sfm Documentation

Release 1.2.0

The George Washington University Libraries

November 02, 2016

Social Feed Manager is open source software for libraries, archives, cultural heritage institutions and research organizations. It empowers those communities' researchers, faculty, students, and archivists to define and create collections of data from social media platforms. Social Feed Manager will harvest from Twitter, Tumblr, Flickr, and Sina Weibo and is extensible for other platforms. In addition to collecting data from those platforms' APIs, it will collect linked web pages and media.

This site provides documentation for installation and usage of SFM. See the Social Feed Manager project site for full information about the project's objectives, roadmap, and updates.

CHAPTER 1

Quick Start Guide

This quick start guide describes how you can start using Social Feed Manager to select, harvest, explore, export, process and analyze social media data. This covers just the basics of using the software; technical information about installing and administering SFM can be found in the technical-documentation.

1.1 Prerequisites

1.1.1 SFM in operation

This quick start guide assumes SFM is already set up and running. For details about installing and administering SFM, see technical-documentation.

1.1.2 An SFM account

You can sign up for an account by clicking the *Sign Up* link from within SFM.

If you'd like to set up shared collecting at your institution, you'll need to have your systems administrator set up groups in SFM.

1.1.3 API credentials

You will need API credentials for each of the social media platforms from which you want to collect. This is more than the Twitter/Flickr/Weibo account that you may already have. To get API credentials:

- Request credentials from the social media platform and enter them into Credentials section. The API Credentials page provides instructions for each platform.
- For some social media platforms, your administrator may have enabled an option that will allow you to connect your account without leaving SFM. With your permission, SFM will get credentials on your behalf. Click *Credentials* and then *Connect [Twitter, Tumblr, or Weibo] Account*.
- If you are part of a group, you'll be able to use the credentials already provided by another member of the group.

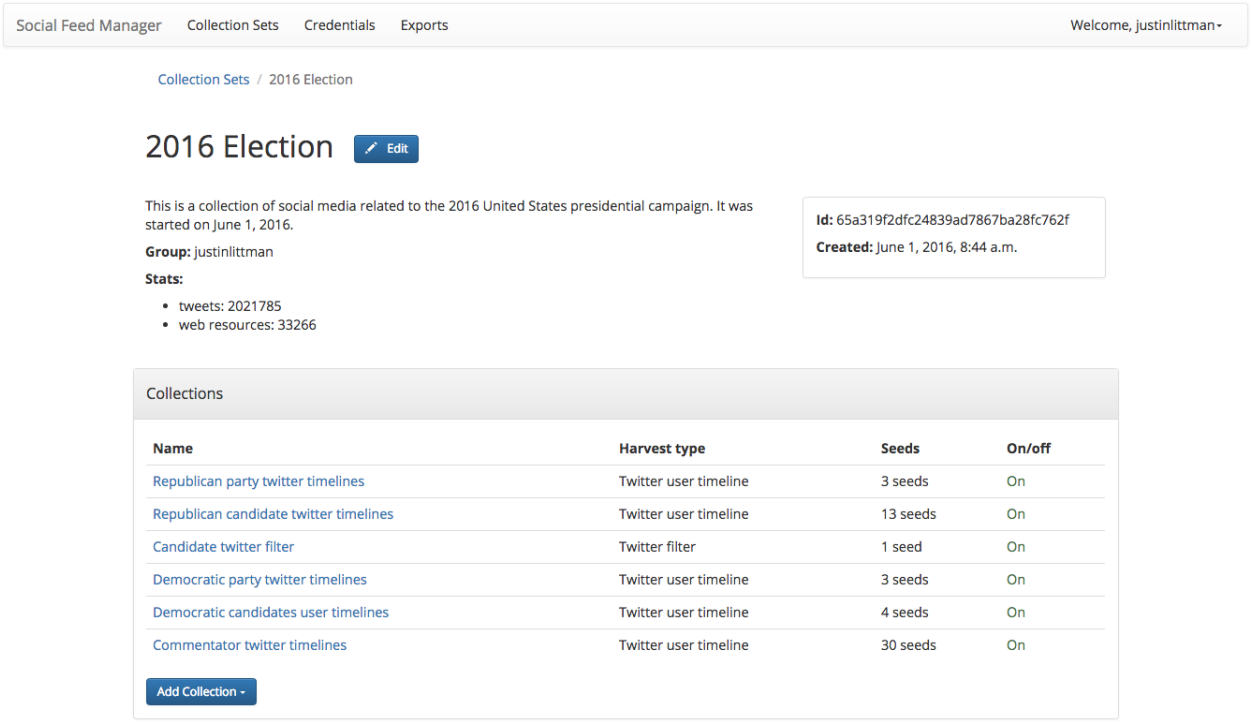
1.2 Setting up collections

Hopefully you've considered what you want to use SFM to collect: which social media accounts, which queries/hashtags/searches/etc., and on which platform(s). You may also have learned a bit about the social media

platforms' APIs and best practices for collecting from social media APIs. Now you'd like to set up your collections in SFM.

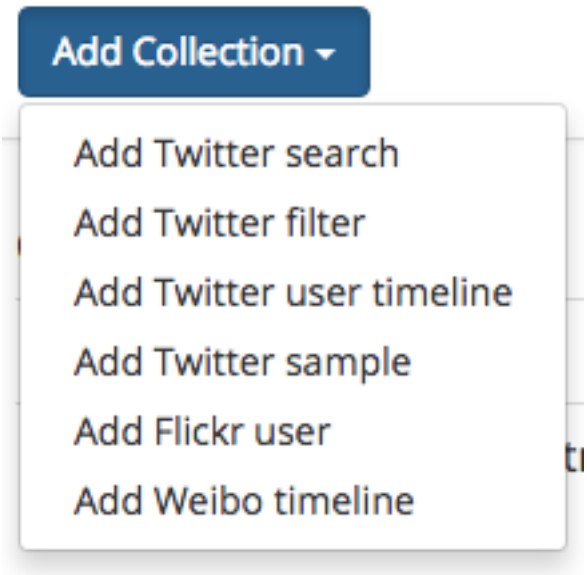
1.2.1 Create a collection set

At the top of the page, go to *Collection Sets* and click the *Add Collection Set* button. A collection set is just a group of collections around a particular topic or theme. For example, you might set up a "2016 U.S. Elections" collection set.



1.2.2 Create a collection

On the collection set detail page, under *Collections* click the *Add Collection* button and select a type.



Collection types differ based on the social media platform and the part of the API from which the social media is to be collected. For more information, see Collection types.

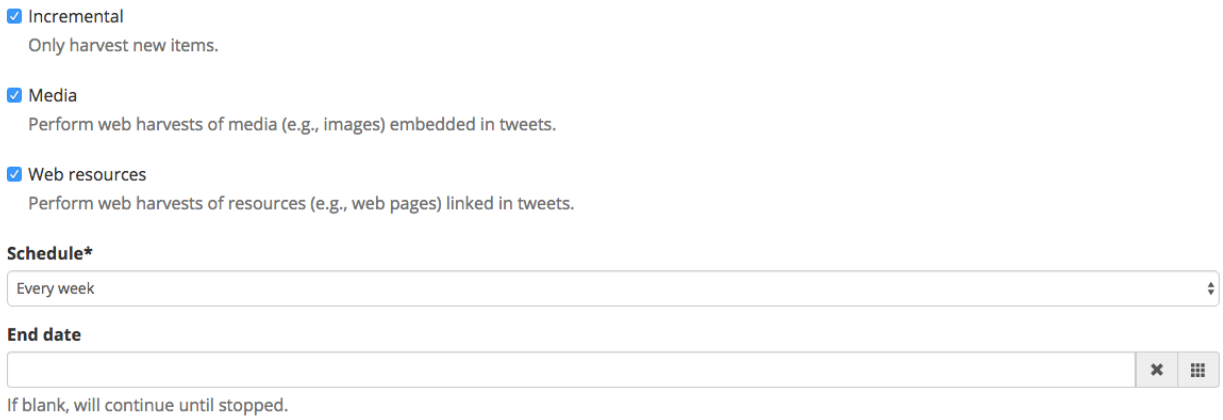
The collection types supported by SFM include:

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

SFM allows you to create multiple collections of each type within a collection set. For example, you might create a "Democratic candidate Twitter user timelines" collection and a "Republican candidate Twitter user timelines" collection. Collections are one way of organizing harvested content.

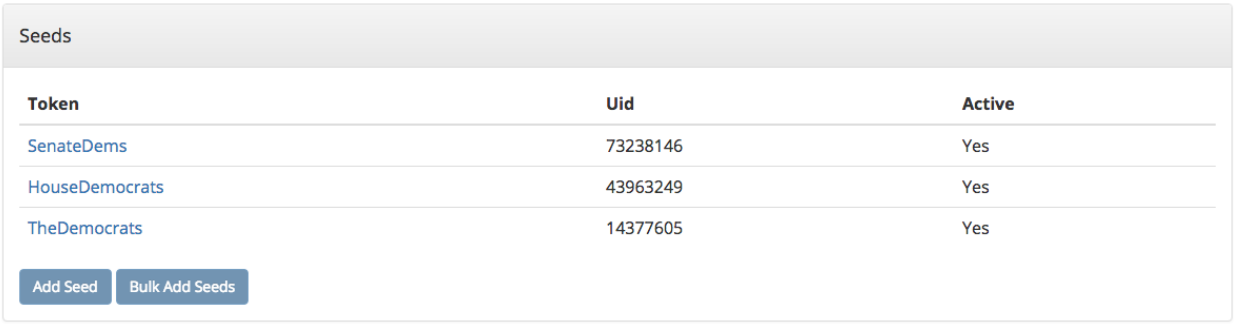
Each collection's harvest type has specific options, which may include:

- Schedule of how often to collect (e.g. daily, monthly). Streaming harvest types such as Twitter filter don't have a schedule – they're either on or off.
- Whether to perform web harvests of images, videos, or web pages embedded or linked from the posts.
- Whether to harvest incrementally. For example, each time a Twitter user timeline harvest runs, it can either collect only new items since the last harvest, or it can try to re-collect each entire timeline.



1.2.3 Add seeds

Some harvest types require seeds, which are the specific targets for collection.



As shown in the chart below, what a seed is and the number of seeds varies by harvest type. Note that some harvest types don't have any seeds.

Harvest type	Seed	How many?
Twitter search	Search query	1 or more
Twitter filter	Track/Follow/Locations	1 or more
Twitter user timeline	Twitter Account Name or ID	1 or more
Twitter sample	None	None
Flickr user	Flickr Account Name or ID	1 or more
Weibo timeline	None	None

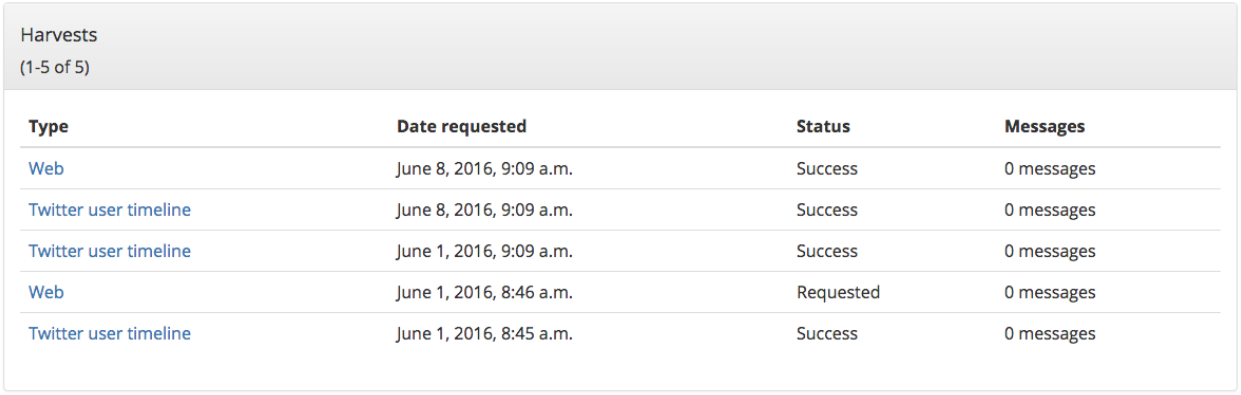
1.3 Start harvesting

Each collection's detail page has a *Turn On* button.

