Social Feed Manager

Digital Preservation Coalition
Web Archiving & Preservation Webinar Series
February 6, 2020

Dan Kerchner
kerchner@gwu.edu
@DanKerchner

Laura Wrubel
lwrubel@gwu.edu
@liblaura
Agenda

- Project background
- Current uses
- Collecting from Twitter
- SFM walkthrough
- Ethical and technical considerations
Social Feed Manager software

- Open source software by GW Libraries.
- User interface for collecting, managing, and exporting social media data.
- Collect from Twitter, Tumblr, Flickr, Sina Weibo.
- Libraries run this for their users as a service. (Not typically a local install on your laptop.)

More: go.gwu.edu/sfm
Project Background

- Initial development: 2013
- Version 1.0: 2016
- Grant funding:
  - IMLS
  - National Historical Publications & Records Commission (NARA)
  - CEAL
Institutions using SFM include

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Stanford University

VIRGINIA TECH™

PennState

RADCLIFFE INSTITUTE FOR ADVANCED STUDY
HARVARD UNIVERSITY
Congratulations to CCAS alums Sally Nuamah, BA '11, and Tara Dorfman, BA '11, who were both named to Forbes 2019 30 Under 30 lists! #GWCCAS #GWU
Congratulations to CCAS alums Sally Nuamah, BA '11, and Tara Dorman named to Forbes 30 Under 30 lists! #GWCCAS #GWU https://t.co/kPWyKzJ9UX
Collecting from Twitter's APIs

- SFM uses the free standard APIs
- API responses are JSON
- SFM handles rate limiting, authentication, organizing collections
- Requires one set of Twitter developer keys for the application and Twitter authentication for each user collecting data.
Twitter collection types

- User timeline: up to 3,200 tweets in the past, for each account
- Search: sample of tweets from the past 7-10 days
- Filter stream: real-time, forward-looking only
- Sample stream ~ 0.5% sample of all tweets; ~2GB per day
Collecting from Twitter: Data Extracts

- JSON
- CSV and Excel (selected fields from JSON)
- Filtered by date range, screen name, etc.
<table>
<thead>
<tr>
<th>id</th>
<th>tweet_url</th>
<th>created_at</th>
<th>parsed_cre</th>
<th>user_scren</th>
<th>text</th>
<th>tweet_type</th>
<th>coordin</th>
<th>hashtags</th>
<th>media</th>
<th>urls</th>
<th>favorite_cc</th>
</tr>
</thead>
<tbody>
<tr>
<td>106970559017435</td>
<td><a href="https://twitter.com/KyrahAltman">https://twitter.com/KyrahAltman</a>, a senior majoring in human services and social justice, was a original</td>
<td>Mon Dec 0 2018-12-03</td>
<td></td>
<td></td>
<td>CCASOutFr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>106967441805249</td>
<td><a href="https://twitter.com/KyrahAltman">https://twitter.com/KyrahAltman</a>, a senior majoring in human services and social justice, was a original</td>
<td>Mon Dec 0 2018-12-03</td>
<td></td>
<td></td>
<td>CCASOutFr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1069344206655754</td>
<td>The new #GWCCAS International Buddy Program wants to help internation original</td>
<td>Sun Dec 02 2018-12-02</td>
<td></td>
<td></td>
<td>GWCCAS G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>106926719931821</td>
<td>Congratulations to CCAS alums Sally Numaah, BA '11, and Tara Dorfman, Br original</td>
<td>Sun Dec 02 2018-12-02</td>
<td></td>
<td></td>
<td>GWCCAS G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1068969237308424</td>
<td>A #GWCCAS junior and her peers got the opportunity to see Michelle Obama original</td>
<td>Sat Dec 01 2018-12-01</td>
<td></td>
<td></td>
<td>GWCCAS O</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>106888619715248</td>
<td>Last week, @CorcoranGW hosted Washington D.C.'s annual National PortFoi original</td>
<td>Sat Dec 01 2018-12-01</td>
<td></td>
<td></td>
<td>GWU GWC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>in_reply_to in_reply_to in_reply_to lang</th>
<th>place</th>
<th>possibly_reply</th>
<th>retweet_cc</th>
<th>retweet_or_quot</th>
<th>retweet_or_quot</th>
<th>retweet_or_quot</th>
<th>source</th>
<th>user_id</th>
<th>user_create</th>
<th>user_default</th>
<th>user_description</th>
<th>user_follow</th>
<th>user_friends</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>FALSE</td>
<td>0</td>
<td>1</td>
<td>106966607483C</td>
<td>CorcoranGW</td>
<td>216058845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>FALSE</td>
<td>0</td>
<td>1</td>
<td>106966607483C</td>
<td>CorcoranGW</td>
<td>216058845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>FALSE</td>
<td>3</td>
<td>1</td>
<td>106966607483C</td>
<td>CorcoranGW</td>
<td>216058845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>FALSE</td>
<td>0</td>
<td>1</td>
<td>106966607483C</td>
<td>CorcoranGW</td>
<td>216058845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>FALSE</td>
<td>1</td>
<td>1</td>
<td>106966607483C</td>
<td>CorcoranGW</td>
<td>216058845</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Log In

If you have not created an account yet, then please sign up first.

