Feeding Hungry Researchers: A Library Ecosystem for Social Media Data

CNI Spring Meeting 2021

Dan Kerchner
Laura Wrubel

George Washington University Libraries & Academic Innovation
From experiment to ecosystem

- Developed initially to help one professor at GW, not self-service
- More faculty, students, researchers
- Rebuild as a self-service web application (w/grants from IMLS, NHPRC, CEAL)
- Consultations
- Workshops
- External collaborations
- TweetSets to facilitate access to, and subsetting of data sets
- Publication of data sets on Dataverse
From experiment to ecosystem

Consultations

Workshops

Research Collaborations

SFM 2.0

TweetSets

Dataverse
Social Feed Manager

- Self-service (following consultation)
  - set up Twitter credentials
  - specify collecting strategy
  - export data as CSV, Excel, JSON, etc.
Access to Twitter Data Sets: TweetSets and Dataverse

Select source datasets
Select the source datasets from which to create a new dataset.

- 2016 United States Presidential Election: Democratic Candidate User Timeline
- 2016 United States Presidential Election: Democratic Convention
- 2016 United States Presidential Election: Democratic Party User Timeline
- 2016 United States Presidential Election: Election Day Filter
- 2016 United States Presidential Election: Election Filter
- 2016 United States Presidential Election: First Presidential Debate Filter
- 2016 United States Presidential Election: Republican Candidate User Timeline
- 2016 United States Presidential Election: Republican Convention
- 2016 United States Presidential Election: Republican Party User Timeline
- 2016 United States Presidential Election: Second Presidential Debate Filter

Select tweets in dataset
List the tweets selected from the source datasets. Once complete, create the dataset.

Tweet text
For a selected or excluded tweet, text includes the text of the source tweet.

Contains all
Test must contain all of these terms (AND). Comma separate multiple terms. Spaces separate words are treated as wildcards.

Contains any
Test must contain one of these terms (OR). Comma separate multiple terms. Spaces separate words are treated as wildcards.

Excludes
Text may not contain (NOT). Comma separate multiple terms. Spaces separate words are treated as wildcards.

Hash tags
Test must contain one of these hash tags (OR). # is optional. Case insensitive.

Mentions
Test must contain one of these mentions (OR). @ is optional. Case sensitive.

Any of these source names must be mentioned (OR). If 0 is optional. Case sensitive.

Description
This dataset contains the tweet ids of 36,405,486 tweets related to Coronavirus or COVID-19. They were collected between March 3, 2020 and December 3, 2020 from the Twitter API using Social Feed Manager.

Please note that this is VERSION 6 of this data set. See the Versions tab below for all versions.


These tweets were collected using the FIDST dataset filter method of the Twitter Stream API, using the track parameter with the following keyword: Coronavirus, COVID19 because of the size of the collection, the list of identifiers is split into 24 files of up to 10 million tweets each, with a tweet identifier on each line. There is a covid19filter README on containing additional documentation of how the tweets were collected.

Data from the first and last days of the collection do not represent complete days.

The GMT timestamp method suggests removing the complete tweet for a tweet id, known as hybirding. Tools such as tweet or hybird can be used to hydrate tweets.

Per Twitter’s Developer Policy, tweet ids may be publicly shared for academic purposes; tweets may not.
Library Services

- 1:1 Social Media research consultations
- "Collecting Social Media Data" workshops
- Guest lectures in GW courses
  - First-year university writing program
  - Data Analytics/Text mining
  - Political Science
- Research collaborations
What do we collect?

- Coronavirus (354M tweets)
- 2020 U.S. Presidential election
- 2018 U.S. midterm election
- 2016 U.S. election (280M tweets)
- U.S. Congress (all senators and representatives)
- U.S. Federal govt (3,000 U.S. government accounts)
- Hurricanes

- News outlets (4,500 media organization accounts)
- Climate Change
- Immigration and travel ban
- Women's March
- #WITBragDay
- Charlottesville
- Winter Olympics
- Ireland 8th amendment
Tracking and understanding impact

- Internal tracking of research support and SFM usage
- Dataset downloads from TweetSets and Dataverse
- Email from external users
- Publication searches for usage of SFM, TweetSets or Dataverse citations
## Source tweetsets: number of users

<table>
<thead>
<tr>
<th>Source dataset</th>
<th>All users</th>
<th>All users (last 6 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Outlets</td>
<td>1,178</td>
<td>162</td>
</tr>
<tr>
<td>Coronavirus Filter</td>
<td>1,087</td>
<td>539</td>
</tr>
<tr>
<td>End of Term 2016 U.S. Government Twitter Archive</td>
<td>509</td>
<td>15</td>
</tr>
<tr>
<td>2016 United States Presidential Election: Election Filter</td>
<td>487</td>
<td>51</td>
</tr>
<tr>
<td>Climate change</td>
<td>388</td>
<td>121</td>
</tr>
<tr>
<td>Women's March</td>
<td>291</td>
<td>42</td>
</tr>
<tr>
<td>2016 United States Presidential Election: Democratic Candidate User Timelines</td>
<td>268</td>
<td>49</td>
</tr>
</tbody>
</table>
2016 United States Presidential Election Tweet Ids

This dataset contains the tweet ids of approximately 280 million tweets related to the 2016 United States presidential election. They were collected between July 13, 2016 and November 10, 2016 from the Twitter API using Social Feed Manager.