Once you turn on the collection, harvesting will proceed in the background according to the collection's schedule. It will stop when it hits the end date or you turn it off.

The collection's detail page will also show a message noting when the next harvest is scheduled for.

As harvesting progresses, SFM will list the results of harvests on the collection's detail page.



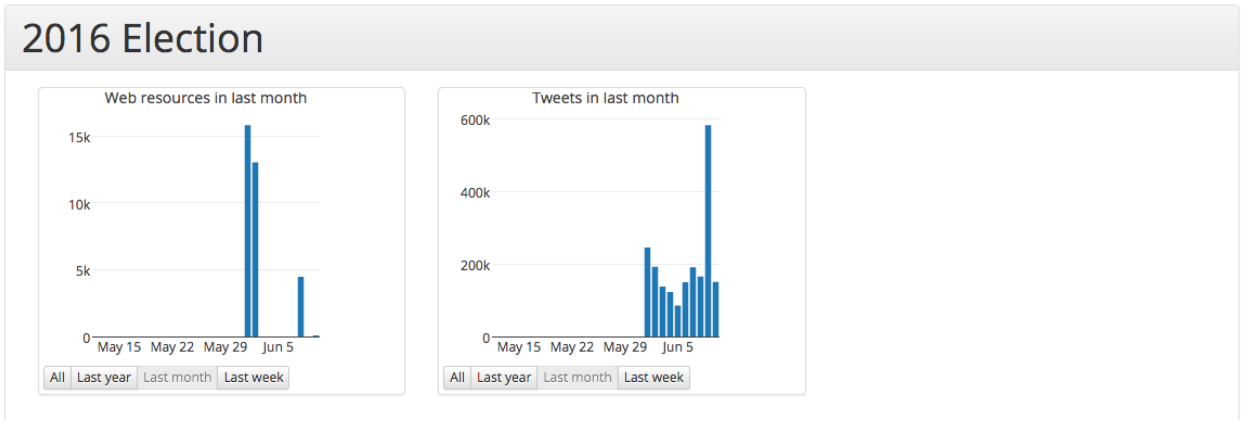
1.4 During harvesting

Within SFM, harvesting is performed by (you guessed it) harvesters. Harvesters make calls to the social media platforms' APIs and records the social media data in WARC files. (WARC is a standard file format used for web archiving.)

Depending on the collection options you selected, SFM may also extract URLs from the posts; these URLs link to web resources such as images, web pages, etc. SFM passes the URLs to the web harvester, which will collect these web resources (similar to more traditional web archiving).

To monitor harvesting:

- View details on each harvest in the Harvests section of the collection detail page.
- Check the visualizations of the number of items harvested for each collection on the home page. (Click *Social Feed Manager* in the top left of the page).



If you want to make changes to the collection's options and/or its seeds after harvesting is started, turn off the collection and then click the *Edit* button.

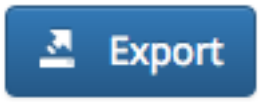
You'll be able to turn it back on and resume collecting afterwards.

1.5 Exploring, exporting, processing and analyzing your social media data

SFM provides several mechanisms for exporting collected social media data or feeding the social media data into your own processing pipelines. It also provides some basic tools for exploring and analyzing the collected content within the SFM environment.

1.5.1 Exports

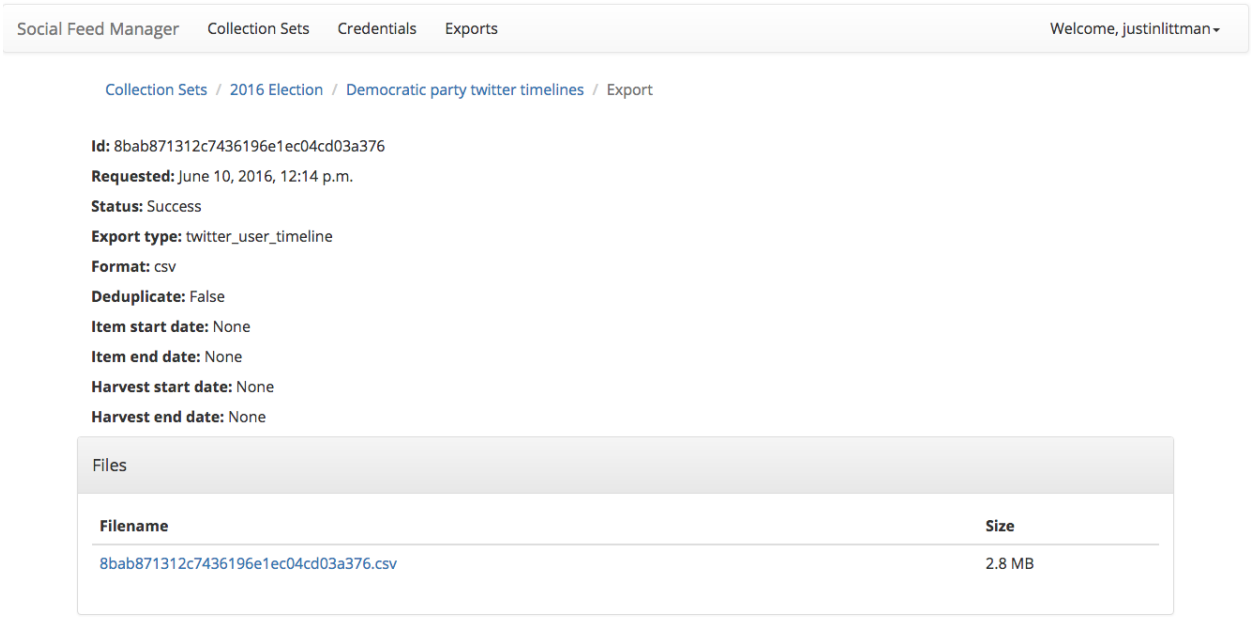
To export collected social media data, click the *Export* button on the collection detail page. Exports are available in a number of formats, including Excel, CSV, and JSON.

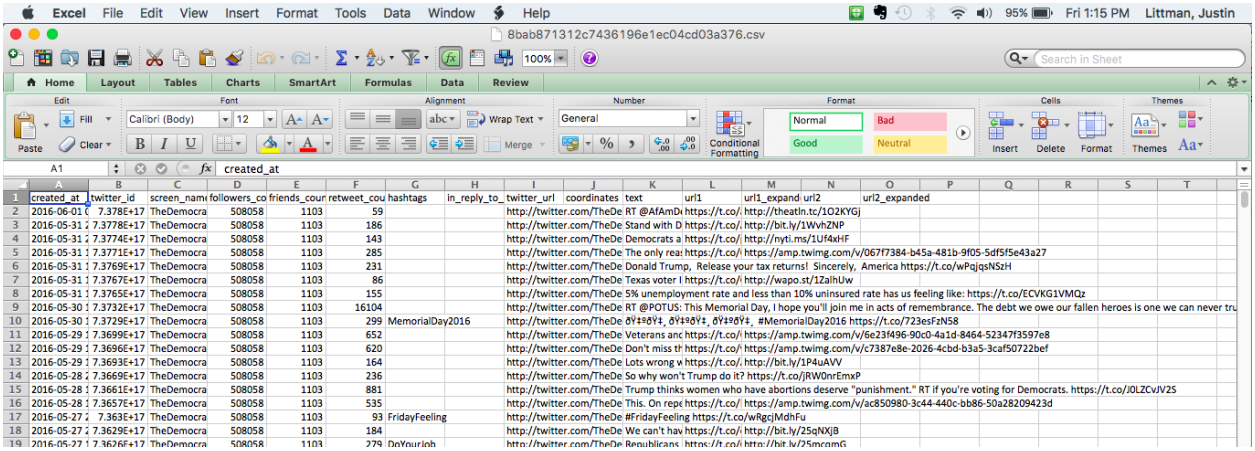


The "Full JSON" format provides the posts (e.g. tweets) in their original form, whereas the other export formats provide a subset of the metadata for each social media item. For example, for a tweet, the CSV export includes the tweet's "coordinates" value but not the "geo" value.

Dehydration (exporting a list of just the IDs of social media items) is supported for certain data-sharing purposes.

Exports are run in the background, and larger exports may take a significant amount of time. You will receive an email when it is completed or you can monitor the status on the Exports page, where you can vew details about the export. This is also where you will find a link to download the export file once it becomes available.





1.5.2 Processing

If you've set up a processing container, or if you've installed SFM tools locally, then you have access to the collected social media data from the command line. You can then feed the data into your own processing pipeline and use your own tools.

More on this topic can be found in the Processing section.

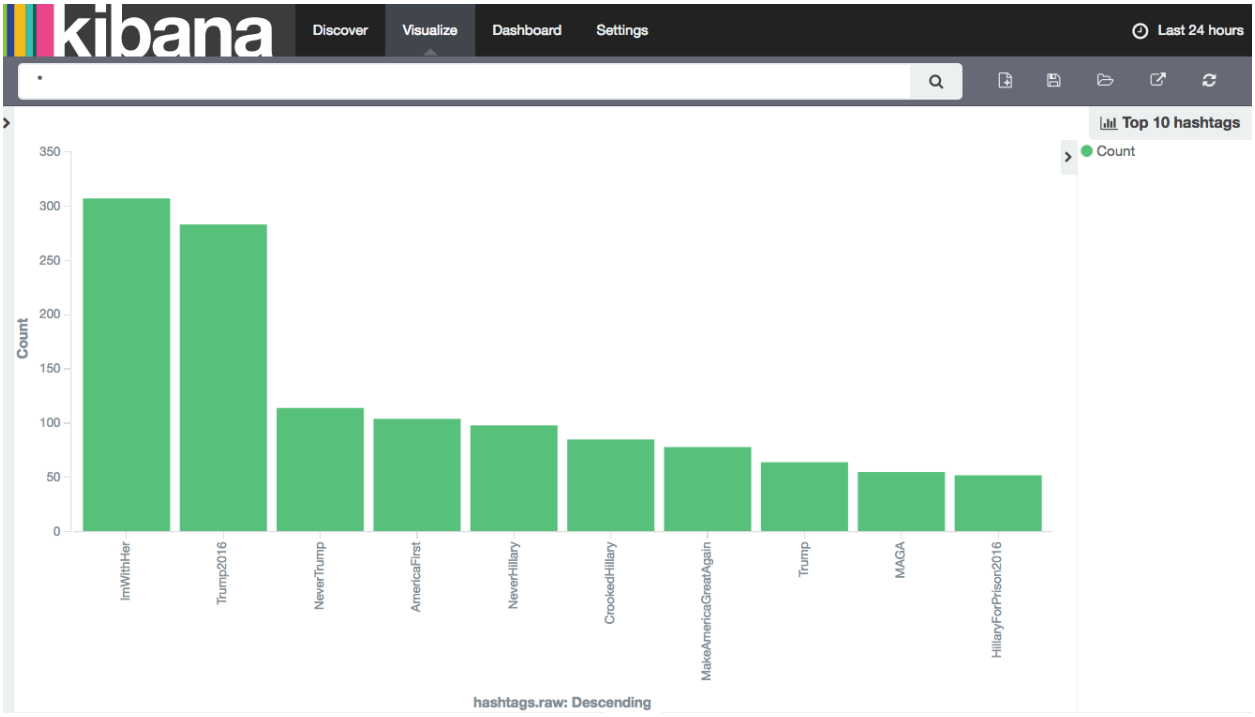
1.5.3 Exploration and analysis

While SFM does not provide a comprehensive toolset for exploring and analyzing the collected social media data, it provides some basic exploration and analysis tools and allows you to export social media data for use with your own tools.

Tools provided by SFM are:

- ELK (Elasticsearch, Logstash, Kibana)

The ELK stack is a general-purpose framework for exploring data. It provides support for loading, querying, analysis, and visualization. SFM provides an instance of ELK that has been customized for exploring social media data, in particular, Twitter and Weibo data.