Username

Password

Forgot your password?

☐ Remember me

Log In

Collecting and using data from social media platforms is subject to those platforms' terms (Twitter, Flickr, Sina Weibo, Tumblr), as you agreed to them when you created your social media account. Social Feed Manager respects those platforms' terms as an application (Twitter, Flickr, Sina Weibo, Tumblr).

Social Feed Manager provides data to you for your research and academic use. Social media platforms' terms of service generally do not allow republishing of full datasets, and you should refer to their terms to understand what you may share. Authors typically retain rights and ownership to their content.

In addition to respecting the platforms' terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are some ethical and privacy guidelines you may want to consider.
Credentials

Credentials are used to authorize Social Feed Manager to collect data from Twitter.
Authorize Social Feed Manager by connecting your account or adding credentials.

Twitter Credential   Flickr Credential   Weibo Credential   Tumblr Credential

Name                  Date Added
GW Kerchner’s twitter credential   Oct. 12, 2018, 10:31:34 a.m. EDT

Connect Twitter Account   Add Twitter Credential

Authorize Social Feed Manager (GW Sandbox) to use your account?

Username or email
Password

Remember me   Forgot password?

Sign In   Cancel

This application will be able to:
- Read Tweets from your timeline.
- See who you follow.

Will not be able to:
- Follow new people.
- Update your profile.
- Post Tweets for you.
- Access your direct messages.
- See your email address.
- See your Twitter password.
GW Twitter Accounts

Data collected: 0 files (0 bytes)

Collections

No collections yet.

Add Collection

- Add Twitter user timeline
- Add Twitter search
- Add Twitter filter
- Add Twitter sample
- Add Tumblr blog posts
- Add Flickr user
- Add Weibo timeline
Add Twitter user timeline

* indicates a required field.

Collection name*
GWU official accounts

Description

Public link
Link to a public version of this collection, e.g., in a data repository.

Credential*
GWKerchner's twitter credential

Schedule*
Every week

End date
If blank, will continue until stopped.

Sharing*
Group only

Who else can view and export from this collection. Select "All other users" to share with all Social Feed Manager users.

Change Note

Further information about this addition.

Save  Cancel
Request Export

Seed choice*
- All seeds
- Active seeds only
- Selected seeds only

Export format*
- Excel (XLSX)

Maximum number of items per file
- 250,000

- Deduplicate (remove duplicate posts)

Limit by item date range
- Item date start
- Item date end

The timezone for dates entered here are America/New_York. Adjustments will be made to match the time zone of the items. For example, dates in tweets are UTC.

Limit by harvest date range
- Harvest date start
- Harvest date end

Collecting and using data from social media platforms is subject to those platforms' terms (Twitter, Flickr, Sina Weibo, Tumblr), as you agreed to them when you created your social media account. Social Feed Manager respects those platforms' terms as an application (Twitter, Flickr, Sina Weibo, Tumblr).

Social Feed Manager provides data to you for your research and academic use. Social media platforms' terms of service generally do not allow republishing of full datasets, and you should refer to their terms to understand what you may share. Authors typically retain rights and ownership to their content.

In addition to respecting the platforms' terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are some ethical and privacy guidelines you may want to consider.
GWU official accounts

Twitter user timeline

New collection added. You can now add seeds.

At least 1 active seed must be added before harvesting can be turned on.

Data collected: 0 files (0 bytes)

<table>
<thead>
<tr>
<th>Active</th>
<th>Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Link</th>
<th>Twitter accounts</th>
<th>User ID</th>
<th>Messages</th>
</tr>
</thead>
</table>

Add Seed  Bulk Add Seeds
Add Twitter user timeline seeds

**Seeds type**

- [ ] Screen Name
- [x] User id

**Bulk Seeds**

- gw_sports
- GWmedia
- GWtweets
- gwucolumbian

Enter each seed on a separate line.

**Change Note**

Further information about this addition.

[Save] [Cancel]
**GWU official accounts**

Twitter user timeline
Collection is active. Turn off to edit.

Next harvest at Dec. 4, 2018, 5:29:18 p.m. EST

**Data collected:** 1 file (4.7 MB)

**Stats:**
- tweets: 12,865

**Details**

### Seeds

<table>
<thead>
<tr>
<th>Active</th>
<th>Deleted</th>
<th>Download seed list</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Link</th>
<th>Twitter accounts</th>
<th>User ID</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>@</td>
<td>GWmedia</td>
<td>59863790</td>
<td></td>
</tr>
<tr>
<td>@</td>
<td>gw_sports</td>
<td>54627173</td>
<td></td>
</tr>
<tr>
<td>@</td>
<td>GWtweets</td>
<td>28101965</td>
<td></td>
</tr>
<tr>
<td>@</td>
<td>gwucolumbian</td>
<td>54639174</td>
<td></td>
</tr>
</tbody>
</table>
Export files for GWU official accounts

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>6b0da4c9bb0e4087a48cb793effe0bab-README.txt</td>
<td>2.0 KB</td>
</tr>
<tr>
<td>6b0da4c9bb0e4087a48cb793effe0bab_001.xlsx</td>
<td>2.8 MB</td>
</tr>
</tbody>
</table>

See the data dictionary for more information about the fields in the export.