These tweet ids are broken up into 12 collections. Each collection was collected either from the GET statuses/user_timeline method of the Twitter REST API or the POST statuses/filter method of the Twitter Stream API. The collections are:

- Candidates and key election hashtags (Twitter filter): election-filter[1-6].txt
- Democratic candidates (Twitter user timeline): democratic-candidate-timelines.txt
- Democratic Convention (Twitter filter): democratic-convention-filter.txt
- Democratic Party (Twitter user timeline): democratic-party-timelines.txt
- Election Day (Twitter filter): election-day.txt
- First presidential debate (Twitter filter): first-debate.txt
- GOP Convention (Twitter filter): republican-convention-filter.txt
- Republican candidates (Twitter user timeline): republican-candidate-timelines.txt
- Republican Party (Twitter user timeline): republican-party-timelines.txt
- Second presidential debate (Twitter filter): second-debate.txt
- Third presidential debate (Twitter filter): third-debate.txt
- Vice Presidential debate (Twitter filter): vp-debate.txt

There is also a README.txt file for each collection containing additional documentation on how it was collected.
GW research using tweet datasets


Worldwide usage of tweet datasets


Research support and collaborations

● Involved at start of project as researchers plan data collection and methods

● Supporting social media research more broadly
  ○ Analyzing hate speech in tweets by candidates for U.S. Congress
  ○ Analysis of COVID-19 "misinfodemic" claim: Do social media posts about COVID-19 contain more, or fewer, links to credible sources than posts about other topics?
  ○ Support for Institute for Data, Democracy & Politics
Research support and collaborations

“A queer black woman invented rock-and-roll”: Rosetta Tharpe, memes, and memory practices in the digital age

Gayle Wald
American Studies Department, George Washington University, Washington, DC, USA

ABSTRACT
In recent years, musician Sister Rosetta Tharpe (1915–73) has gone from being a “forgotten” rock trailblazer to being a viral video sensation and the subject of the Internet meme “A queer black woman invented rock and roll.” This article traces the “meme-ing” of Tharpe, paying attention to how the meme constructs, circulates, and mobilizes knowledge about her in online and offline spaces. After briefly reviewing the challenges of writing about Tharpe’s sexuality in “Shout, Sister, Shout!” my 2007 biography, I turn to the history of the meme, showing how it supports a radical politics of remembering by articulating Tharpe’s erasure from the annals of rock to the ongoing struggles of marginalized black and queer subjects. In particular, the boundary-pushing musicians Janelle Monáe and Lizzo and the visionary Haitian-American fashion designer Kerby Jean-Raymond illustrate how the meme has fueled engagements with the erasure, appropriation, and commodification of black queer creativity in the 21st century. These examples show how contests over rock-and-roll history and memory “vibrate,” or continue to resonate, across media in the post-rock moment.

ARTICLE HISTORY
Received 5 July 2020
Revised 11 November 2020
Accepted 19 November 2020

KEYWORDS
Supporting social media archiving and research

- Education for other libraries and archives
  - Continuing Education to Advance Web Archiving (CEDWARC)
  - Webinars for Digital Preservation Coalition
- Archives Unleashed
- DocNow community
Benefits to campus researchers

1. Researchers get data they need for research questions from a range of disciplines.
2. Access to otherwise unavailable free historical data.
3. Opens up access to data to users of all technical backgrounds.
4. Students learn about data considerations.
5. Datasets easily available for teaching data skills to students.
Benefits to the library

1. Positions library as data creator and data publisher
2. Leading in collecting datasets of cultural value
3. Opportunities to become more deeply involved with research projects
4. Library part of conversation around data ethics and privacy
5. Librarians and developers get firsthand experience and involvement with research support using larger datasets
Challenges in providing services and support

● Time and effort
  ○ Within GW: Consultations, Workshops, etc.
  ○ Beyond GW: External inquiries/requests
  ○ Software maintenance

● Making work visible, measurable, and making its value clear

● Policy - ensuring appropriate data use / privacy

● Changing Twitter API

● Tweetsets is a computationally-intensive public service
Future Directions

- Community Sprint for software development
- Twitter datasets on GW ArchiveSpace
- Accessioning data sets
Get in touch with us!

sfm@gwu.edu

Social Feed Manager project: gwu-libraries.github.io/sfm-ui
Tweetsets: tweetsets.library.gwu.edu
GWU Libraries @ Harvard Dataverse: dataverse.harvard.edu/dataverse/gwu-libraries