ELK may be particularly useful for monitoring and adjusting the targets of ongoing social media collections. For example, it can be used to discover additional relevant Twitter hashtags or user accounts to collect, based on what has been collected so far.

ELK requires some additional setup. More on this topic can be found in the Exploring social media data with ELK section.

- Processing container

A processing container allows you to have access to the collected social media content from the command line. The processing container has been provisioned with a handful of analysis tools such as Twarc utils.

The following shows piping some tweets into a wordcloud generator from within a processing container:

```
# find_warcs.py 4f4d1 | xargs twitter_rest_warc_iter.py | python /opt/twarc/utils/wordcloud.py
```

More on this topic can be found in the Processing section.

1.6 Access and display

SFM does not currently provide a web interface to the collected social media content. However, this should be possible, and we welcome your ideas and contributions.

CHAPTER 2

API Credentials

Accessing the APIs of social media platforms requires credentials for authentication (also knows as API keys). Social Feed Manager supports managing those credentials.

Most API credentials have two parts: an application credential and a user credential. (Flickr is the exception – only an application credential is necessary.)

It is important to understand how credentials/authentication affect what API methods can be invoked and rate limits. For more information, consult the documentation for each social media platform's API.

2.1 Managing credentials

SFM supports two approaches to managing credentials: adding credentials and connecting credentials. Both of these options are available from the Credentials page.

2.1.1 Adding credentials

For this approach, a user gets the application and/or user credential from the social media platform and provide them to SFM by completing a form. More information on getting credentials is below.

2.1.2 Connecting credentials

For this approach, SFM is configured with the application credentials for the social media platform. The user credentials are obtained by the user being redirected to the social media website to give permission to SFM to access her account.

SFM is configured with the application credentials in the `docker-compose.yml`. If additional management is necessary, it can be performed using the Social Accounts section of the Admin interface.

This is the easiest approach for users. Configuring application credentials is encouraged.

2.2 Platform specifics

2.2.1 Twitter

Twitter credentials can be obtained from https://apps.twitter.com/. It is recommended to change the application permissions to read-only. You *must* provide a callback URL, but the URL you provide doesn't matter.

2.2.2 Weibo

For instructions on obtaining Weibo credentials, see this guide.

To use the connecting credentials approach for Weibo, the redirect URL must match the application's actual URL and use port 80.

2.2.3 Flickr

Flickr credentials can be obtained from https://www.flickr.com/services/api/keys/.

Flickr does not require user credentials.

2.2.4 Tumblr

Tumblr credentials can be obtained from https://www.tumblr.com/oauth/apps.

CHAPTER 3

Collection types

Each collection type connects to one of a social media platform's APIs, or methods for retrieving data. Understanding what each collection type provides is important to ensure you collect what you need and are aware of any limitations. Reading the social media platform's documentation provides further important details.

3.1 Twitter search

Queries the Twitter Search API to retrieve public tweets from a sampling of tweets from the most recent 7-9 days. This is not a comprehensive search of all tweets.

To formulate a search query, use the Twitter Advanced Search query builder. Pay careful attention to the query syntax described in Twitter's documentation. This query is the seed for the collection; Twitter search collections only have one seed.

Due to Twitter's rate limits and the amount of data available from the Search API, broad Twitter searches may take a long time to complete (up to multiple days). In choosing a schedule, make sure that there is enough time between searches. In some cases, you may only want to run the search once and then turn off the collection.

If the incremental option is selected, only new tweets (i.e., tweets that have not yet been harvested in this collection) will be harvested. In general, you will want to select the incremental option.

See the *Collecting Web resources* guidance below for deciding whether to collection media or web resources.

3.2 Twitter filter

Collects public tweets from the Twitter filter streaming API, matching keywords, locations, or users. The tweets are from the current time going forward. Tweets from the past are not available with this collection type. (Create a Twitter search collection with the same terms to collect tweets from the recent past).

When creating a Twitter filter pay careful attention to query syntax described in Twitter's documentation. The filter query is the seed for the collection; Twitter filter collections only have one seed.

There are limits on how many tweets Twitter will supply, so filters on high-volume terms/hashtags will not return all tweets available. Thus, you will want to strategize about how broad/narrow to construct your filter. Twitter only allows you to run one filter at a time with a set of Twitter API credentials; SFM enforces this for you.

SFM captures the filter stream in 30 minute chunks and then momentarily stops. Between rate limiting and this momentary stop, you should never assume that you are getting every tweet.

Unlike other collection types, Twitter filter collections are either turned on or off; they do not operate according to a schedule.

See the *Collecting Web resources* guidance below for deciding whether to collection media or web resources.

3.3 Twitter user timeline

Collects tweets by a particular Twitter user account using Twitter's user_timeline API. Twitter provides up to the most recent 3,200 tweets by that account, provided a limited ability to collect tweets from the past.

Each Twitter user timeline collection can have multiple seeds, where each seed is a user timeline. To identify a user timeline, you can provide a screen name or Twitter user ID (UID). If you provide one, the other will be looked up and displayed in SFM UI. The Twitter user ID is a number and does not change; a user may change her screen name. The user ID will be used for retrieving the user timeline.

While large number of user timeline seeds are supported in a collection, they may take a long time to collect due to Twitter's rate limits.

If the incremental option is selected, only new tweets (i.e., tweets that have not yet been harvested for that user timeline) will be harvested, meaning you will not collect duplicate tweets. If the incremental option is not selected, you will collect the most 3,200 recent tweets, meaning you will get duplicates across harvests. However, you may be able to examine differences across time in a user's timeline, e.g., deleted tweets, or track changes in follower or retweet counts.

In choosing a schedule, you may want to consider how prolific a tweeter the Twitter user is. In general, the more frequent the tweeter, the more frequent you'll want to schedule harvests.

See the *Collecting Web resources* guidance below for deciding whether to collection media or web resources.

3.4 Twitter sample

Collects tweets from the Twitter sample stream. The Twitter sample stream returns approximately 0.5-1% of public tweets, which is approximately 3GB a day (compressed).

See the *Collecting Web resources* guidance below for deciding whether to collection media or web resources.

3.5 Flickr user

Collects metadata about public photos by a specific Flickr user.

Each Flickr user collection can have multiple seeds, where each seed is a Flickr user. To identify a user, you can provide a either a username or an NSID. If you provide one, the other will be looked up and displayed in the SFM UI. The NSID is a unique identifier and does not change; a user may change her username.

For each user, the user's information will be collected using Flickr's people.getInfo API and the list of her public photos will be retrieved from people.getPublicPhotos. Information on each photo will be collected with photos.getInfo.

Depending on the image sizes you select, the actual photo files will be collected as well.

If the incremental option is selected, only new photos will be collected.

3.6 Weibo timeline

Collects Weibos by the user and friends of the user whose credentials are provided using the Weibo friends_timeline API.

Note that because collection is determined by the user whose credentials are provided, there are no seeds for a Weibo timeline collection. To change what is being collected, change the user's friends from the Weibo website or app.

See the *Collecting Web resources* guidance below for deciding whether to collect image or web resources.

3.7 Tumblr blog posts

Collects blog posts by a specified Tumblr blog.

Each Tumblr blog post collection can have multiple seeds, where each seed is a blog. The blog can be specified with or without the .tumblr.com extension.

If the incremental option is selected, only new blog posts will be collected.

See the *Collecting Web resources* guidance below for deciding whether to collect image or web resources.

3.8 Collecting Web resources

Each collection type allows you to select an option to collect web resources such as images, web pages, etc. that are included in the social media post. When a social media post includes a URL, SFM will harvest the web page at that URL. It will harvest only that web page, not any pages linked from that page.

Be very deliberate in collecting web resources. Performing a web harvest both takes longer and requires significantly more storage than collecting the original social media post.

CHAPTER 4

Processing

Your social media data can be used in a processing/analysis pipeline. SFM provides several tools and approaches to support this.

4.1 Tools

4.1.1 Warc iterators

A warc iterator tool provides an iterator to the social media data contained in WARC files. When used from the commandline, it writes out the social items one at a time to standard out. (Think of this as `cat`-ing a line-oriented JSON file. It is also equivalent to the output of Twarc.)

Each social media type has a separate warc iterator tool. For example, `twitter_rest_warc_iter.py` extracts tweets recorded from the Twitter REST API. For example:

```
root@0ac9caaf7e72:/sfm-data# twitter_rest_warc_iter.py
usage: twitter_rest_warc_iter.py [-h] [--pretty] [--dedupe]
                                 [--print-item-type]
                                 filepaths [filepaths ...]
```

Here is a list of the warc iterators:

- `twitter_rest_warc_iter.py`: Tweets recorded from Twitter REST API.
- `twitter_stream_warc_iter.py`: Tweets recorded from Twitter Streaming API.
- `flickr_photo_warc_iter.py`: Flickr photos
- `weibo_warc_iter.py`: Weibos
- `tumblr_warc_iter.py`: Tumblr posts

Warc iterator tools can also be used as a library.

4.1.2 Find Warcs

`find_warcs.py` helps put together a list of WARC files to be processed by other tools, e.g., warc iterator tools. (It gets the list of WARC files by querying the SFM API.)

Here is arguments it accepts:

```
root@0ac9caaf7e72:/sfm-data# find_warcs.py
usage: find_warcs.py [-h] [--include-web] [--harvest-start HARVEST_START]
                     [--harvest-end HARVEST_END] [--api-base-url API_BASE_URL]
                     [--debug [DEBUG]]
                     collection [collection ...]
```

For example, to get a list of the WARC files in a particular collection, provide some part of the collection id:

```
root@0ac9caaf7e72:/sfm-data# find_warcs.py 4f4d1
/sfm-data/collections/b06d164c632d405294d3c17584f03278/4f4d1a6677f34d539bbd8486e22de33b/2016/05/04/14
```

(In this case there is only one WARC file. If there was more than one, it would be space separated.)

The collection id can be found from the SFM UI.

Note that if you are running `find_warcs.py` from outside a Docker environment, you will need to supply `--api-base-url`.

4.2 Approaches

4.2.1 Processing in container

To bootstrap processing, a processing image is provided. A container instantiated from this image is Ubuntu 14.04 and pre-installed with the warc iterator tools, `find_warcs.py`, and some other use tools. It will also have read-only access to the data from `/sfm-data`.

The other tools are:

- jq for JSON processing.
- twarc for access to the Twarc utils.
- JWAT Tools for processing WARCs.
- warctools for processing WARCs.
- parallel for parallelizing processing.

To instantiate:

```
docker-compose run --rm processing /bin/bash
```

You will then be provided with a bash shell inside the container from which you can execute commands:

```
root@0ac9caaf7e72:/sfm-processing# find_warcs.py 4f4d1 | xargs twitter_rest_warc_iter.py | python /op
```

Setting `PROCESSOR_VOLUME` in `.env` to a host volume will link `/sfm-processing` to your local filesystem. You can place scripts in this directory to make them available inside the processing container or write output files to this directory to make them available outside the processing container.

Note that once you exit the processing container, the container will be automatically removed. However, if you have saved all of your scripts and output files to `/sfm-processing`, they will be available when you create a new processing container.

4.2.2 Processing locally

In a typical Docker configuration, the data directory will be linked into the Docker environment. This means that the data is available both inside and outside the Docker environment. Given this, processing can be performed locally

(i.e., outside of Docker).

The various tools can be installed locally:

```
GLSS-F0G5RP:tmp justinlittman$ virtualenv ENV
GLSS-F0G5RP:tmp justinlittman$ source ENV/bin/activate
(ENV)GLSS-F0G5RP:tmp justinlittman$ pip install git+https://github.com/gwu-libraries/sfm-utils.git
(ENV)GLSS-F0G5RP:tmp justinlittman$ pip install git+https://github.com/gwu-libraries/sfm-twitter-harv
(ENV)GLSS-F0G5RP:tmp justinlittman$ twitter_rest_warc_iter.py
usage: twitter_rest_warc_iter.py [-h] [--pretty] [--dedupe]
                                 [--print-item-type]
                                 filepaths [filepaths ...]
twitter_rest_warc_iter.py: error: too few arguments
```

4.3 Recipes

4.3.1 Extracting URLs

The "Extracting URLs from #PulseNightclub for seeding web archiving" blog post provides some useful guidance on extracting URLs from tweets, including unshortening and sorting/counting.

