

Web scraping and social media scraping – APIs

Jacek Lewkowicz, Dorota Celińska-Kopczyńska

University of Warsaw

April 14, 2019

Internet as a source of data for research

- Imagine a highly popular website, e.g., Twitter, Reddit, etc
- There are millions of users, myriads of activities, tons of text messages...
- Which seems not only to be a great source of data for the needs of your master or doctoral thesis, but if utilized correctly might even earn you a tenure!
- The sea of immeasurable gains lies at the tip of your fingers... just scrape the data!

Problem #1

- The problem is... it is not only you who had the same idea

The screenshot shows a Google Scholar search for 'twitter'. The search results are displayed on a Firefox browser window. The search bar contains 'twitter' and the search button is highlighted. The results show about 6,230,000 results in 0.11 seconds. The first article is 'What is Twitter, a social network or a news media?' by H. Kwak, C. Lee, H. Park, and S. Moon, published in 2010. The second article is 'Why we twitter: understanding microblogging usage and communities' by A. Java, X. Song, T. Finin, B. Tseng, published in 2007. The third article is 'Earthquake shakes Twitter users: real-time event detection by social sensors' by T. Sakaki, M. Okazaki, and Y. Matsuo, published in 2010. The fourth article is 'Measuring user influence in twitter: The million follower fallacy' by M. Cha, H. Kwak, P. Benerchi, E. Bakshy, and D. Sussman, published in 2010. The fifth article is 'Twitter mood predicts the stock market' by J. Bolton, H. Mao, and Z. Zeng, published in 2011.

1 | 2 | 4 |

twitter - Google Scholar - Firefox Nightly

twitter - Google Scholar x +

https://scholar.google.com/scholar?as_vis=1&q=twitter&hl=en&as_sdt=1,5

Google Scholar

twitter

Articles About 6,230,000 results (0.11 sec)

Any time
Since 2018
Since 2017
Since 2014
Custom range...

Sort by relevance
Sort by date

Include patents
 Include citations

Create alert

What is Twitter, a social network or a news media?
H. Kwak, C. Lee, H. Park, S. Moon - ... of the 19th international conference on ..., 2010 - dl.acm.org
Abstract **Twitter**, a microblogging service less than three years old, commands more than 41 million users as of July 2009 and is growing fast. **Twitter** users tweet about any topic within the 140-character limit and follow others to receive their tweets. The goal of this paper is to ...
☆ Cited by 5902 Related articles All 33 versions [PDF](#) bgu.ac.il

Why we twitter: understanding microblogging usage and communities
A. Java, X. Song, T. Finin, B. Tseng - Proceedings of the 9th WebKDD and ..., 2007 - dl.acm.org
Abstract Microblogging is a new form of communication in which users can describe their current status in short posts distributed by instant messages, mobile phones, email or the Web. **Twitter**, a popular microblogging tool has seen a lot of growth since it launched in ...
☆ Cited by 3441 Related articles All 25 versions [PDF](#) umbc.edu

Earthquake shakes Twitter users: real-time event detection by social sensors
T. Sakaki, M. Okazaki, Y. Matsuo - ... of the 19th international conference on ..., 2010 - dl.acm.org
Abstract **Twitter**, a popular microblogging service, has received much attention recently. An important characteristic of **Twitter** is its real-time nature. For example, when an earthquake occurs, people make many **Twitter** posts (tweets) related to the earthquake, which enables ...
☆ Cited by 3315 Related articles All 33 versions [PDF](#) ethz.ch

Measuring user influence in twitter: The million follower fallacy.
M. Cha, H. Kwak, P. Benerchi, E. Bakshy, D. Sussman - Iccsm, 2010 - aaai.org
Abstract Directed links in social media could represent anything from intimate friendships to common interests, or even a passion for breaking news or celebrity gossip. Such directed links determine the flow of information and hence indicate a user's influence on others—a ...
☆ Cited by 2812 Related articles All 43 versions [PDF](#) aaai.org

Twitter mood predicts the stock market
J. Bolton, H. Mao, Z. Zeng - Journal of computational science, 2011 - Elsevier
Abstract Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, ie can societies experience mood states that affect their collective decision making? By extension is the ...
☆ Cited by 3124 Related articles All 28 versions [PDF](#) arxiv.org

Problem #2

- ... and sometimes “publish or perish” is really innocent
- Cambridge Analytica issue – the company is accused of buying millions of Americans’ data from a researcher who told Facebook he was collecting it strictly for academic purposes
- Which generally made accessing social media data real pain...

