Journal of East Asian Libraries

Volume 2017 | Number 165 Article 12

10-2017

# Collecting Social Media Data from the Sina Weibo Api

Justin Littman

Daniel Kerchner

Yan He

Yecheng Tan

Cathy Zeljak

*Journal of East Asian Libraries*, No. 165, Oct. 2017

# Collecting Social Media Data from the Sina Weibo Api

**Justin Littman**
**Daniel Kerchner**
**Yan He**
**Yecheng Tan**
**Cathy Zeljak**

## Abstract

As part of a larger collaborative project to preserve non-official voices in China's anti-corruption campaign, a team from George Washington University Libraries was tasked with collecting content from Sina Weibo, a massively popular Chinese social media site. Based on their experience developing Social Feed Manager (SFM), an open-source tool for collecting social media data to support research and constructing archives, the team added support for Sina Weibo. This paper provides introductions to both the Sina Weibo platform and the approach of collecting social media from an Application Programming Interface (API). Further, it describes SFM's support for Sina Weibo, the challenges in working with the Sina Weibo API, guidance for other institutions interested in conducting similar work, and a brief characterization of the content collected about China's anti-corruption campaign.

## Introduction

Sina Weibo[1] is a massively popular Chinese social media site, in use by over 100 million users per day (Freier 2015). Given its prominent role in communicating the "everyday voices" of the Chinese public, the value of collecting content from Sina Weibo was self-evident when a team from Johns Hopkins University, George Washington (GW) University, and Georgetown University initiated a collaborative project to preserve non-official voices in China's anti-corruption campaign. (For background on and motivation for the grant, see the companion report, "Web-archiving Chinese Social Media: Final Project Report.") Awarded a Mellon innovation grant administered by the Council on East Asian Libraries,[2] this team took a two-prong approach:

1. Collect blog content using Archive-It, a service provided by the Internet Archive for performing web archiving.
2. Enhance Social Feed Manager (SFM), an open-source tool developed at GW Libraries, to collect weibos (posts) from the application programming interface (API) of Sina Weibo.

[1] http://weibo.com/
[2] http://www.eastasianlib.org/MellonGrants.htm

The goal of this paper is to explore the second prong, the enhancement of Social Feed Manager, by:

1. Giving a brief overview of Sina Weibo.
2. Explaining API-based social media collecting.
3. Describing the affordances of the Weibo API.
4. Providing an overview of SFM and the Weibo social media harvester.
5. Document some challenges in working with the Weibo API.
6. Characterize the anti-corruption campaign-related social media data that has been collected to date.

**Sina Weibo**

Sina Weibo is a micro-blogging social media platform similar to Twitter. The core of Sina Weibo is the "weibo" ("微博", meaning microblog), a post analogous to a tweet. Users on Sina Weibo perform two main functions with weibos: they author weibos and they read weibos.

When authoring a weibo, users are restricted to 2000 Chinese characters, an increase from the earlier limitation of 140 Chinese characters, which mirrored Twitter's character limit. However, when viewing a weibo, only the first 140 Chinese characters are displayed, with a link allowing the rest of the text to be viewed. To mention another user, the user's username is preceded by an "@". To mark a term in a weibo as a topic, it is bracketed by "#" (similar to Twitter's hashtag -- more on this below). Weibos can include URLs; embedded images or multimedia items; and/or emoticons.

English version:


Chinese version:


The primary mechanism for reading weibos is a user's timeline. A timeline is a list of weibos that is primarily determined by the user accounts that the user follows. A user's timeline contains the weibos posted by the followed user accounts. Upon reading a weibo, a user can comment on it or repost the weibo, similar to retweeting on Twitter.

An example user page would appear as follows:


Like other social media platforms, Sina Weibo can be accessed from a website or smartphone application (i.e., mobile). Sina Weibo has other social features, but those described above are the most relevant for this discussion.

**API-based social media collecting**

When a user goes to a website of a social media platform, e.g., http://us.weibo.com/gb, she interacts with the web interface of the platform. This interface is intended to facilitate the interaction between a human user (using a web browser) and the social media platform. However, many social media platforms provide an additional interface, the API, to facilitate the interaction between software and the social media platform. So, for example, when the Twitter app on your phone needs to fetch your tweets, it requests those tweets from Twitter's API. Unlike the web interface, which returns HTML that is optimized for rendering in a browser, the API returns structured data, usually in a format called JSON, which is optimized for processing by software.[3]

Both the web interface and the API can be used to collect content from a social media platform. Collecting from the HTML/web interface requires using web harvesting software such as Heritrix[4] or WebRecorder;[5] collecting from an API requires software such as SFM or Twarc.[6] Social media harvesting and web harvesting can be considered complementary, with different strengths and weaknesses (Littman et al. 2016). Some of the strengths of collecting from the API include:

- Data collected by the API is in a structured format such as JSON or XML. These formats are amenable to research that involves applying computational techniques. (This will be more evident when we look at the anatomy of a weibo below.) Extracting similar data from web pages would be extremely difficult.
- Some social media APIs provide metadata that is not available from the website. For example, the Twitter API provides the name of the device/application used to author a tweet.
- Social media platforms tend to keep their APIs stable and when making a change will generally announce it in advance. This is because there are potentially numerous other application (e.g., smartphone apps) that depend on the API. In contrast, social media web pages regularly change in small ways that make extracting data problematic.
- To optimize the user experience, social media websites make heavy use of Javascript code and other client-side techniques that are not amenable to web harvesting. Because of this many web harvesters do a poor job of capturing social media websites (Thomson 2016).
- Social media data can generally be collected more efficiently from the API than the website. Some APIs expose API methods that crosscut the social media data in ways

[3] http://json.org/
[4] https://github.com/internetarchive/heritrix3
[5] https://webrecorder.io/
[6] https://github.com/docnow/twarc. Twarc is a tool for collecting from the Twitter API. While web harvesting software is generally intended to collect content from any website, social media harvesting software is not general-purpose; it can usually only collect from one to a small number of APIs.

that aren't possible from the website. For example, via its API, Twitter allows filtering the stream of all tweets currently passing through the platform.

