is actually quite simple to obtain developer permissions for your own Twitter account, but the approval process can take up to a day or two. So for now, click "Easy setup."
- This will redirect to a new tab asking for Twitter account authorization; click "Sign in with Twitter."
- Another new tab! Type in your account name and password, then click "Authorize app."
- When prompted, close the extra tabs and return to Google Sheets. Change the title of the spreadsheet in the upper left to something more recognizable, like "twitter archvie test." Setup complete!

Creating an Archive

Note in the bottom left of your window that there are two sheets in the TAGS spreadsheet: the current Readme/Settings sheet and the Archive sheet. Switch to this sheet. This is what TAGS results will look like: one row (or observation) for each tweet, and one column (or variable) for each piece of information about it.

A couple obvious fields are missing here, but TAGS allows users to specify what metadata variables it should collect for each tweet. All possible fields are listed at this webpage. Add `retweet_count` and `favorite_count` now - and whatever other variables you'd like - by selecting the last column, clicking the "Insert" tab and "Add column right," and then typing in the name of the variable you're adding.

Creating an Archive

Now return to the Readme/Settings sheet to look at the (Advanced) Settings.

- Period is the time span to search; the Twitter API limits term-search results to six to nine days, but you can set this range back up to seven days in the past.
- Follower count filter instructs the search to exclude all results from accounts with fewer than *n* followers.
- Number of tweets limits the results returned.
- Type allows a choice between three different kinds of search:
  - search/tweets, for instances of a particular word.
  - statuses/user_timeline, for a single user's posts, replies, and retweets.
  - favorites/list, for all the tweets a single user has liked.

Making an Archive

Formulating a Research Question

TAGS thus gives us three very different types of approach to choose from in order to sift through a large field of complex data. Before committing to a search, it's important to consider the kind of research question you're asking and what approach will be most useful in generating the archive you want.

- Searching by hashtag
  - Hashtags (#) are used to indicate a tweet's participation in a particular discourse. By tracking hashtag use, then, a researcher is tracking the way users signal participation in a particular discourse.
  - Some hashtags are topical and only crop up for brief spans of time while others are recurring or situational.
  - Hashtag use doesn't indicate the tweet's position on the discourse in question (ex., approval or disapproval), though there are some methods of computational analysis that can help guess.
  - Hashtags are often used ironically, and irony is much more difficult to identify. Again, hashtags are discursive markers: making claims about the meanings of those markers will require further analysis.

Formulating a Research Question

- Searching by term or phrase
  - Searching for a term can be more granular than searching for a hashtag, but they aren't subject to the same kind of explicit signaling as hashtags.
  - The most precise term-based searches focus on a term that is used distinctively. This minimizes but by no means eliminates interference from uses that aren't pertinent to the research question.
  - Memes can be interesting to track if you catch them early enough, but they can also morph beyond their initial defining terms.
  - Once again, term use doesn't indicate the tweet's position on a particular topic (ex., approval or disapproval); attempting to determine this would be the work of further analysis.
  - One quirk of term-based search is that it will also pick up all tweets by any user whose user name includes the term.
  - Searching for a very commonly-used term will only return the most recent results.

Formulating a Research Question

- Searching by user(s)
  - The two search options here result in slightly different emphasis.
  - A user's likes primarily tell us about how they interact with their broader Twitter ecosystem.
  - A user's tweets primarily tell us about their own content and the way other users interact with them (via likes and retweets).
  - Identifying a group of important users for a particular research question and generating archives for each is one way to circumvent the narrow time window, at least up to 3,200 tweets each. Additional users of interest can be identified by replies. For this kind of inquiry, it is important to establish clear conditions for user inclusion/exclusion.

Running a Search

Take a few minutes to consider a possible search; you may wish to try searching on Twitter itself first to get an idea of the possible results.

When you're ready, edit the settings accordingly and enter the search term or user name in the "Enter term" field. Click the "TAGS" tab on the upper navigation bar and select "Run now!"

Once the script has finished running (this will take a minute or two), go to the Archive sheet.

Borrowing an Archive

Locating an Archive

Annoyed by the time window? Many researchers have assembled their own Twitter datasets: one of them might better fit your needs.

Twitter doesn't allow the publishing of entire datasets of tweets online, but it does allow for the publishing of datasets of tweet IDs, which can then be matched with the tweets themselves with applications like Hydrator and even TAGS.

DocNow, a project developed through the University of Maryland, University of California Riverside, and Washington University in St. Louis for chronicling digital content, has an extensive catalog of datasets of tweet IDs at https://www.docnow.io/catalog/. Take a minute to scroll through; select a small dataset and download it.

Get Hydrated

- Once you've downloaded a dataset of tweets, unzip it and open it; it should be in a simple .txt format. This is what dehydrated tweets look like: just a long list of ID numbers. Highlight everything and copy it.
- Go back to your TAGS browser. Click the "TAGS" tab and select "Build archive from Tweet IDs"; this will generate a new sheet called "ID" that you can select in the bottom left.
- Select the cell in the first row and the first column and paste the tweet IDs.
- If you'd prefer to look at your own search, stop here. If you'd rather view the results of the borrowed archive, click the "TAGS" tab and select "Wipe Archive Sheet."
- Go to the "TAGS" tab and select "Build archive from Tweet IDs" again. When prompted, identify whether the first row contains a header or a tweet ID. Depending on the size of the dataset, this will take several minutes to hydrate.