4.3.2 Exporting to line-oriented JSON files

This recipe is for exporting social media data from WARC files to line-oriented JSON files. There will be one JSON file for each WARC. This may be useful for some processing or for loading into some analytic tools.

This recipe uses parallel for parallelizing the export.

Create a list of WARC files:

```
find_warcs.py 7c37157 | tr ' ' '\n' > source.lst
```

Replace *7c37157* with the first few characters of the collection id that you want to export. The collection id is available on the colllection detail page in SFM UI.

Create a list of JSON destination files:

```
cat source.lst | xargs basename -a | sed 's/.warc.gz/.json/' > dest.lst
```

This command puts all of the JSON files in the same directory, using the filename of the WARC file with a .json file extension.

If you want to maintain the directory structure, but use a different root directory:

```
cat source.lst | sed 's/sfm-data\/collection_set/sfm-processing\/export/' | sed 's/.warc.gz/.json/'
```

Replace *sfm-processing/export* with the root directory that you want to use.

Perform the export:

```
parallel -a source.lst -a dest.lst --xapply "twitter_stream_warc_iter.py {1} > {2}"
```

Replace *twitter_stream_warc_iter.py* with the name of the warc iterator for the type of social media data that you are exporting.

You can also perform a filter on export using jq. For example, this only exports tweets in Spanish:

```
parallel -a source.lst -a dest.lst --xapply "twitter_stream_warc_iter.py {1} | jq -c 'select(.lang ==
```

And to save space, the JSON files can be gzip compressed:

```
parallel -a source.lst -a dest.lst --xapply "twitter_stream_warc_iter.py {1} | gzip > {2}"
```

You might also want to change the file extension of the destination file to ".json.gz" by adjusting the commmand use to create the list of JSON destination files. To access the tweets in a gzipped JSON file, use:

```
gzip -c <filepath>
```

4.3.3 Counting posts

wc -l can be used to count posts. To count the number of tweets in a collection:

```
find_warcs.py 7c37157 | xargs twitter_stream_warc_iter.py | wc -l
```

To count the posts from line-oriented JSON files created as described above:

```
cat dest.lst | xargs wc -l
```

wc -l gotcha: When doing a lot of counting, *wc -l* will output a partial total and then reset the count. The partial totals must be added together to get the grand total. For example:

```
[Some lines skipped ...]
    1490 ./964be41e1714492bbe8ec5793e05ec86-20160725070757217-00000-7932-62ebe35d576c-8002.json
    4514 ./5f78a79c6382476889d1ed4734d6105a-20160722202703869-00000-5110-62ebe35d576c-8002.json
   52043 ./417cf950a00d44408458c93f08f0690e-20160910032351524-00000-1775-c4aea5d70c14-8000.json
54392684 total
[Some lines skipped ...]
   34778 ./30bc1c34880d404aa3254f82dd387514-20160806132811173-00000-21585-62ebe35d576c-8000.json
   30588 ./964be41e1714492bbe8ec5793e05ec86-20160727030754726-00000-10044-62ebe35d576c-8002.json
21573971 total
```

4.3.4 Using jq to process JSON

For tips on using jq with JSON from Twitter and other sources, see:

- Getting Started Working with Twitter Data Using jq
- Reshaping JSON with jq

CHAPTER 5

Exploring social media data with ELK

The ELK (Elasticsearch, Logstash, Kibana) stack is a general-purpose framework for exploring data. It provides support for loading, querying, analysis, and visualization.

SFM provides an instance of ELK that has been customized for exploring social media data. It currently supports data from Twitter and Weibo.

One possible use for ELK is to monitor data that is being harvested to discover new seeds to select. For example, it may reveal new hashtags or users that are relevant to a collection.

Though you can use Logstash and Elasticsearch directly, in most cases you will interact exclusively with Kibana, which is the exploration interface.

5.1 Enabling ELK

ELK is not available by default; it must be enabled as described here.

You can enable one or more ELK Docker containers. Each container can be configured to be loaded with all social media data or the social media data for a single collection set.

To enable an ELK Docker container it must be added to your `docker-compose.yml` and then started by:

```
docker-compose up -d
```

An example container is provided in `example.docker-compose.yml` and `example.prod.docker-compose.yml`. These examples also show how to limit to a single collection set by providing the collection set id.

By default, Kibana is available at http://<your hostname>:5601/app/kibana. (Also, by default Elasticsearch is available on port 9200 and Logstash is available on port 5000.)

If enabling multiple ELK containers, add multiple containers to your `docker-compose.yml`. Make sure to give each a unique name and map to different ports.

5.2 Loading data

ELK will automatically be loaded as new social media data is harvested. (Note, however, that there will be some latency between the harvest and the data being available in Kibana.)

Since only new social media data is added, it is recommended that you enable the ELK Docker container before beginning harvesting.

If you would like to load social media data that was harvested before the ELK Docker container was enabled, use the `resendwarccreatedmsgs` management command:

```
usage: manage.py resendwarccreatedmsgs [-h] [--version] [-v {0,1,2,3}]
                                       [--settings SETTINGS]
                                       [--pythonpath PYTHONPATH] [--traceback]
                                       [--no-color]
                                       [--collection-set COLLECTION_SET]
                                       [--harvest-type HARVEST_TYPE] [--test]
                                       routing_key
```

The `resendwarccreatedmsgs` command resends warc_created messages which will trigger the loading of data by ELK.

To use this command, you will need to know the routing key. The routing key is `elk_loader_<container id>.warc_created`. The container id can be found with `docker ps`.

The loading can be limited by collection set (`--collection-set`) and/or (`--harvest-type`). You can get collection set ids from the collection set detail page. The available harvest types are twitter_search, twitter_filter, twitter_user_timeline, twitter_sample, and weibo_timeline.

This shows loading the data limited to a collection set:

```
docker exec docker_sfmuiapp_1 python sfm/manage.py resendwarccreatedmsgs --collection-set b438a62cbc
```

5.3 Overview of Kibana

The Kibana interface is extremely powerful. However, with that power comes complexity. The following provides an overview of some basic functions in Kibana. For some advanced usage, see the Kibana Reference or the Kibana 101: Getting Started with Visualizations video.

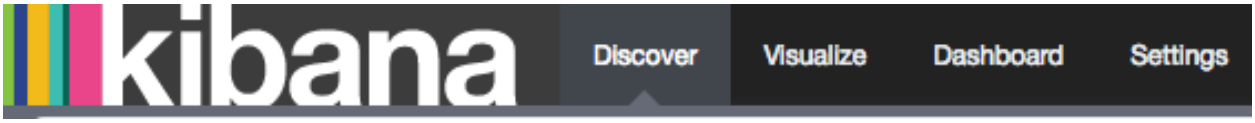
When you start Kibana, you probably won't see any results.

This is because Kibana defaults to only showing data from the last 15 minutes. Use the date picker in the upper right corner to select a more appropriate time range.

Tip: At any time, you can change the date range for your query, visualization, or dashboard using the date picker.

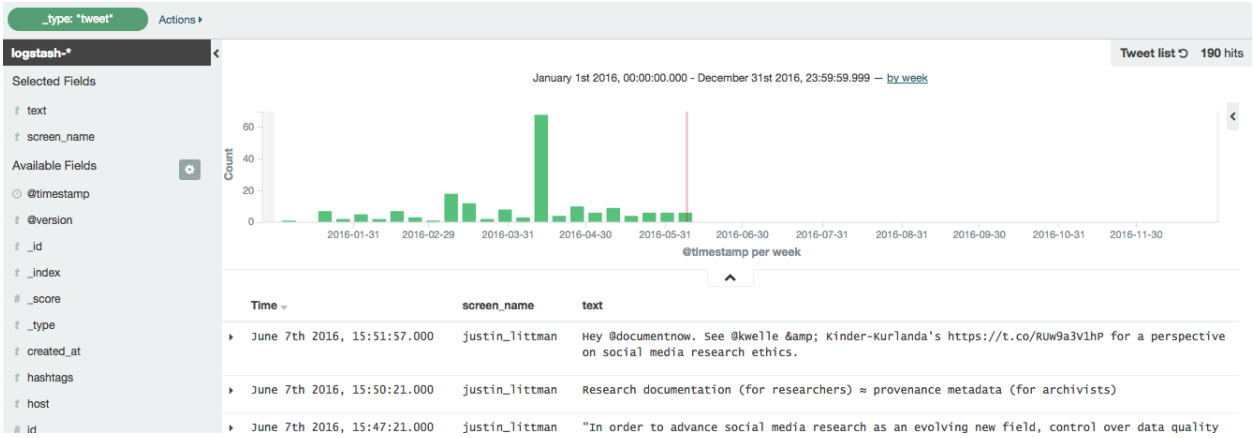
5.3.1 Discover

The Discover tab allows you to query the social media data.

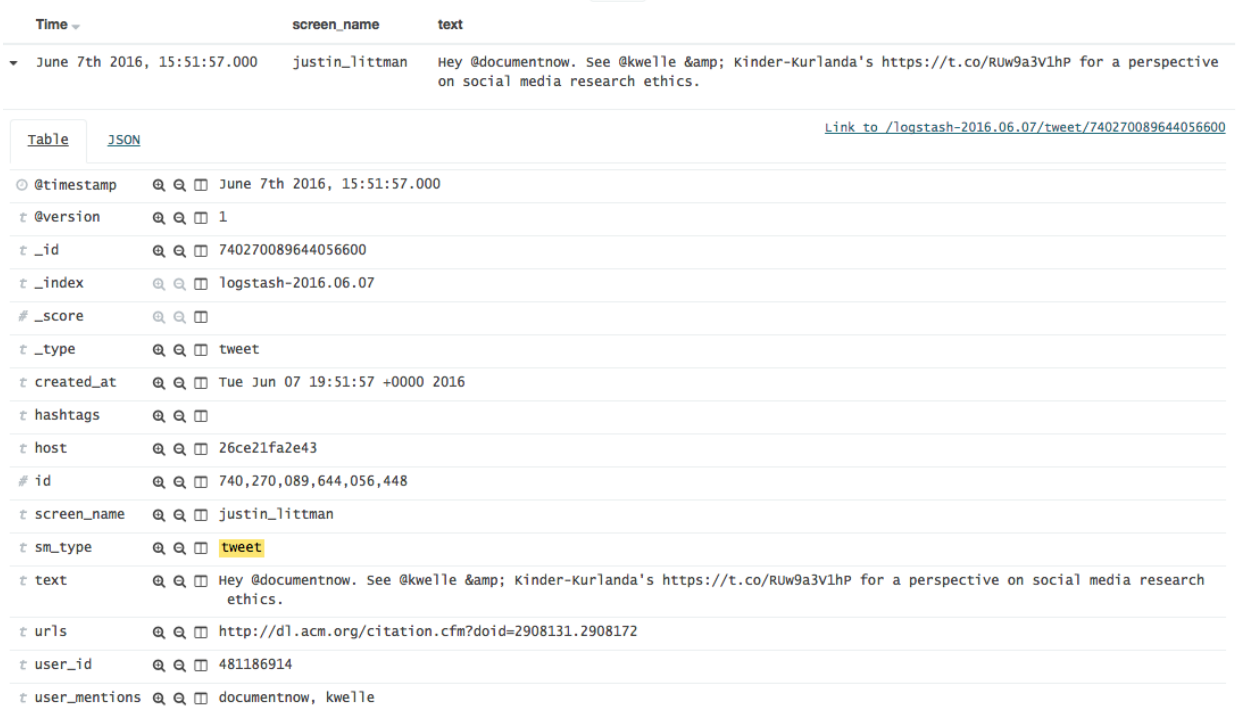


By default, all social media types are queried. By limit to a single type (e.g., tweets), click the folder icon and select the appropriate filter.

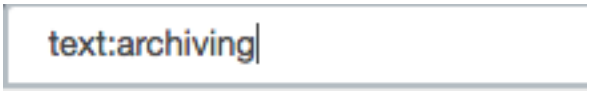
You will now only see results for that social media type.