Some of the weaknesses of collecting from an API include:

- Not all social media platforms have complete, public APIs. This will be discussed further below as it relates to the Weibo API.
- Each API is different.
- Some platforms, notably Twitter, place explicit requirements and limits on the use and sharing of data harvested using the API. These may be more explicit or limiting than the terms of service for the websites.
- Limitations placed by the platform on the amount of data that can harvested via the API can make it difficult or impossible to capture older content.

**Affordances of the Weibo API**

With this basic understanding of collecting from the API of a social media platform, let's consider the affordances of the Weibo API. For starters, here is part of a weibo retrieved from the Weibo API[7]:

{
"created_at": "Thu May 11 15:43:13 +0800 2017",
"id": 4106245138832598,
"text": "#民生服务# 【中央机关公开遴选选调360名公务员】2017年中央机关公开遴选和公开选调公务员工作今日开始报名。此次公开遴选和公开选调共有56个中央机关参加计划选拔360名公务员。[心]报名时间截止5月22日18:00。[心]笔试时间为2017年6月25日[心]考试地点设在北京、上海、西安、兰州等17个城市。详情 ",
"textLength": 314,
"source": "<a href=\"http://app.weibo.com/t/feed/6ghA0p\" rel=\"nofollow\">搜狗高速浏览器</a>",
"favorited": false,
"truncated": false,
"in_reply_to_status_id": "",
"in_reply_to_user_id": "",
"in_reply_to_screen_name": "",
"geo": null,
"user": {
"id": 5083132536,
"screen_name": "西安国际港务区",

[7] The full JSON for this weibo is available at: https://gist.github.com/justinlittman/db0280402a2ded54cc3d539a798c16e0. For comparison, here is the JSON for a tweet: https://gist.github.com/justinlittman/462a398d161002a8caff0905bf4e5f7f.

```
            "name": "西安国际港务区",
            "province": "61",
            "city": "1000",
            "location": "陕西",
            "description":  "承东启西贯通欧亚西部雄心全球视野中国有了内陆港——西安国际港务区。",
            "domain": "guojigangwuqu",
            "weihao": "",
            "gender": "m",
            "followers_count": 22224,
            "friends_count": 278,
            "pagefriends_count": 12,
            "statuses_count": 11597,
            "favourites_count": 1570,
            "created_at": "Mon Mar 24 13:42:11 +0800 2014"
        },
        "reposts_count": 0,
        "comments_count": 1,
        "attitudes_count": 1,
        "isLongText": true,
        "url_objects": [{
            "url_ori": "http://t.cn/Rax18fd",
            "object_id": "3000000225:8dfcfb1971c9f2e4801cd99705139555",
            "info": {
                "url_short": "http://t.cn/Rax18fd",
                "url_long": "http://www.gov.cn/xinwen/2017-05/11/content_5192758.htm"
            }
        }]
}
```


The meanings of some fields are described in the API documentation[8] (Sina Corp 2017).

Here is how this weibo is rendered on the Sina Weibo website:

[8] http://open.weibo.com/wiki/2/statuses/home_timeline


As mentioned previously, the weibo is in JSON, a simple format that is widely used to exchange data on the web. In the case of a weibo, some of the more significant fields include:

- created_at: A timestamp for when the weibo was posted.
- id: A unique identifier for the weibo.
- text: The text of the weibo.
- user: Information on the author.
- reposts_count: The number of times the weibo has been reposted by other users. This is as of the time that the weibo was retrieved from the API and may change over time.
- comments_count: The number of comments by other users on the weibo. This also may change over time.
- geo: Optional geotagging of the location from which the weibo was posted.
- isLongText: Marks whether the JSON text field is truncated text or is long text. When true, the text of the weibo contained in the text field is truncated to 140 characters. The API does not provide a mechanism for getting the complete text of the weibo, which is a significant limitation.

Note that in this example, the text of the weibo is 161 characters, while the entire JSON-formatted weibo is 10,460 characters.

Like other social media APIs, Sina Weibo provides a number of different methods for interacting with the API. These include functions like "statuses update" for posting a weibo and "friendships destroy" for unfollowing another user (again, think the functions that would be needed by a Weibo smartphone app).[9] For the purposes of collecting social media data, our focus is on methods that allow retrieving weibos. Unfortunately, this is one of the fundamental challenges of collecting content from the Sina Weibo API: the methods for retrieving weibos are extremely limited.

The Weibo API is split into basic methods and advanced methods.[10] Anyone can get access to basic methods; access to the advanced API involves more hurdles and will be discussed further below.

The basic API provides the "friends timeline" method[11], which returns the weibos of the user and the user's friends. (On Sina Weibo, users can have followers. These are more like following a user on Twitter than having a friend on Facebook, as the assent of the friend is not necessary to establish the friend relationship.) The "friends timeline" method can be used as indirect mechanism for collecting the posts of a set of users. So, for example, for the anti-corruption campaign we want to collect the posts of a set of users that we identified that post about the topic. To use the friends timeline method to collect weibos for a single collection, a new, "dummy" account must be created and that account must follow only the desired users.[12]

Using the "friends timeline" method to collect the tweets of other users can be contrasted with Twitter API's "statuses/user_timeline" method[13]. The "statuses/user_timeline" method allows collecting the most recent tweets of any user (if the account is not protected). It does not require the indirection of having to have a "dummy" account follow the users. Put more directly, this approach to collecting with the "friends timeline" method is a creative "kludge" due to the limitations of the Sina Weibo API.

Based on experimentation, we have determined that the friends timeline method only returns the last 150 weibos from the user and the user's friends. (Sometimes it might be 1 or 2 less due to unknown reasons). For example, if User 张三 posts 200 weibos and User 李四

[9] For a complete list, see http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI.

[10] In the documentation, advanced methods are indicated by "高" (advanced).