Results

What Is Our Data?

Our data is extremely tidy: each observation - in this case, each tweet - is a single row in the spreadsheet and most variables - in this case, a piece of information about that tweet - have their own column. This tidy format will greatly expedite any future quantitative analysis.

The archive sheet can be re-organized in the same ways as any other spreadsheet. Try sorting based on `favorite_count` by clicking the arrow at the top of the column. Sorting by `retweet_count` may also be of interest, as the tweet with the most retweets in any archive is probably not the most retweeted tweet in that archive. Sort by `user_followers_count` to focus only on tweets from users whom many other users follow.

What Is Our Data?

As small and ephemeral as individual tweets seem, we can see that each is quite data-rich. A few things to note:

- The `geo_coordinates` field is often rather poor; `user_location` is often better, but users will often make up all sorts of locations. This limits and skews the possibilities for geo-spatial analysis without quite eliminating it.
- Each observation (or row) is indeed a single tweet, but note that each retweet counts as its own observation. Retweets are distinguished by an RT and the user name of the original poster (never the user name of an intermediary retweet).
- Tweets in reply to other tweets are distinguished by the existence of values for the `in_reply_to_user_id_str` variable, which is otherwise blank. Replies could be eliminated in further analysis on this basis if desired.

Summary and Dashboard

TAGS includes several built-in tools for representing the general contours of any archive it produces. Click the "TAGS" tab on the top navigation bar and select "Add summary sheet." Click on the new sheet in the lower left and then select the red "Off" button by "Sheet Calculation" to turn it "On." This page shows a number of general statistics about the archive as well as a user breakdown for each user of the total number of their tweets, their retweets, and replies to them that contained the search term.

Return to the "TAGS" tab and now select "Add dashboard sheet." The Dashboard sheet includes a couple more useful summaries, in particular the Tweet Volume Over Time graph and the table for the most retweeted tweets in the archive.

TAGS Archive

The initial Readme/Settings sheet contains two additional interfaces for exploring the collected tweets: TAGS Archive and TAGSExplorer. In order to use these, it's necessary to make the spreadsheet public. This is the same process as sharing any other Google Doc. If you aren't familiar, follow these commands as prompted beginning in the upper navigation bar: File > Share > Get sharable link > Anyone with the link can view > Done.

Clicking the TAGS Archive cell and following the link will generate a more Twitter-like viewer than the Archive sheet itself.

TAGSExplorer

The second option, TAGSExplorer, is somewhat more itneresting. It produces a network graph in which each node is a user and each edge (or connection) is a reply. This visualization gives us a clearer idea of the major actors in a particular discursive network. In the lower right you can click to add (dotted) edges for mentions and (blue) edges for retweets. The tabs in the upper left overlay related graphs.

Further Analysis?

TAGS is nice, but its options for analysis are preset, and there are many avenues for analysis beyond what it supports. This might include more robust network analysis, geo-spatial analysis, or analysis of the textual content of each tweet, each of which would require additional software or coding. We don't have time in this workshop to explore these possibilities, but if you attend the semester's final hands-on Data 101 workshop you'll learn a variety of methods for analysis in the R programming language; you can even use the Twitter archive you produced today as sample data. (Alternatively, you can work through this workshop on your own by downloading it from GitHub.)

Wrapping Up

Keep It Coming, or Not!

One of the handiest aspects of TAGS is that it can continue re-running your search at regular intervals automatically. Click the "TAGS" tab in the upper navigation bar, then select "Update archive every hour." To undo this, return to the same place and select "Stop updating archive every hour."

If you want to reset this archive rather than start a new one, click "TAGS" and then "Wipe Archive Sheet."

You can also "Disconnect Twitter Access" in the "TAGS" tab.

Exporting

If you'd like to download your archive for further analysis, click the "File" tab in the upper left and select "Download as." You can download the entire file, with both sheets, as a Microsoft Excel document (`.xlsx`).

I recommend that you instead make sure the Archive sheet is currently open with the tab at the bottom, and then download it as a Comma-separated values file (`.csv`). You can still open the `.csv` file format in spreadsheet editors like Excel or Sheets, but it has the added benefit of being more amenable to computational analysis in programming languages (such as Python or R).

Other Methods of Collection

There are many other possible tools for collecting Twitter data.

For R users, there is the `rvest` package; for Python users, there is the popular `Tweepy` package. These tools allow users to conduct a wider range of analyses and obtain data that more basic tools like TAGS do not (for example, all of an account's followers or all of the accounts a particular account follows). There are even some Python packages that attempt to use web scraping techniques to get around the Twitter API's 3,200 tweet limit on user searches.

Finally, I am indebted to a prior workshop by Francesca Giannetti, which I recommend to anyone looking for a wider range of tools or scholarship.

Sharing

If you intend to publicize or publish Twitter datasets in any way, you'll need to follow the same protocol of dehydrating everything first; consult the Twitter API user agreement. Tweet IDs are stored in TAGS archives as `id_str`.

Thanks for Coming!

We would really appreciate it if you took a minute to fill out our brief feedback survey.

If you'd like to return to this workshop in more detail, visit https://github.com/azleslie/TwitterData.