Notice that each social media item has a number of fields.

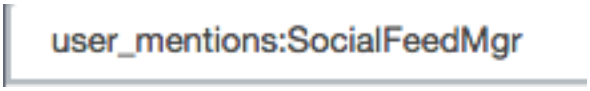


You can search against a field. For example, to find all tweets containing the term "archiving":



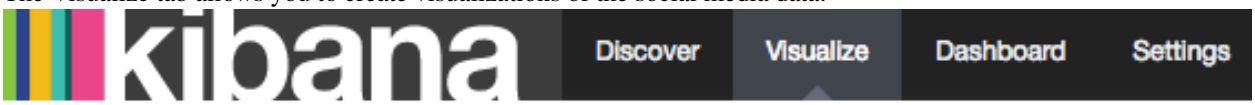
or having the hashtag #SaveTheWeb:

or mentioning @SocialFeedMgr:



5.3.2 Visualize

The Visualize tab allows you to create visualizations of the social media data.



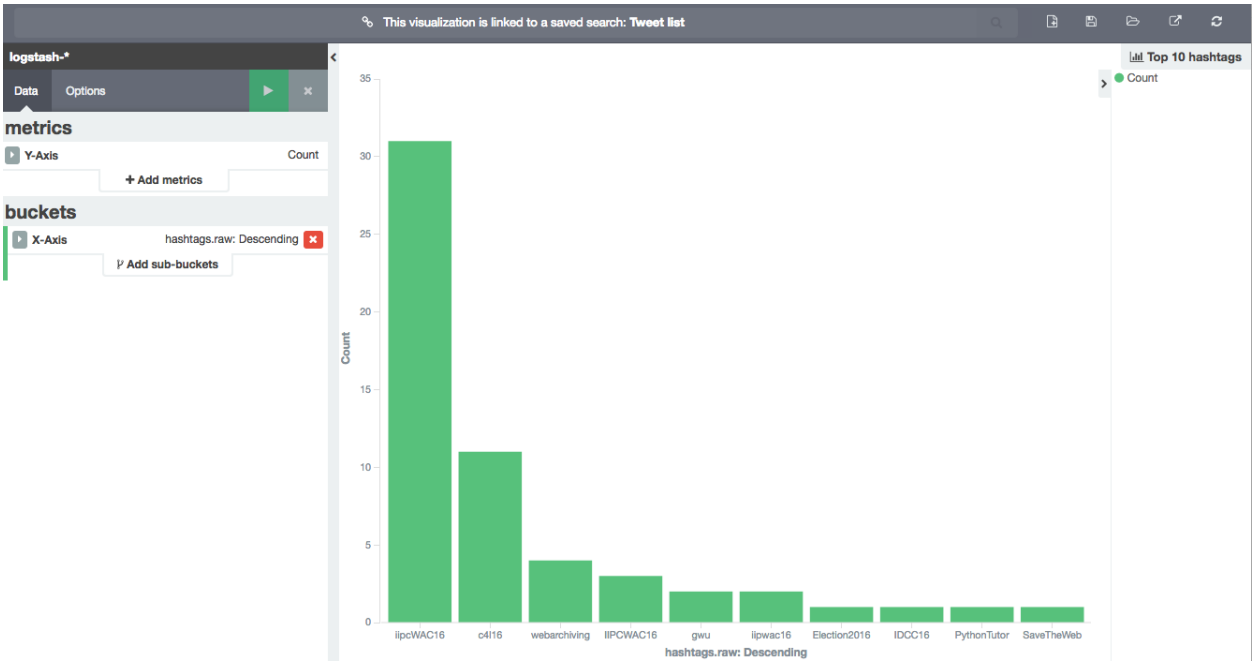
The types of visualizations that are supported include:

- Area chart
- Data table
- Line chart
- Pie chart
- Map
- Vertical bar chart

Describing how to create visualizations is beyond the scope of this overview.

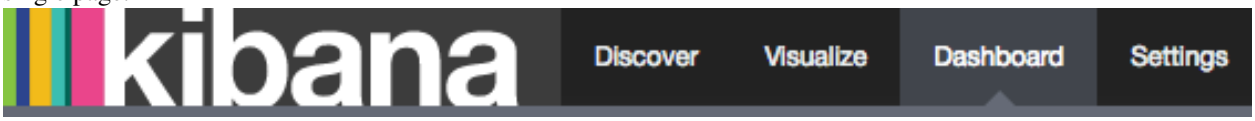
A number of visualizations have already been created for social media data. (The available visualizations are listed on the bottom of the page.)

For example, here is the Top 10 hashtags visualization:

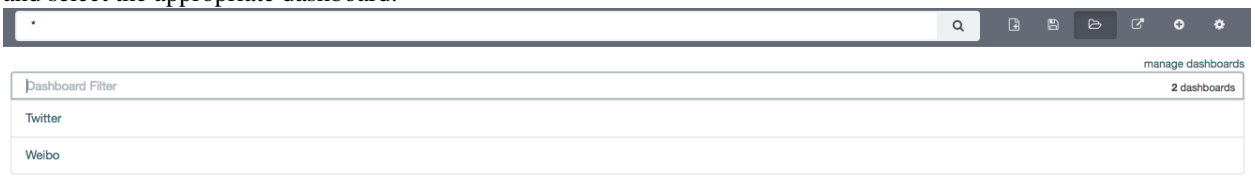


5.3.3 Dashboard

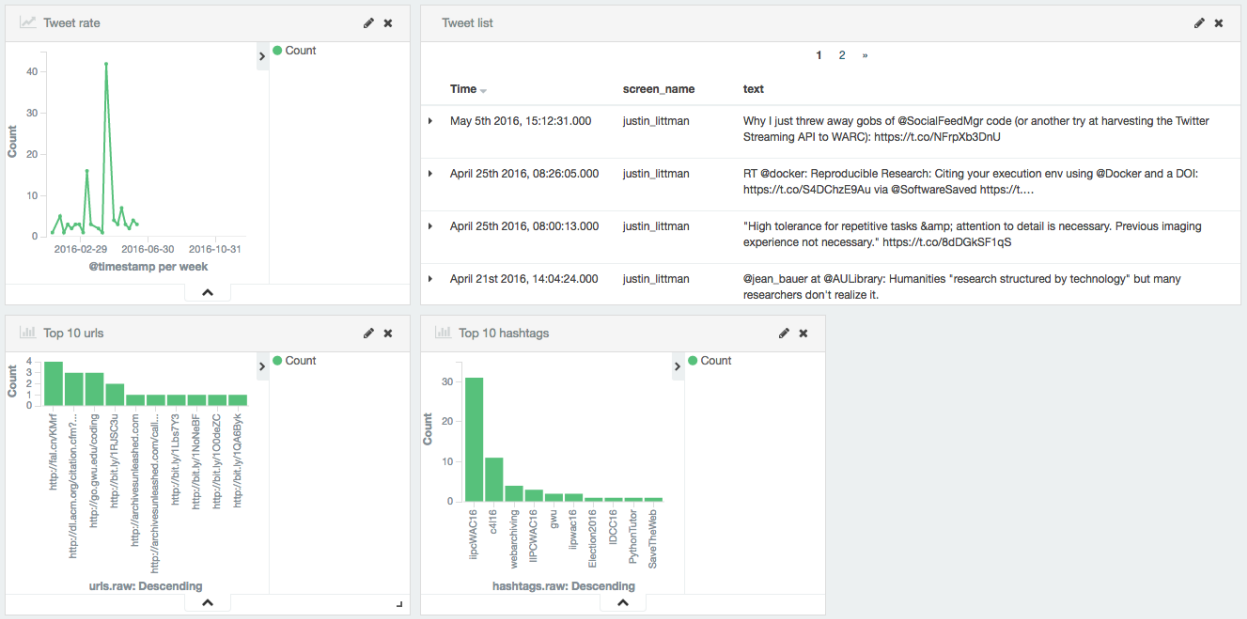
The Dashboard tab provides a summary view of data, bringing together multiple visualizations and searches on a single page.



A number of dashboards have already been created for social media data. To select a dashboard, click the folder icon and select the appropriate dashboard.



For example, here is the top of the Twitter dashboard:



5.4 Caveats

- This is experimental. We have not yet determined the level of development that will be performed in the future.
- Approaches for administering and scaling ELK have not been considered.
- No security or access restrictions have been put in place around ELK.

CHAPTER 6

Installation and configuration

6.1 Overview

The supported approach for deploying SFM is Docker containers. For more information on Docker, see Docker.

Each SFM service will provide images for the containers needed to run the service (in the form of `Dockerfile` s). These images will be published to Docker Hub. GWU created images will be part of the GWUL organization and be prefixed with *sfm-*.

sfm-docker provides the necessary `docker-compose.yml` files to compose the services into a complete instance of SFM.

The following will describe how to setup an instance of SFM that uses the latest release (and is suitable for a production deployment.) See the development documentation for other SFM configurations.

SFM *can* be deployed without Docker. The various `Dockerfile` s should provide reasonable guidance on how to accomplish this.

6.2 Local installation

Installing locally requires Docker and Docker-Compose. See *Installing Docker*.

1. Either clone the sfm-docker repository and copy the example configuration files:

```
git clone https://github.com/gwu-libraries/sfm-docker.git
cd sfm-docker
cp example.prod.docker-compose.yml docker-compose.yml
cp example.env .env
```

or just download `example.prod.docker-compose.yml` and `example.env`:

```
curl -L https://github.com/gwu-libraries/sfm-docker/raw/master/example.prod.docker-compose.yml > doc
curl -L https://github.com/gwu-libraries/sfm-docker/raw/master/example.env > .env
```

2. Update configuration in `.env` as described in *Configuration*.
3. Bring up the containers:

```
docker-compose up -d
```

4. It is also recommended that you scale up the Twitter REST Harvester container:

```
docker-compose scale twitterrestharvester=2
```

Notes:

- The first time you bring up the containers, their images will be pulled from Docker Hub. This will take several minutes.
- For instructions on how to make configuration changes *after* the containers have been brought up, see *Configuration*.
- To learn more about scaling , see *Scaling up with Docker*.

6.3 Amazon EC2 installation

To launch an Amazon EC2 instance running SFM, follow the normal procedure for launching an instance. In *Step 3: Configure Instance Details*, under *Advanced Details* paste the following in user details and modify as appropriate as described in *Configuration*:

```
#cloud-config
repo_update: true
repo_upgrade: all

packages:
 - python-pip

runcmd:
 - curl -sSL https://get.docker.com/ | sh
 - usermod -aG docker ubuntu
 - pip install -U docker-compose
 - mkdir /sfm-data
 - mkdir /sfm-processing
 - cd /home/ubuntu
# This brings up the latest production release. To bring up master, remove prod.
 - curl -L https://github.com/gwu-libraries/sfm-docker/raw/master/example.prod.docker-compose.yml > d
 - curl -L https://github.com/gwu-libraries/sfm-docker/raw/master/example.env > .env
# Set config below by uncommenting.
# Don't forget to escape $ as \$.
# COMMON CONFIGURATION
# - echo TZ=America/New_York >> .env
# VOLUME CONFIGURATION
# Don't change these.
 - echo DATA_VOLUME=/sfm-data:/sfm-data
 - echo PROCESSING_VOLUME=/sfm-processing:/sfm-processing
# SFM UI CONFIGURATION
# Don't change this.
 - echo SFM_HOSTNAME=`curl http://169.254.169.254/latest/meta-data/public-hostname` >> .env
 - echo SFM_PORT=80 >> .env
# To send email, set these correctly.
# - echo SFM_SMTP_HOST=smtp.gmail.com >> .env
# - echo SFM_EMAIL_USER=someone@gmail.com >> .env
# - echo SFM_EMAIL_PASSWORD=password >> .env
# To enable connecting to social media accounts, provide the following.
# - echo TWITTER_CONSUMER_KEY=mBbq9ruffgEcfsktgQztTHUir8Kn0 >> .env
# - echo TWITTER_CONSUMER_SECRET=Pf28yReB9Xgz0fpLVO4b46r5idZnKCKQ6xlOomBAjD5npFEQ6Rm >> .env
# - echo WEIBO_API_KEY=13132044538 >> .env
# - echo WEIBO_API_SECRET=68aea49fg26ea5072ggec14f7c0e05a52 >> .env
# - echo TUMBLR_CONSUMER_KEY=Fki09cW957y56h6fhRtCnig14QhpM0pjuHbDWMrZ9aPXcsthVQq >> .env
```

```
# - echo TUMBLR_CONSUMER_SECRET=aPTpFRE2O7sVl46xB3difn8kBYb7EpnWfUBWxuHcB4gfvP >> .env
# For automatically created admin account
# - echo SFM_SITE_ADMIN_NAME=sfmadmin >> .env
# - echo SFM_SITE_ADMIN_EMAIL=nowhere@example.com >> .env
# - echo SFM_SITE_ADMIN_PASSWORD=password >> .env
# RABBIT MQ CONFIGURATION
# - echo RABBITMQ_USER=sfm_user >> .env
# - echo RABBITMQ_PASSWORD=password >> .env
# - echo RABBITMQ_MANAGEMENT_PORT=15672 >> .env
# DB CONFIGURATION
# - echo POSTGRES_PASSWORD=password >> .env
# WEB HARVESTER CONFIGURATION
# - echo HERITRIX_USER=sfm_user >> .env
# - echo HERITRIX_PASSWORD=password >> .env
# - echo HERITRIX_ADMIN_PORT=8443 >> .env
# - echo HERITRIX_CONTACT_URL=http://library.myschool.edu >> .env
 - docker-compose up -d
 - docker-compose scale twitterrestharvester=2
```

When the instance is launched, SFM will be installed and started.