[11] http://open.weibo.com/wiki/2/statuses/friends_timeline

[12] It might seem like the "user timeline" method (http://open.weibo.com/wiki/2/statuses/user_timeline) would be useful for collecting. However, this API call only return the weibos of the authorized user, requiring that each users need to authorize your application to collect his timeline. This isn't feasible for large scale collecting.

[13] https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline

posts 150 weibos, the result would be the 150 latest weibos among the total 350 weibos. To attempt to compensate for this, weibos can be collected from the "friends timeline" on a regular schedule, e.g., every half hour. Still, weibos collected via this method should be considered a sample and not a complete set of all weibos authored by the user and the user's friends.

The advanced API provides the "topic search" method[14], which returns the most recent 200 posts matching a topic (again, sometimes minus 1 or 2 posts due to unknown reasons). A topic is a single keyword bracketed by "#" in the text of a weibo, e.g., "#keyword#" or "#你好#"; topics on Weibo are like hashtags on Twitter, although they are bracketed by hashtag symbols rather than preceded by a single hashtag symbol.[15] Marking keywords as topics allows users to contribute to a folksonomy on the Sina Weibo platform, as other users can discover their weibos based on the presence of the topic. Since each call to the "topic search" API only returns a limited number of weibos, collecting on a regular schedule is essential to building as complete a topic collection as possible.

Compared to a social media platform such as Twitter, Sina Weibo is rather limited in terms of the API methods that would be desired in order to perform comprehensive collecting. Still, as will be shown below, given the volume of weibos passing through the Sina Weibo platform and when coupled with software such as SFM, the Weibo API does permit creating large scale and valuable datasets for research.

**Social Feed Manager**

Social Feed Manager[16] (SFM) is an open source software developed by GW Libraries[17] that harvests social media data and web resources from Twitter, Tumblr, and Flickr, in addition to Sina Weibo.[18] SFM empowers researchers, faculty, students, and archivists to define and create collections of social media data. While SFM can be run on a laptop, it is intended to be hosted on a server and operated by a collecting organization as a service for the members of its community. For example, GW Libraries operates an instance of SFM that is available to campus researchers, including students, faculty, and library staff.

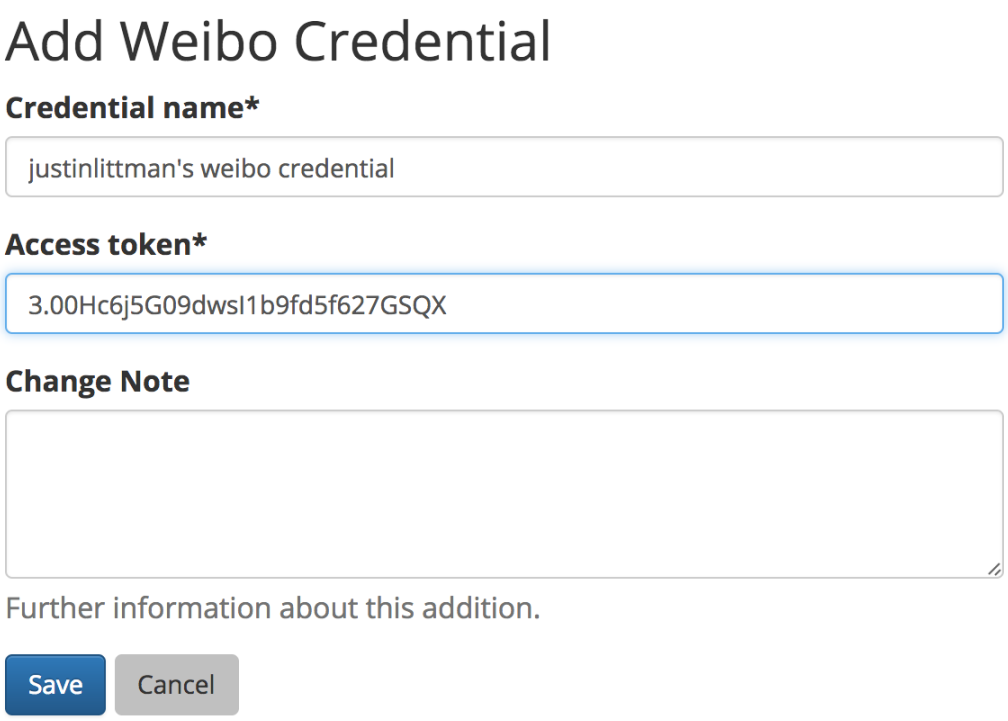
[14] http://open.weibo.com/wiki/2/search/topics
[15] The closing hashtag is needed due to the lack of spacing in Chinese characters.
[16] For project information see http://go.gwu.edu/sfm. For software documentation, see http://sfm.readthedocs.io/en/latest/. For a conceptual and technical overview of SFM, see (Littman et al. 2016).
[17] SFM is supported by grant #NARDI-14-50017-14 from the National Historical Publications and Records Commission.
[18] The Sina Weibo harvester (https://github.com/gwu-libraries/sfm-weibo-harvester) was developed by Yecheng Tan.

One of the primary goals of SFM is to lower the barriers to collecting social media data. This has two aspects: First, users interact with a user-friendly website to create, manage, and export their collections. Users do not need to have technical skills or detailed knowledge of social media APIs (although an understanding of the affordances of the APIs is recommended for successful research). Extensive documentation[19] has been created to assist users in creating social media collections with SFM (GW Libraries 2015).