See the guidance on citing SFM and datasets.

**Status:** Success

**Format:** xlsx

**Selected seeds:** All seeds

Collecting and using data from social media platforms is subject to those platforms’ terms (Twitter, Flickr, Sina Weibo, Tumblr), as you agreed to them when you created your social media account. Social Feed Manager respects those platforms’ terms as an application (Twitter, Flickr, Sina Weibo, Tumblr).

Social Feed Manager provides data to you for your research and academic use. Social media platforms’ terms of service generally do not allow republishing of full datasets, and you should refer to their terms to understand what you may share. Authors typically retain rights and ownership to their content.

In addition to respecting the platforms’ terms, as a user of Social Feed Manager and data collected within it, it is your responsibility to consider the ethical aspects of collecting and using social media data. Your discipline or professional organization may offer guidance. Here are some ethical and privacy guidelines you may want to consider.
Collection Sets

A collection set is a group of collections around a particular topic or theme. Collections sets are active when there is at least one active collection within them. Collection sets are inactive when all collections have been deactivated and are no longer harvesting.

<table>
<thead>
<tr>
<th>Name</th>
<th>Collections</th>
<th>Date Added</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>115th U.S. Congress</td>
<td>3 collections</td>
<td>Jan. 27, 2017, 10:47:12 a.m. EST</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Alaska Earthquake</td>
<td>2 collections</td>
<td>Nov. 30, 2018, 1:37:25 p.m. EST</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>China Anti-Corruption</td>
<td>48 collections</td>
<td>June 29, 2016, 11:49:14 p.m. EDT</td>
<td>CEAL Grant</td>
</tr>
<tr>
<td>Climate change</td>
<td>1 collection</td>
<td>Sept. 21, 2017, 8:11:54 a.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Corcoran</td>
<td>1 collection</td>
<td>March 29, 2017, 5:17:54 p.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Foreign leaders</td>
<td>1 collection</td>
<td>March 11, 2018, 9:43:54 p.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
<tr>
<td>Governors</td>
<td>1 collection</td>
<td>March 11, 2018, 9:34:03 p.m. EDT</td>
<td>GW Libraries Scholarly Technology Group</td>
</tr>
</tbody>
</table>
Technical info

Requires:

- Twitter application credentials
- Linux server
- Storage

Social Feed Manager runs as a set of Docker containers

- Application written in Python
- Small database (postgresql) for user info and metadata
- Collected content is stored on the file server as WARC files
Ethical and privacy concerns
Social media data comes from people

• Consider the impact of your work on the creator of the social media. Be thoughtful collecting social media of vulnerable individuals.
• When publishing, get permission from creator for quotes from social media if possible, do not rely on anonymizing posts.
• Respect a user’s deleting their content.
• Be familiar with platform terms of use. Don’t republish full datasets and share in accordance with terms (e.g., tweet ids only)
SFM software

- **Open source on GitHub**
- Issues tracked in the sfm-ui repository: we welcome requests!
- Documentation for installation and use: [sfm.readthedocs.io/](https://sfm.readthedocs.io/)
- Main project site (blog posts and more): [go.gwu.edu/sfm](https://go.gwu.edu/sfm)
Thank you

Social Feed Manager
sfm@gwu.edu
go.gwu.edu/sfm

Dan Kerchner
kerchner@gwu.edu
@DanKerchner

Laura Wrubel
lwrubel@gwu.edu
@liblaura
TweetSets
tweetsets.library.gwu.edu
Select source datasets

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

Other datasets that we’ve collected:
- Members of the 116th Congress
- Healthcare
- Tax reform
- Hurricane Florence
- Hurricane Michael
- Solar Eclipse

Contact us if you’d like any of these loaded into TweetSets.

<table>
<thead>
<tr>
<th>Dataset Description</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016 United States Presidential Election: Democratic Candidate User Timelines</td>
<td>22,251</td>
</tr>
<tr>
<td>2016 United States Presidential Election: Democratic Convention</td>
<td>8,340,668</td>
</tr>
<tr>
<td>2016 United States Presidential Election: Democratic Party User Timelines</td>
<td>11,846</td>
</tr>
<tr>
<td>2016 United States Presidential Election: Election Filter</td>
<td>251,077,140</td>
</tr>
<tr>
<td>2016 United States Presidential Election: First Presidential Debate Filter</td>
<td>3,182,249</td>
</tr>
<tr>
<td>2016 United States Presidential Election: Republican Candidate User Timelines</td>
<td>52,395</td>
</tr>
<tr>
<td>2016 United States Presidential Election: Republican Convention</td>
<td>6,248,658</td>
</tr>
</tbody>
</table>