Note the following:

- Starting up the EC2 instance will take several minutes.
- This has been tested with *Ubuntu Server 14.04 LTS*, but may work with other AMI types.
- We don't have recommendations for sizing, but providing multiple processors even for testing/experimentation is suggested.
- If you need to make additional changes to your `docker-compose.yml`, you can ssh into the EC2 instance and make changes. `docker-compose.yml` and `.env` will be in the default user's home directory.
- Make sure to configure a security group that exposes the proper ports. To see which ports are used by which services, see example.prod.docker-compose.yml.
- To learn more about configuring EC2 instances with user data, see the AWS user guide.

6.4 Configuration

Configuration is documented in `example.env`. For a production deployment, pay particular attention to the following:

- Set new passwords for `SFM_SITE_ADMIN_PASSWORD`, `RABBIT_MQ_PASSWORD`, `POSTGRES_PASSWORD`, and `HERITRIX_PASSWORD`.
- The data volume strategy is used to manage the volumes that store SFM's data. By default, normal Docker volumes are used. To use a host volume instead, change the `DATA_VOLUME` and `PROCESSING_VOLUME` settings. Host volumes are recommended for production because they allow access to the data from outside of Docker.
- Set the `SFM_HOSTNAME` and `SFM_PORT` appropriately. These are the public hostname (e.g., sfm.gwu.edu) and port (e.g., 80) for SFM.
- Email is configured by providing `SFM_SMTP_HOST`, `SFM_EMAIL_USER`, and `SFM_EMAIL_PASSWORD`. (If the configured email account is hosted by Google, you will need to configure the account to "Allow less secure apps." Currently this setting is accessed, while logged in to the google account, via https://myaccount.google.com/security#connectedapps).

- Application credentials for social media APIs are configured in by providing the `TWITTER_CONSUMER_KEY`, `TWITTER_CONSUMER_SECRET`, `WEIBO_API_KEY`, `WEIBO_API_SECRET`, and/or `TUMBLR_CONSUMER_KEY`, `TUMBLR_CONSUMER_SECRET`. These are optional, but will make acquiring credentials easier for users. For more information and alternative approaches see API Credentials.
- Set an admin email address with `SFM_SITE_ADMIN_EMAIL`.
- Provide a contact URL (e.g., http://library.gwu.edu) to be used when web harvesting with `HERITRIX_CONTACT_URL`.

Note that if you make a change to configuration *after* SFM is brought up, you will need to restart containers. If the change only applies to a single container, then you can stop the container with `docker kill <container name>`. If the change applies to multiple containers (or you're not sure), you can stop all containers with `docker-compose stop`. Containers can then be brought back up with `docker-compose up -d` and the configuration change will take effect.

CHAPTER 7

Authentication

Social Feed Manager allows users to self-sign up for accounts. Those accounts are stored and managed by SFM. Future versions of SFM will support authentication against external systems, e.g., Shibboleth.

By default, a group is created for each user and the user is placed in group. To create additional groups and modify group membership use the Admin interface.

In general, users and groups can be administered from the Admin interface.

The current version of SFM is not very secure. Future versions of SFM will more tightly restrict what actions users can perform and what they can view. In the meantime, it is encouraged to take other measures to secure SFM such as restricting access to the IP range of your institution.

CHAPTER 8

Docker

This page contains information about Docker that is useful for installation, administration, and development.

8.1 Installing Docker

Docker Engine and Docker Compose

On OS X:

- Install the Docker for Mac.
- If you are using Docker Toolbox, switch to Docker for Mac.

On Ubuntu:

- If you have difficulties with the `apt` install, try the `pip` install.
- The docker group is automatically created. Adding your user to the docker group avoids having to use sudo to run docker commands. Note that depending on how users/groups are set up, you may need to manually need to add your user to the group in `/etc/group`.

8.2 Helpful commands

`docker-compose up -d` Bring up all of the containers specified in the docker-compose.yml file. If a container has not yet been pulled, it will be pulled. If a container has not yet been built it will be built. If a container has been stopped ("killed") it will be re-started. Otherwise, a new container will be created and started ("run").

`docker-compose pull` Pull the latest images for all of the containers specified in the docker-compose.yml file with the *image* field.

`docker-compose build` Build images for all of the containers specified in the docker-compose.yml file with the *build* field. Add `--no-cache` to re-build the entire image (which you might want to do if the image isn't building as expected).

`docker ps` List running containers. Add `-a` to also list stopped containers.

`docker-compose kill` Stop all containers.

`docker kill <container name>` Stop a single container.

`docker-compose rm -v --force` Delete the containers and volumes.

`docker rm -v <container name>` Delete a single container and volume.

docker rm $(docker ps -a -q) -v Delete all containers.

docker-compose logs List the logs from all containers. Add `-f` to follow the logs.

docker logs <container name> List the log from a single container. Add `-f` to follow the logs.

docker-compose -f <docker-compose.yml filename> <command> Use a different docker-compose.yml file instead of the default.

docker exec -it <container name> /bin/bash Shell into a container.

docker rmi <image name> Delete an image.

docker rmi $(docker images -q) Delete all images

docker-compose scale <service name>=<number of instances> Create multiple instances of a service.

8.3 Scaling up with Docker

Most harvesters and exporters handle one request at a time; requests for exports and harvests queue up waiting to be handled. If requests are taking too long to be processed you can scale up (i.e., create additional instances of) the appropriate harvester or exporter.

To create multiple instances of a service, use docker-compose scale.

The harvester most likely to need scaling is the Twitter REST harvester since some harvests (e.g., broad Twitter searches) may take a long time. To scale up the Twitter REST harvester to 3 instances use:

```
docker-compose scale twitterrestharvester=3
```

To spread containers across multiple containers, use Docker Swarm.

Using compose in production provides some additional guidance.

CHAPTER 9

Limitations and Known Issues

To make sure you have the best possible experience with SFM, you should be aware of the limitations and known issues:

- Twitter REST harvesters can become congested when there are too many long-running harvests using too few unique credentials (Ticket 472)
- Better monitoring and logging for harvesting and exporting is needed (Ticket #303 and Ticket #229)
- Collections are not portable between SFM instances (Ticket #326)
- SFM is not secure. It does not currently run with HTTPS and enforcement of authorizations is not consistent in the UI (Ticket #362).
- Web harvester (Heritrix) does not handle deduping (Ticket #438)
- Huge exports are not segmented into multiple files (Ticket #454)
- Because of the need to link a Heritrix container and a web harvester container, the web harvester cannot be scaled with `docker-compose scale command` (Ticket 408)
- Changes to the hostname of server (e.g., from the reboot of an AWS EC2 instance) are not handled (Ticket 435)

We are planning to address these in future releases. In the meantime, there are work-arounds for many of these issues. For a complete list of tickets, see https://github.com/gwu-libraries/sfm-ui/issues

In addition, you should be aware of the following:

- Access to the Weibo API is limited, so make sure you understand what can be collected.
- SFM does not currently provide a web interface for "replaying" the collected social media or web content.
- ELK is only experimental. Scaling and administration of ELK have not been considered.

CHAPTER 10

Troubleshooting

10.1 General tips

- Upgrade to the latest version of Docker and Docker-Compose.
- Make sure expected containers are running with `docker ps`.
- Check the logs with `docker-compose logs` and `docker logs <container name>`.
- Additional information is available via the admin interface that is not available from the UI. To access the admin interface, log in as an account that has superuser status and under "Welcome, <your name>," click Admin. By default, a superuser account called *sfmadmin* is created. The password can be found in `.env`.

10.2 Specific problems

10.2.1 Bind error

If when bringing up the containers you receive something like:

```
ERROR: driver failed programming external connectivity on endpoint docker_sfmuiapp_1 (98caab29b4ba3c2
```

it means another application is already using a port configured for SFM. Either shut down the other application or choose a different port for SFM. (Chances are the other application is Apache.)

10.2.2 Bad Request (400)

If you receive a Bad Request (400) when trying to access SFM, your `SFM_HOST` environment variable is not configured correctly. For more information, see ALLOWED_HOSTS.

10.2.3 Social Network Login Failure for Twitter

If you receive a Social Network Login Failure when trying to connect a Twitter account, make sure that the Twitter app from which you got the Twitter credentials is configured with a callback URL. The URL you provide doesn't matter.

10.2.4 Docker problems

If you are having problems bringing up the Docker containers (e.g., `driver failed programming external connectivity on endpoint`), restart the Docker service. On Ubuntu, this can be done with:

```
# service docker stop
docker stop/waiting
# service docker status
docker stop/waiting
# service docker start
docker start/running, process 15039
```

10.3 Still stuck?

Contact the SFM team. We're happy to help.

CHAPTER 11

Development

11.1 Setting up a development environment

SFM is composed of a number of components. Development can be performed on each of the components separately.

For SFM development, it is recommended to run components within a Docker environment (instead of directly in your OS, without Docker).

11.1.1 Step 1: Install Docker and Docker Compose

See *Installing Docker*.

11.1.2 Step 2: Clone sfm-docker and create copies of docker-compose files

For example:

```
git clone https://github.com/gwu-libraries/sfm-docker.git
cd sfm-docker
cp example.docker-compose.yml docker-compose.yml
cp example.env .env
```

For the purposes of development, you can make changes to `docker-compose.yml` and `.env`. This will be described more below.

11.1.3 Step 3: Clone the component repos

For example:

```
git clone https://github.com/gwu-libraries/sfm-ui.git
```

Repeat for each of the components that you will be working on. Each of these should be in a sibling directory of sfm-docker.

11.2 Running SFM for development

To bring up an instance of SFM for development, change to the sfm-docker directory and execute:

```
docker-compose up -d
```

You may not want to run all of the containers. To omit a container, simply comment it out in `docker-compose.yml`.

By default, the code that has been committed to master for each of the containers will be executed. To execute your local code (i.e., the code you are editing), you will want to link in your local code. To link in the local code for a container, uncomment the volume definition that points to your local code. For example:

```
volumes:
    - "../sfm-twitter-harvester:/opt/sfm-twitter-harvester"
```

sfm-utils and warcprox are dependencies of many components. By default, the code that has been committed to master for sfm-utils or warcprox will be used for a component. To use your local code as a dependency, you will want to link in your local code. Assuming that you have cloned sfm-utils and warcprox, to link in the local code as a dependency for a container, change `SFM_REQS` in `.env` to "dev" and comment the volume definition that points to your local code. For example:

```
volumes:
    - "../sfm-twitter-harvester:/opt/sfm-twitter-harvester"
    - "../sfm-utils:/opt/sfm-utils"
    - "../warcprox:/opt/warcprox"
```

Note: * As a Django application, SFM UI will automically detect code changes and reload. Other components must be killed and brought back up to reflect code changes.