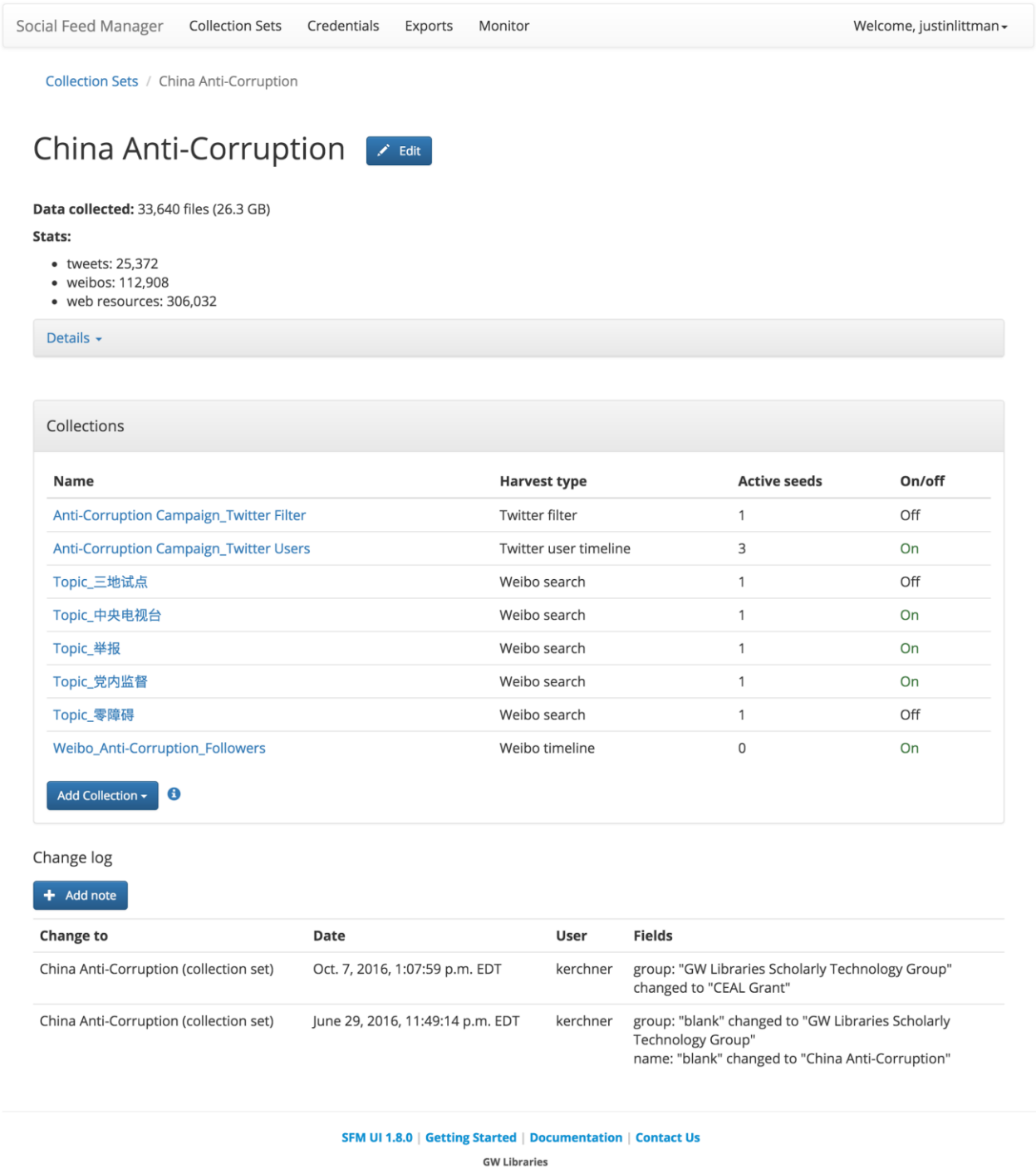
Second, deploying SFM has been made as simple as possible. Using a technology called Docker, SFM can easily be deployed on a server with a minimal number of steps. (Instructions are also provided for deploying "in the cloud" using Amazon Web Services.) Still, SFM does require a moderately technical administrator and the availability of a server. Just as for users, extensive documentation is provided for administering SFM[20].

Setting up a Sina Weibo collection in SFM involves five steps. The first step is to create an account on Sina Weibo's website and follow the users that you would like to collect. The next step is to provide a credential (aka, API key). Like other social media platforms, Sina Weibo requires a credential to access the API. (The challenges of acquiring a credential will be discussed below.) Once a user has a credential, she supplies it to SFM so SFM can use it for collecting.

[19] http://sfm.readthedocs.io/en/latest/#user-docs
[20] http://sfm.readthedocs.io/en/latest/#admin-documentation

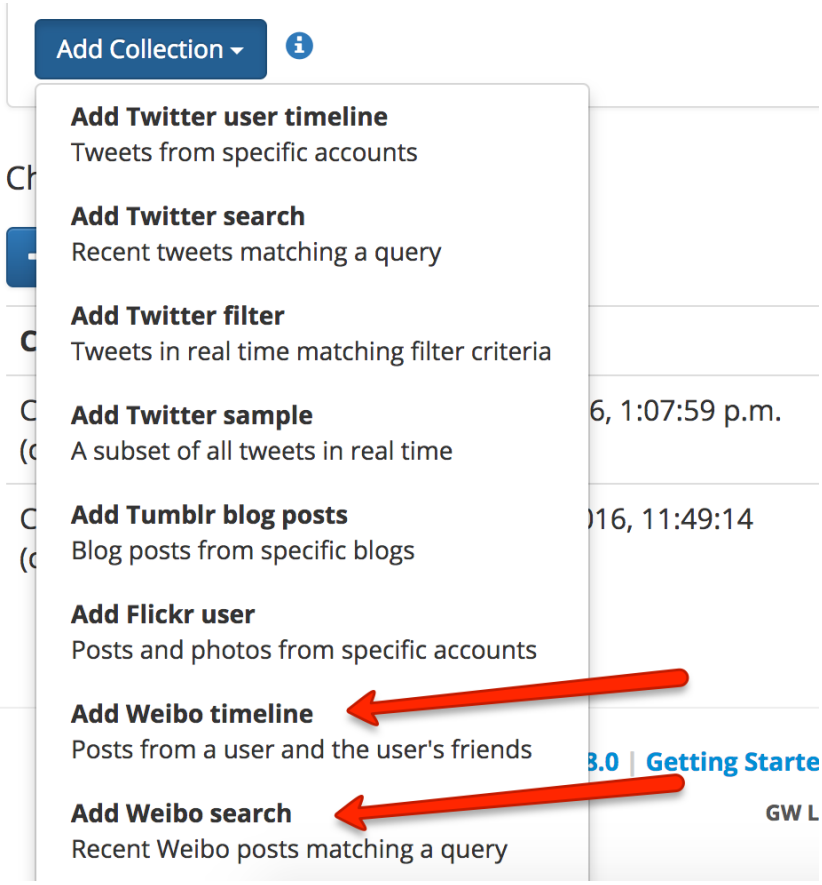



The third step is to create a collection set. A collection set is merely a container for collections. In the case of China's anti-corruption campaign, we have a collection set that contains a Sina Weibo timeline collection, two Twitter collections, and a number of Sina Weibo topic searches. (The list of collections is truncated in this screenshot.)



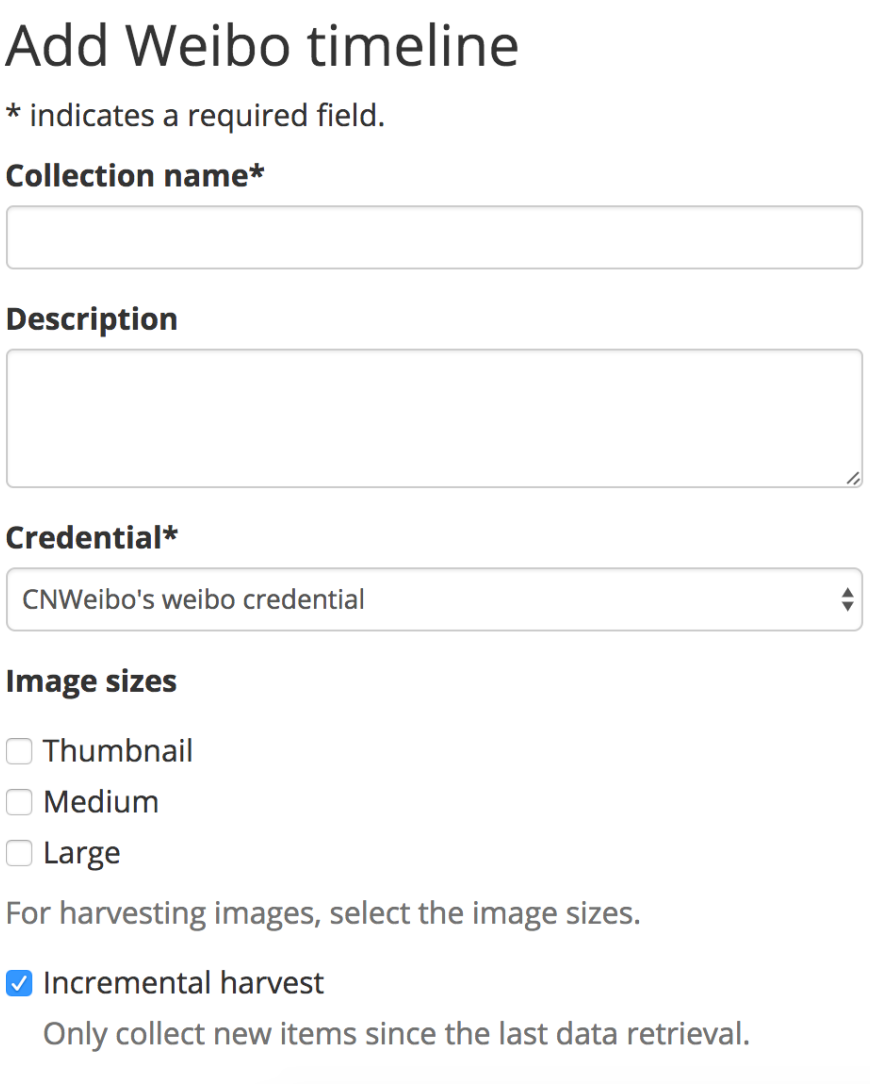

The fourth step is to create a collection. Each collection is for a particular part of a single social media API. For example, for Twitter, SFM supports four different collection types (user

timeline, search, filter, and sample). For Sina Weibo, SFM supports timeline and search[21] collection types corresponding to the API methods described previously.



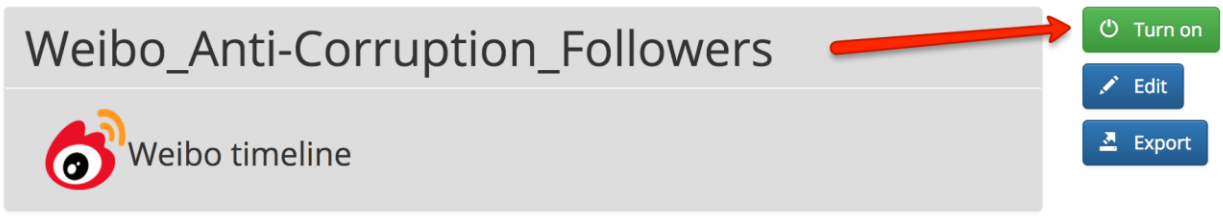

For each collection, the user selects the credential to be used, a schedule (e.g., every 30 minutes, every day, every week), and the collecting parameters (e.g., whether to collect web resources for images or URLs found in weibos).

[21] For reasons that will be obvious when we discuss Sina Weibo challenges, the search collection type is not enabled by default.