11.3 Running tests

11.3.1 Unit tests

Some components require a `test_config.py` file that contains credentials. For example, sfm-twitter-harvester requires a `test_config.py` containing:

```
TWITTER_CONSUMER_KEY = "EHdoTksBfgGflP5nUalEfhaeo"
TWITTER_CONSUMER_SECRET = "ZtUpemtBkf2cEmaqiy52Dd343ihFu9PAiLebuMOmqN0QtXeAlen"
TWITTER_ACCESS_TOKEN = "411876914-c2yZjbk1np0Z5MWEFYYQKSQNFFGBXd8T4k90YkJl"
TWITTER_ACCESS_TOKEN_SECRET = "jK9QOmn5VRF5mfgAN6KgfmCKRqThXVQ1G6qQg8BCejvp"
```

Note that if this file is not present, unit tests that require it will be skipped. Each component's README will describe the `test_config.py` requirements.

Unit tests for most components can be run with:

```
python -m unittest discover
```

The notable exception is SFM UI, which can be run with:

```
cd sfm
./manage.py test --settings=sfm.settings.test_settings
```

11.3.2 Integration tests

Many components have integration tests, which are run inside docker containers. These components have a `ci.docker-compose.yml` file which can be used to bring up a minimal environment for running the tests.

As described above, some components require a `test_config.py` file.

To run integration tests, bring up SFM:

```
docker-compose -f docker/dev.docker-compose.yml up -d
```

Run the tests:

```
docker exec docker_sfmtwitterstreamharvester_1 python -m unittest discover
```

You will need to substitute the correct name of the container. (`docker ps` will list the containers.)

And then clean up:

```
docker-compose -f docker/dev.docker-compose.yml kill
docker-compose -f docker/dev.docker-compose.yml rm -v --force
```

For reference, see each component's `.travis.yml` file which shows the steps of running the integration tests.

11.3.3 Smoke tests

sfm-docker contains some smoke tests which will verify that SFM is running correctly.

To run the smoke tests, first bring up SFM:

```
docker-compose up -d
```

and then run the tests:

```
docker-compose -f docker-compose.yml -f smoketests.docker-compose.yml run --rm smoketests python -m u
```

Note that the smoke tests are not yet complete.

For reference, the continuous integration deploy instructions shows the steps of running the smoke tests.

11.4 Requirements files

This will vary a depending on whether a project has warcprox and sfm-utils as a dependency, but in general:

- `requirements/common.txt` contains dependencies, except warcprox and sfm-utils.
- `requirements/release.txt` references the last released version of warcprox and sfm-utils.
- `requirements/master.txt` references the master version of warcprox and sfm-utils.
- `requirements/dev.txt` references local versions of warcprox and sfm-utils in development mode.

To get a complete set of dependencies, you will need `common.txt` and either `release.txt`, `master.txt` or `dev.txt`. For example:

```
virtualenv ENV
source ENV/bin/activate
pip install -r requirements/common.txt -r requirements/dev.txt
```

11.5 Development tips

11.5.1 Admin user accounts

Each component should automatically create any necessary admin accounts (e.g., a django admin for SFM UI). Check `.env` for the username/passwords for those accounts.

11.5.2 RabbitMQ management console

The RabbitMQ management console can be used to monitor the exchange of messages. In particular, use it to monitor the messages that a component sends, create a new queue, bind that queue to *sfm_exchange* using an appropriate routing key, and then retrieve messages from the queue.

The RabbitMQ management console can also be used to send messages to the exchange so that they can be consumed by a component. (The exchange used by SFM is named *sfm_exchange*.)

For more information on the RabbitMQ management console, see *RabbitMQ*.

11.5.3 Blocked ports

When running on a remote VM, some ports (e.g., 15672 used by the RabbitMQ management console) may be blocked. SSH port forwarding can help make those ports available.

11.5.4 Django logs

Django logs for SFM UI are written to the Apache logs. In the docker environment, the level of various loggers can be set from environment variables. For example, setting *SFM_APSCHEDULER_LOG* to *DEBUG* in the *docker-compose.yml* will turn on debug logging for the apscheduler logger. The logger for the SFM UI application is called ui and is controlled by the *SFM_UI_LOG* environment variable.

11.5.5 Apache logs

In the SFM UI container, Apache logs are sent to stdout/stderr which means they can be viewed with *docker-compose logs* or *docker logs <container name or id>*.

11.5.6 Initial data

The development and master docker images for SFM UI contain some initial data. This includes a user ("testuser", with password "password"). For the latest initial data, see *fixtures.json*. For more information on fixtures, see the Django docs.

11.5.7 Runserver

There are two flavors of the the development docker image for SFM UI. *gwul/sfm-ui:master* runs SFM UI with Apache, just as it will in production. *gwul/sfm-ui:master-runserver* runs SFM UI with runserver, which dynamically reloads changed Python code. To switch between them, change *UI_TAG* in *.env*.

Note that as an byproduct of how runserver dynamically reloads Python code, there are actually 2 instances of the application running. This may produce some odd results, like 2 schedulers running. This will not occur with Apache.

11.5.8 Job schedule intervals

To assist with testing and development, a 5 minute interval can be added by setting *SFM_FIVE_MINUTE_SCHEDULE* to *True* in the *docker-compose.yml*.

11.5.9 Connecting to the database

To connect to postgres using psql:

```
docker exec -it sfm_db_1 psql -h db -U postgres -d sfmdatabase
```

You will be prompted for the password, which you can find in *.env*.

11.6 Docker tips

11.6.1 Building vs. pulling

Containers are created from images. Images are either built locally or pre-built and pulled from Docker Hub. In both cases, images are created based on the docker build (i.e., the Dockerfile and other files in the same directory as the Dockerfile).

In a docker-compose.yml, pulled images will be identified by the *image* field, e.g., *image: gwul/sfm-ui:master*. Built images will be identified by the *build* field, e.g., *build: app-dev*.

In general, you will want to use pulled images. These are automatically built when changes are made to the Github repos. You should periodically execute *docker-compose pull* to make sure you have the latest images.

You may want to build your own image if your development requires a change to the docker build (e.g., you modify fixtures.json).

11.6.2 Killing, removing, and building in development

Killing a container will cause the process in the container to be stopped. Running the container again will cause process to be re-started. Generally, you will kill and run a development container to get the process to be run with changes you've made to the code.

Removing a container will delete all of the container's data. During development, you will remove a container to make sure you are working with a clean container.

Building a container creates a new image based on the Dockerfile. For a development image, you only need to build when making changes to the docker build.

CHAPTER 12

Writing a harvester

12.1 Requirements

- Implement the Messaging Specification for harvesting social media content. This describes the messages that must be consumed and produced by a harvester.
- Write harvested social media to a WARC, following all relevant guidelines and best practices. The message for announcing the creation of a WARC is described in the Messaging Specification. The WARC file must be written to *<base path>/<harvest year>/<harvest month>/<harvest day>/<harvest hour>/*, e.g., */data/test_collection_set/2015/09/12/19/*. (Base path is provided in the harvest start message.) Any filename may be used but it must end in *.warc* or *.warc.gz*. It is recommended that the filename include the harvest id (with file system unfriendly characters removed) and a timestamp of the harvest.
- Extract urls for related content from the harvested social media content, e.g., a photo included in a tweet. The message for publishing the list of urls is described in the Messaging Specification.
- Document the harvest types supported by the harvester. This should include the identifier of the type, the API methods called, the required parameters, the optional parameters, what is included in the summary, and what urls are extracted. See the Flickr Harvester as an example.
- The smoke tests must be able to prove that a harvester is up and running. At the very least, the smoke tests should check that the queues required by a harvester have been created. (See test_queues().)
- Be responsible for its own state, e.g., keeping track of the last tweet harvested from a user timeline. See sfmutils.state_store for re-usable approaches to storing state.
- Create all necessary exchanges, queues, and bindings for producing and consuming messages as described in Messaging.
- Provide master and production Docker images for the harvester on Docker Hub. The master image should have the *master* tag and contain the latest code from the master branch. (Setup an automated build to simplify updating the master image.) There must be a version specific production images, e.g., *1.3.0* for each release. For example, see the Flickr Harvester's dockerfiles and Docker Hub repo.

12.2 Suggestions

- See sfm-utils for re-usable harvester code. In particular, consider subclassing BaseHarvester.
- Create a development Docker image. The development Docker images links in the code outside of the container so that a developer can make changes to the running code. For example, see the Flickr harvester development image.

- Create a development *docker-compose.yml*. This should include the development Docker image and only the additional images that the harvester depends on, e.g., a Rabbit container. For example, see the Flickr harvester development docker-compose.yml.
- When possible, use existing API libraries.
- Consider write integration tests that test the harvester in an integration test environment. (That is, an environment that includes the other services that the harvester depends on.) For example, see the Flickr Harvester's integration tests.
- See the Twitter harvester unit tests for a pattern on configuring API keys in unit and integration tests.

12.3 Notes

- Harvesters can be written in any programming language.
- Changes to gwu-libraries/* repos require pull requests. Pull requests are welcome from non-GWU developers.

CHAPTER 13

Messaging

13.1 RabbitMQ

RabbitMQ is used as a message broker.

The RabbitMQ managagement console is exposed at `http://<your docker host>:15672/`. The username is `sfm_user`. The password is the value of `RABBITMQ_DEFAULT_PASS` in `secrets.env`.

13.2 Publishers/consumers

- The hostname for RabbitMQ is `mq` and the port is 5672.
- It cannot be guaranteed that the RabbitMQ docker container will be up and ready when any other container is started. Before starting, wait for a connection to be available on port 5672 on `rabbit`. See appdeps.py for docker application dependency support.
- Publishers/consumers may not assume that the requisite exchanges/queues/bindings have previously been created. They must declare them as specified below.

13.3 Exchange

`sfm_exchange` is a durable topic exchange to be used for all messages. All publishers/consumers must declare it.:

```
#Declare sfm_exchange
from kombu import Connection

exchange = Exchange(name="sfm_exchange,
                    type="topic", durable=True)
exchange(channel).declare()
```

13.4 Queues

All queues must be declared durable.:

```
#Declare harvester queue
from kombu import Queue
queue = Queue(name="harvester",
              exchange=exchange,
              channel=channel,
              durable=True)
queue.declare()
queue.bind_to(exchange=exchange,
              routing_key="harvest.status.*.*")
```

CHAPTER 14

Messaging Specification

14.1 Introduction

SFM is architected as a number of components that exchange messages via a messaging queue. To implement functionality, these components send and receive messages and perform certain actions. The purpose of this document is to describe this interaction between the components (called a "flow") and to specify the messages that they will exchange.

Note that as additional functionality is added to SFM, additional flows and messages will be added to this document.

14.2 General

- Messages may include extra information beyond what is specified below. Message consumers should ignore any extra information.
- RabbitMQ will be used for the messaging queue. See the Messaging docs for additional information. It is assumed in the flows below that components receive messages by connecting to appropriately defined queues and publish messages by submitting them to the appropriate exchange.

14.3 Harvesting social media content

Harvesting is the process of retrieving social media content from the APIs of social media services and writing to WARC files. It also includes extracting urls for other web resources from the social media so that they can be harvested by a web harvester. (For example, the link for an image may be extracted from a tweet.)

14.3.1 Background information

- A requester is an application that requests that a harvest be performed. A requester may also want to monitor the status of a harvest. In the current architecture, the SFM UI serves the role of requester.
- A stream harvest is a harvest that is intended to continue indefinitely until terminated. A harvest of a Twitter public stream is an example of a stream harvest. A stream harvest is different from a non-stream harvest in that a requester must both start and optionally stop a stream harvest. Following the naming conventions from Twitter, a harvest of a REST, non-streaming API will be referred to as a REST harvest.
- Depending on the implementation, a harvester may produce a single warc or multiple warcs. It is likely that in general stream harvests will result in multiple warcs, but REST harvest will result in a single warc.