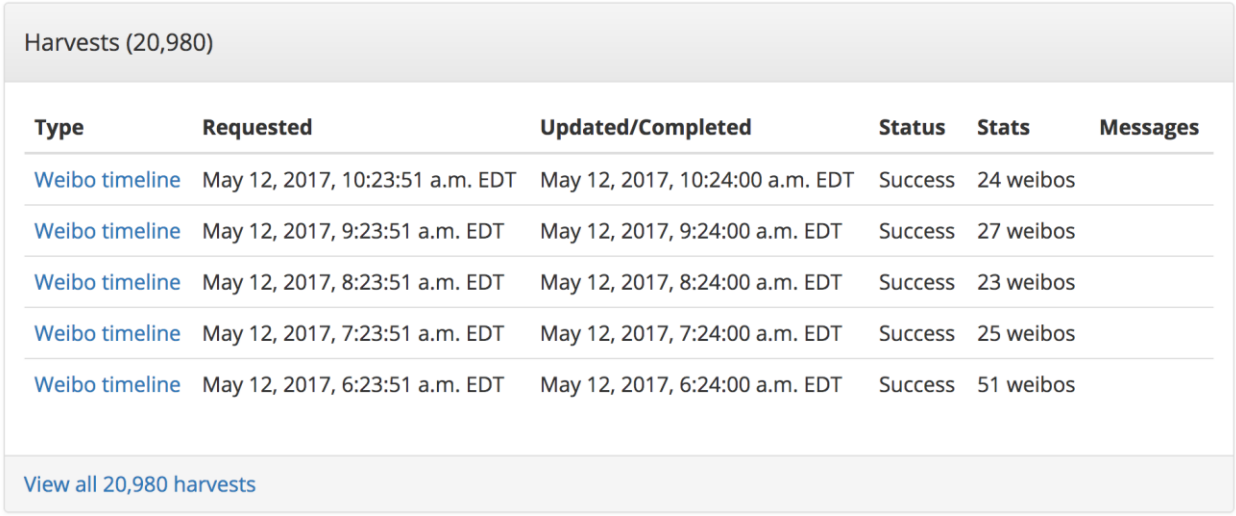



Most other types of collections require the user to provide seeds. For example, a Twitter user timeline collection requires the screen names of the accounts to collect (e.g., @SocialFeedMgr). However, since the weibos that are collected for a weibo timeline collection are determined by the friends of the user whose credentials are used, seeds are not applicable.

Lastly, the collection must be turned on.



SFM initiates harvests as determined by the schedule. In these harvests, SFM's "harvester" components request content from the social media APIs. SFM provides the user with information about the harvests that were performed.

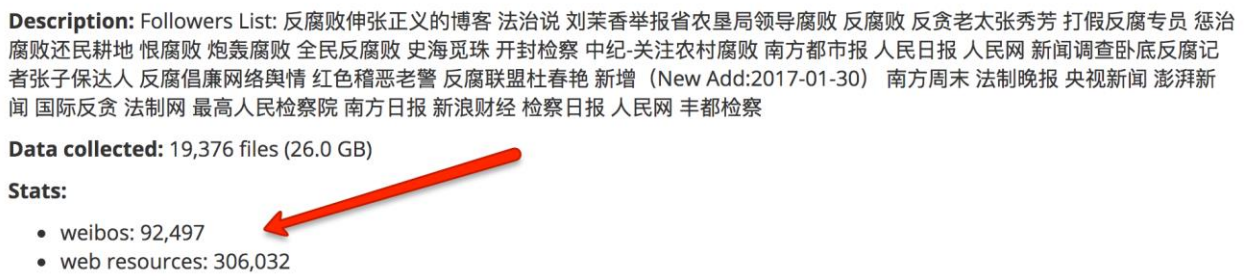


Harvests (20,980)

| Type | Requested | Updated/Completed | Status | Stats | Messages |
|---|---|---|---|---|---|
| Weibo timeline | May 12, 2017, 10:23:51 a.m. EDT | May 12, 2017, 10:24:00 a.m. EDT | Success | 24 weibos | |
| Weibo timeline | May 12, 2017, 9:23:51 a.m. EDT | May 12, 2017, 9:24:00 a.m. EDT | Success | 27 weibos | |
| Weibo timeline | May 12, 2017, 8:23:51 a.m. EDT | May 12, 2017, 8:24:00 a.m. EDT | Success | 23 weibos | |
| Weibo timeline | May 12, 2017, 7:23:51 a.m. EDT | May 12, 2017, 7:24:00 a.m. EDT | Success | 25 weibos | |
| Weibo timeline | May 12, 2017, 6:23:51 a.m. EDT | May 12, 2017, 6:24:00 a.m. EDT | Success | 51 weibos | |

View all 20,980 harvests

SFM also provides summary statistics for a collection.



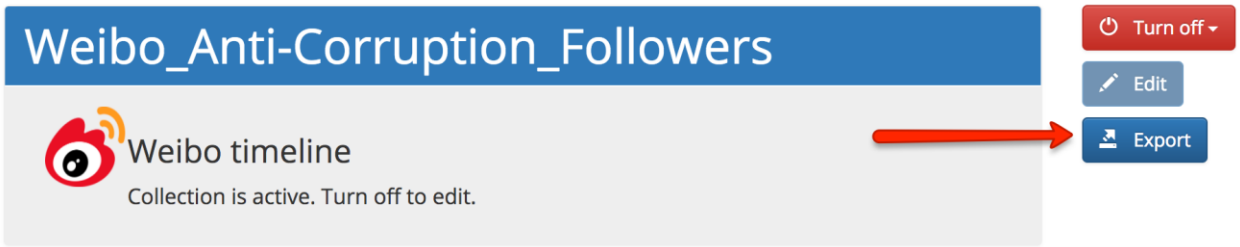
**Description:** Followers List: 反腐败伸张正义的博客 法治说 刘茉香举报省农垦局领导腐败 反腐败 反贪老太张秀芳 打假反腐专员 惩治腐败还民耕地 恨腐败 炮轰腐败 全民反腐败 史海觅珠 开封检察 中纪-关注农村腐败 南方都市报 人民日报 人民网 新闻调查卧底反腐记者张子保达人 反腐倡廉网络舆情 红色稽恶老警 反腐联盟杜春艳 新增（New Add:2017-01-30） 南方周末 法制晚报 央视新闻 澎湃新闻 国际反贪 法制网 最高人民检察院 南方日报 新浪财经 检察日报 人民网 丰都检察

**Data collected:** 19,376 files (26.0 GB)

**Stats:**

- weibos: 92,497
- web resources: 306,032

There are several mechanisms for accessing social media data. By far the most common and simplest is to export it to a spreadsheet format such as Excel or CSV. Other options include a JSON-formatted export, or access from the command line.




Key fields are extracted from each weibo to flatten it into a spreadsheet row. For weibos, this is documented in the data dictionary.[22]


**Challenges in working with the Weibo API**

Getting SFM to this point in supporting collecting social media data from Sina Weibo required navigating a number of challenges. The first challenge is the Weibo documentation, which is available in Mandarin[23] and English[24] (Sina Corp 2017). The English documentation is sorely out of date. While the Mandarin documentation is less out of date, it seems to often be inaccurate. The SFM team fortunately includes a native Chinese speaker and was able to overcome the errors in the documentation through careful experimentation.

The second challenge is the process of getting credentials, especially for those not fluent in Mandarin. While we did encounter some existing English instructions for creating an account and applying for credentials, they were out of date. To support SFM users and other Sina Weibo researchers, we wrote our own instructions (Tan 2016).

Applying for credentials to the advanced API requires some additional steps (documented in the instructions mentioned above). The software application that is requesting the

[22] http://sfm.readthedocs.io/en/latest/data_dictionary.html#weibo-dictionary
[23] http://open.weibo.com/wiki/%E9%A6%96%E9%A1%B5
[24] http://open.weibo.com/wiki/%E9%A6%96%E9%A1%B5/en

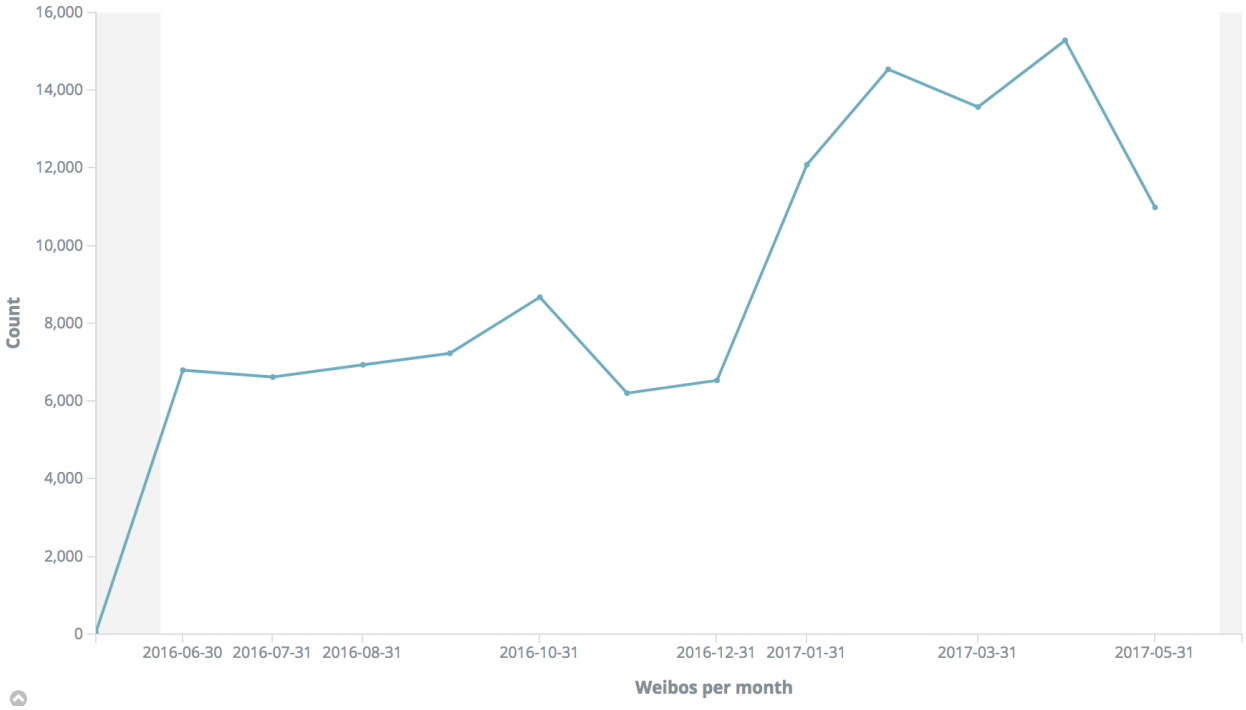
credentials must have at least 1,000 authorized users. This is a considerable hurdle; options include manually creating accounts or purchasing accounts from the grey market that exists around Sina Weibo and then performing authorization.[25] There is no easy way to automate authorization due to the anti-robot/CAPTCHA. It should be recognized that this represents a significant barrier to other institutions using the topic search API method to collect weibos.

And lastly, for security reasons some parts of the Sina Weibo website are blocked by some organizations. (These are primarily resources included in web pages, not the site itself.) In the case of GW, we could not register for credentials on the GW campus network; we had to use an external network.

**China's anti-corruption campaign collection overview**

While the social media data related to the Chinese anti-corruption campaign has not yet been closely studied, we can at least consider an overview of the collection. Since its inception on June 29, 2016 until May 10, 2017, 91,584 weibos and 23,218 tweets have been collected.

This visualization show the monthly collection rate over the past year.




Here are the top 10 topics from the past year in the dataset:

[25] In early 2017, purchasing accounts used for authorization on Taobao (https://www.taobao.com) cost around $1 per 100 accounts.

| Topic | Count |
|---|---|
| 纪检监察 | 545 |
| 反腐倡廉 | 457 |
| 2017两会 | 417 |
| 2017看两会 | 402 |
| 北电学生举报教授 | 401 |
| 央视新闻微直播 | 380 |
| 关注里约奥运 | 373 |
| 微博看两会 | 294 |
| 北电学生实名举报女教授 | 264 |
| 一起看奥运 | 245 |

When we take a deep look at the top 10 topics, 纪检监察, 反腐倡廉, 北电学生举报教授 and 北电学生实名举报女教授 are closely related to the anti-corruption campaign. The other three topics, 2017两会2017看两会 and 微博看两会, are all related to 两会[26] which refer to two annual political plenary sessions, National People's Congress and National Committee of the Chinese People's Political Consultative Conference. In both conferences, the anti-corruption campaign is one of the most important topics. 央视新闻微直播 is the official Sina Weibo account of CCTV's center for news which covers breaking news and featured news. News related to the anti-corruption campaign is probably one of its foci.