14.3.2 Flow

The following is the flow for a harvester performing a REST harvest and creating a single warc:

1. Requester publishes a harvest start message.
2. Upon receiving the harvest message, a harvester:
   (a) Makes the appropriate api calls.
   (b) Extracts urls for web resources from the results.
   (c) Writes the api calls to a warc.
3. Upon completing the api harvest, the harvester:
   (a) Publishes a web harvest message containing the extracted urls.
   (b) Publishes a warc created message.
   (c) Publishes a harvest status message with the status of *completed success* or *completed failure*.

The following is the message flow for a harvester performing a stream harvest and creating multiple warcs:

1. Requester publishes a harvest start message.
2. Upon receiving the harvest message, a harvester:
   (a) Opens the api stream.
   (b) Extracts urls for web resources from the results.
   (c) Writes the stream results to a warc.
3. When rotating to a new warc, the harvester publishes a warc created message.
4. At intervals during the harvest, the harvester:
   (a) Publishes a web harvest message containing extracted urls.
   (b) Publishes a harvest status message with the status of *running*.
5. When ready to stop, the requester publishes a harvest stop message.
6. Upon receiving the harvest stop message, the harvester:
   (a) Closes the api stream.
   (b) Publishes a final web harvest message containing extracted urls.
   (c) Publishes a final warc created message.
   (d) Publishes a final harvest status message with the status of *completed success* or *completed failure*.

- Any harvester may send harvest status messages with the status of *running* before the final harvest status message. A harvester performing a stream harvest must send harvest status messages at regular intervals.
- A requester should not send harvest stop messages for a REST harvest. A harvester performing a REST harvest may ignore harvest stop messages.

14.3.3 Messages

Harvest start message

Harvest start messages specify for a harvester the details of a harvest. Example:

```
{
    "id": "sfmui:45",
    "type": "flickr_user",
    "path": "/sfm-data/collections/3989a5f99e41487aaef698680537c3f5/6980fac666c54322a2ebdbcb2a9510f5"
    "seeds": [
        {
            "id": "a36fe186fbfa47a89dbb0551e1f0f181",
            "token": "justin.littman",
            "uid": "131866249@N02"
        },
        {
            "id": "ab0a4d9369324901a890ec85f00194ac",
            "token": "library_of_congress"
        }
    ],
    "options": {
        "sizes": ["Thumbnail", "Large", "Original"]
    },
    "credentials": {
        "key": "abddfe6fb8bba36e8ef0278ec65dbbc8",
        "secret": "1642649c54cc3ebe"
    },
    "collection_set": {
        "id": "3989a5f99e41487aaef698680537c3f5"
    }
}
```

Another example:

```
{
    "id": "test:1",
    "type": "twitter_search",
    "path": "/sfm-data/collections/3989a5f99e41487aaef698680537c3f5/6980fac666c54322a2ebdbcb2a9510f5"
    "seeds": [
        {
            "id": "32786222ef374eb38f1c5d56321c99e8",
            "token": "gwu"
        },
        {
            "id": "0e789cddd0fb41b5950f569676702182",
            "token": "gelman"
        }
    ],
    "credentials": {
        "consumer_key": "EHde7ksBGgflbP5nUalEfhaeo",
        "consumer_secret": "ZtUpemtBkf2maqFiy52D5dihFPAiLebuMOmqN0jeQtXeAlen",
        "access_token": "481186914-c2yZjgbk13np0Z5MWEFQKSQNFBXd8T9r4k90YkJl",
        "access_token_secret": "jK9QOmn5Vbbmfg2ANT6KgfmKRqV8ThXVQ1G6qQg8BCejvp"
    },
    "collection_set": {
        "id": "3989a5f99e41487aaef698680537c3f5"
    }
}
```

- The routing key will be *harvest.start.<social media platform>.<type>*. For example, *harvest.start.flickr.flickr_photo*.
- *id*: A globally unique identifier for the harvest, assigned by the requester.

- *type*: Identifies the type of harvest, including the social media platform. The harvester can use this to map to the appropriate api calls.
- *seeds*: A list of seeds to harvest. Each seed is represented by a map containing *id*, *token* and (optionally) *uid*. Note that some harvest types may not have seeds.
- *options*: A name/value map containing additional options for the harvest. The contents of the map are specific to the type of harvest. (That is, the seeds for a flickr photo are going to be different than the seeds for a twitter user timeline.)
- *credentials*: All credentials that are necessary to access the social media platform. Credentials is a name/value map; the contents are specific to a social media platform.
- *path*: The base path for the collection.

Web resource harvest start message

Harvesters will extract urls from the harvested social media content and publish a web resource harvest start message. This message is similar to other harvest start messages, with the differences noted below. Example:

```
{
    "id": "flickr:45",
    "parent_id": "sfmui:45",
    "type": "web",
    "path": "/sfm-data/collections/3989a5f99e41487aaef698680537c3f5/6980fac666c54322a2ebdbcb2a9510f5"
    "seeds": [
        {
            "id": "3724fd97e85345ee84f5175eee09748d",
            "token": "http://www.gwu.edu/"
        },
        {
            "id": "aba6033aafce4fbabd846026ca47f13e",
            "token": "http://library.gwu.edu/"
        }
    ],
    "collection_set": {
        "id": "3989a5f99e41487aaef698680537c3f5"
    }
}
```

- The routing key will be *harvest.start.web*.
- *parent_id*: The id of the harvest from which the urls were extracted.

Harvest stop message

Harvest stop messages tell a harvester perform a stream harvest to stop. Example:

```
{
    "id": "sfmui:45"
}
```

- The routing key will be *harvest.stop.*<*social media platform*>.<*type*>. For example, *harvest.stop.twitter.filter*.

Harvest status message

Harvest status messages allow a harvester to provide information on the harvests it performs. Example:

```
{
    "id": "sfmui:45"
    "status": "completed success",
    "date_started": "2015-07-28T11:17:36.640044",
    "date_ended": "2015-07-28T11:17:42.539470",
    "infos": []
    "warnings": [],
    "errors": [],
    "stats": {
        "2016-05-20": {
            "photos": 12,
        },
        "2016-05-21": {
            "photos": 19,
        },
    },
    "token_updates": {
        "a36fe186fbfa47a89dbb0551e1f0f181": "j.littman"
    },
    "uids": {
        "ab0a4d9369324901a890ec85f00194ac": "671366249@N03"
    },
    "warcs": {
        "count": 3
        "bytes": 345234242
    }
}
```

- The routing key will be *harvest.status.<social media platform>.<type>*. For example, *harvest.status.flickr.flickr_photo*.
- *status*: Valid values are *completed success*, *completed failure*, or *running*.
- *infos*, *warnings*, and *errors*: Lists of messages. A message should be an object (i.e., dictionary) containing a *code* and *message* entry. Codes should be consistent to allow message consumers to identify types of messages.
- *stats*: A count of items that are harvested by date. Items should be a human-understandable labels (plural and lower-cased). Stats is optional for in progress statuses, but required for final statuses.
- *token_updates*: A map of uids to tokens for which a token change was detected while harvesting. For example, for Twitter a token update would be provided whenever a user's screen name changes.
- *uids*: A map of tokens to uids for which a uid was identified while harvesting at not provided in the harvest start message. For example, for Flickr a uid would be provided containing the NSID for a username.
- *warcs*.'count': The total number of WARCs created during this harvest.
- *warcs*.'bytes': The total number of bytes of the WARCs created during this harvest.

Warc created message

Warc created message allow a harvester to provide information on the warcs that are created during a harvest. Example:

```
{
    "warc": {
        "path": "/sfm-data/collections/3989a5f99e41487aaef698680537c3f5/6980fac666c54322a2ebdbcb2a951
        "sha1": "7512e1c227c29332172118f0b79b2ca75cbe8979",
        "bytes": 26146,
        "id": "aba6033aafce4fbabd846026ca47f13e",
```

```
        "date_created": "2015-07-28T11:17:36.640178"
    },
    "collection_set": {
        "id": "3989a5f99e41487aaef698680537c3f5"
    },
    "harvest": {
        "id": "98ddaa6e8c1f4b44aaca95bc46d3d6ac",
        "type": "flickr_user"
    }
}
```

- The routing key will be *warc_created*.
- Each warc created message will be for a single warc.

14.4 Exporting social media content

Exporting is the process of extracting social media content from WARCs and writing to export files. The exported content may be a subset or derivate of the original content. A number of different export formats will be supported.

14.4.1 Background information

- A requester is an application that requests that an export be performed. A requester may also want to monitor the status of an export. In the current architecture, the SFM UI serves the role of requester.
- Depending on the nature of the export, a single or multiple files may be produced.

14.4.2 Flow

The following is the flow for an export:

1. Requester publishes an export start message.
2. Upon receiving the export start message, an exporter:
   (a) Makes calls to the SFM REST API to determine the WARC files from which to export.
   (b) Limits the content is specified by the export start message.
   (c) Writes to export files.
3. Upon completing the export, the exporter publishes an export status message with the status of *completed success* or *completed failure*.

Export start message

Export start messages specify the requests for an export. Example:

```
{
    "id": "f3ddcbfc5d6b43139d04d680d278852e",
    "type": "flickr_user",
    "collection": {
        "id": "005b131f5f854402afa2b08a4b7ba960"
    },
    "path": "/sfm-data/exports/45",
```

```
    "format": "csv",
    "dedupe": true,
    "item_date_start": "2015-07-28T11:17:36.640178",
    "item_date_end": "2016-07-28T11:17:36.640178",
    "harvest_date_start": "2015-07-28T11:17:36.640178",
    "harvest_date_end": "2016-07-28T11:17:36.640178"
}
```

Another example:

```
{
    "id": "f3ddcbfc5d6b43139d04d680d278852e",
    "type": "flickr_user",
    "seeds": [
        {
            "id": "48722ac6154241f592fd74da775b7ab7",
            "uid": "23972344@N05"
        },
        {
            "id": "3ce76759a3ee40b894562a35359dfa54",
            "uid": "85779209@N08"
        }
    ],
    "path": "/sfm-data/exports/45",
    "format": "json"
}
```

- The routing key will be *export.start.<social media platform>.<type>*. For example, *export.start.flickr.flickr_user*.
- *id*: A globally unique identifier for the harvest, assigned by the requester.
- *type*: Identifies the type of export, including the social media platform. The export can use this to map to the appropriate export procedure.
- *seeds*: A list of seeds to export. Each seed is represented by a map containing *id* and *uid*.
- *collection*: A map containing the *id* of the collection to export.
- Each export start message must have a *seeds* or *collection* but not both.
- *path*: A directory into which the export files should be placed. The directory may not exist.
- *format*: A code for the format of the export. (Available formats may change.)
- *dedupe*: If true, duplicate social media content should be removed.
- *item_date_start* and *item_date_end*: The date of social media content should be within this range.
- *harvest_date_start* and *harvest_date_end*: The harvest date of social media content should be within this range.

Export status message

Export status messages allow an exporter to provide information on the exports it performs. Example:

```
{
    "id": "f3ddcbfc5d6b43139d04d680d278852e"
    "status": "completed success",
    "date_started": "2015-07-28T11:17:36.640044",
    "date_ended": "2015-07-28T11:17:42.539470",
    "infos": []
```

```
    "warnings": [],
    "errors": [],
}
```

- The routing key will be *export.status.<social media platform>.<type>*. For example, *export.status.flickr.flickr_user*.
- *status*: Valid values are *completed success* or *completed failure*.
- *infos*, *warnings*, and *errors*: Lists of messages. A message should be an object (i.e., dictionary) containing a *code* and *message* entry. Codes should be consistent to allow message consumers to identify types of messages.

CHAPTER 15

Indices and tables

- genindex
- modindex
- search

15.1 Funding history

- Development of this project has been supported by a grant (#NARDI-14-50017-14) from the National Historical Publications & Records Commission to George Washington University Libraries from 2014-2017.
- Development of the Sina Weibo harvester is supported by a grant from the Council on East Asian Libraries.
- **Prior development of SFM under the previous repository** was supported by a grant (#LG-46-13-0257-13) from the Institute of Museum and Library Services to George Washington University Libraries from 2013-2014.