And a word cloud of the top 30 topics:

[26] https://en.wikipedia.org/wiki/Lianghui


Over the past year, the most prolific weibo authors have been:

| Screen name | Count |
|---|---|
| 人民网 | 21354 |
| 人民日报 | 11925 |
| 刘茉香举报省农垦局领导腐败 | 7817 |
| 打腐败促正义 | 7805 |
| 新浪财经 | 6804 |
| 新闻调查反腐卧底记者张子保 | 6404 |

| | |
|---|---|
| 法制网 | 5560 |
| 央视新闻 | 5437 |
| 法制晚报 | 4420 |
| 新闻调查卧底反腐记者张子保 | 4029 |

Among the 10 accounts, 人民网人民日报央视新闻 represent official government media, which reports comprehensive news that covers the anti-corruption campaign. 新浪财经 is Sino's official account with a focus on finance and economics, and also covers news on corruptions. 法制网 and 法治晚报, as the most prestigious media on law, are major media channels revealing corruption cases. 刘茉香举报省农垦局领导腐败, 打腐败促正义 and 新闻调查反腐卧底记者张子保 are private accounts that focus exclusively on anti-corruption.

The top 10 links includes in weibos were:

| Link | Count |
|---|---|
| https://m.weibo.cn/status/3987396334993894?wm=3333_1001 | 573 |
| http://rp11047911.faisco.cn/ | 491 |
| https://m.weibo.cn/status/3983754928012045?luicode=20000061&lfid=3987396334993894&featurecode=20000180&wm=2468_1006 | 388 |
| https://m.weibo.cn/status/3987396334993894?luicode=20000061&lfid=3983754928012045&featurecode=20000180&wm=3333_1001 | 383 |
| http://weibo.com/ttarticle/p/show?id=230940398009153213957 4 | 317 |
| https://m.weibo.cn/status/4044157926040825?luicode=20000061&lfid=3987396334993894&featurecode=20000180&wm=2468_1007 | 270 |

| | |
|---|---|
| https://m.weibo.cn/status/3950106925327228?luicode=20000061&lfid=4044157926040825&featurecode=20000180&wm=3333_1001 | 191 |
| https://m.weibo.cn/status/3963019648757989?luicode=20000061&lfid=3950106925327228&featurecode=20000180&wm=3333_1001 | 174 |
| https://m.weibo.cn/status/3954573536554737?luicode=20000061&lfid=3963019648757989&featurecode=20000180 | 166 |
| https://m.weibo.cn/status/4022801294191995?luicode=20000061&lfid=3954573536554737&featurecode=20000180&wm=3333_1001 | 154 |

Note that some of these links may require a Sina Weibo account; others may no longer be available.

And lastly, these are the top 10 users mentioned over the past year:

| User | Count |
|---|---|
| 明月清风74545 | 1584 |
| 打腐败促正义 | 1347 |
| 中国政府网 | 986 |
| 感恩 | 969 |
| 素素abc | 676 |
| 少佳小屋2 | 520 |
| 中国政府网 | 518 |
| 最高人民检察院 | 496 |

| | |
|---|---|
| 人民网 | 465 |
| 谢流石 | 431 |

Four of the top five users are private accounts (including 明月清风74545打腐败促正义感恩素素abc) that focus is on anti-corruption. They follow each other, and repost each other's weibos. 中国政府网最高人民检察院 and 人民网 are official accounts of the Chinese government.

Though only a cursory look at the anti-corruption campaign dataset, it indicates that collecting has largely been on-topic. Further, it hopefully demonstrates the research potential of the dataset.

**Conclusion**

When the social media data is combined with the blogs captured by our colleagues, we contend that this represents a unique and valuable collection for current and future scholars to study non-official voices in China's anti-corruption campaign. Though there are significant limitations to the Sina Weibo API and challenges in using the API, we believe that our work in adding Sina Weibo support to SFM and building the China's anti-corruption campaign collection proves that these can be overcome to perform meaningful collection building. Further, we are hopeful that through the demonstration of this collection and by lowering the technical barriers to collecting from Sina Weibo, it will be possible for others in the community to draw from the millions of Sina Weibo users to create other compelling research collections.

**Acknowledgements**

The authors are grateful for the support of the Mellon Foundation-CEAL Innovation Grant Program and the opportunity to collaborate with Yunshan Ye and Ding Ye. In addition, we appreciate the contribution of current and former colleagues on the SFM team: Laura Wrubel, Aditya Dharne, Christie Peterson, Dan Chudnov, Rachel Trent, Soomin Park, and Yonah Bromberg Gaber.

**References**

Freier, Anne. 2015. "Sina Weibo Revenue and Statistics." *Business of Apps*. December 7. http://www.businessofapps.com/sina-weibo-revenue-and-statistics/.

GW Libraries. 2015. "Social Feed Manager (SFM) Documentation." http://sfm.readthedocs.org/en/latest/.

Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. 2016. "API-Based Social Media Collecting as a Form of Web Archiving." *International Journal on Digital Libraries*, December. Springer Berlin Heidelberg, 1–18. doi:10.1007/s00799-016-0201-7.

Sina Corp. 2017. "微博API." Accessed June 14. http://open.weibo.com/wiki/%E9%A6%96%E9%A1%B5.

Tan, Yecheng. 2016. "Weibo API Guide." *Social Feed Manager*. April 26. https://gwu-libraries.github.io/sfm-ui/posts/2016-04-26-weibo-api-guide.

Thomson, Sara Day. 2016. "Preserving Social Media." Edited by Neil Beagrie. Digital Preservation Coalition. doi:10.7207/twr16-01.