Solution

- In theory, the site admins should be prepared not only for the typical activity of the users but also for the activity of third parties that want to get access to the data
- In such cases providing and using APIs may save a lot of time (and trouble)

What is API?

- An **Application Programming Interface** is a set of subroutine definitions, protocols and tools for building application software
- It should make communication between various software components easier
- In web scraping usage of APIs usually boils down to sending requests messages and getting response messages usually in XML or JSON format
- Example:
`https://console.developers.google.com/apis/dashboard`

Data from APIs

- APIs encompass tools which enable programmers to connect their software with “something else”
- They are useful in programming software that relies on external soft- or hardware because the developers do not have to go into the details of external soft- or hardware mechanics

Scheme of usage

- The API provider sets up a service that grants access to data from the application or to the application itself
- The API user accesses the API to gather data or communicate with the application
- It may be necessary to write wrapper software for convenient data exchange with the web service

Common commands

- GET – visiting a website through the address bar in the browser
- POST – filling out a form/submitting information to a backend script on the server
- PUT – updating an object or information
- DELETE – deleting an object

Authentication

- Services may impose some restrictions in accessing data if the user is not authenticated
- The methods of authentication discussed earlier still hold
- If you (for some reason) provide sensitive data, remember to use, e.g., environment variables (or any other “relatively safe” technique)

```
token = "<your api key>"  
webRequest = urllib.request.Request("http://myapi.com", headers={"token":token})  
html = urlopen(webRequest)
```

Why is it interesting?

- Lots of active users
- A lot of topics: from public debate and political issues to breakups...
- Fast and easy access

Terms of Use and available information

- <http://developer.twitter.com/en/docs/>
- Account Activity API: manage account settings and interactions with users (e.g., follow, mute, search users)
- Post, retrieve and engage with tweets
- Direct Message API, Media and Trends

Authentication

- Twitter refers to an open standard of authentication that permits connections without sharing username/password
- All queries require a valid OAuth token
 - New application on dev.twitter.com or apps.twitter.com/app/new
 - Sign in and save your consumer key and consumer secret for future sessions
 - Python twitter tools:
<http://mike.verdone.ca/twitter/#downloads>

Twitter: an example

- We will search for tweets containing hashtag python
- Start with navigating to the directory in which you store twitter tools
- `$ python setup.py install`

```
# import the library
from twitter import Twitter

# authenticate
t = Twitter(auth=OAuth(<Access Token>,<Access Token Secret>, <Consumer Key>,<Consumer Secret>))

pythonTweets = t.search.tweets(q = "#python")
print(pythonTweets)
```

Twitter: an example

- We will retrieve the list of tweets by a given user
- For more examples see also

<http://github.com/sixohsix/twitter/tree/master>

```
# import the library
from twitter import Twitter

# authenticate
t = Twitter(auth=OAuth(<Access Token>,<Access Token Secret>, <Consumer Key>,<Consumer Secret>))

pythonStatuses = t.statuses.user_timeline(screen_name="POTUS", count=10)
print(pythonStatuses)
```

Why is it interesting?

- Lots of active users
- Differently themed boards, popular and less well-known ones
- Availability of searching for a particular user's activity in many boards

Terms of Use and available information

- <https://www.reddit.com/dev/api/>
- Everything a user can do (for your account)
- Getting comments, titles, overviews, upvotes and downvotes...

Authentication

- `https://github.com/reddit-archive/reddit/wiki/API`
- Clients must authenticate with a valid OAuth2 token
 - Details described here: `https://github.com/reddit-archive/reddit/wiki/OAuth2`

Reddit: an example

- We will search for posts and comments in subreddit boardgames

```
# import necessary libraries
import urllib3
import praw

urllib3.disable_warnings()

# provide authentication
r = praw.Reddit(user_agent='insert-here', client_id='your-id', client_secret='your-secret', redirect_uri='your-site',
username='your-login')

submissions = r.subreddit('boardgames').hot(limit=None)

for x in submissions:
    print("SUBMISSION BY "+str(x.author))
    print("ID "+str(x.id))
    print("TITLE "+x.title.encode('utf-8'))
    print("SCORE "+str(x.score))
    print("DATE "+str(x.created_utc))
    print(x.selftext.encode('utf-8'))
    # print(vars(x))
    print("")
    x.comments.replace_more(limit=None, threshold=0)
    comment_queue = x.comments[:] # Seed with top-level
    while comment_queue:
        comment = comment_queue.pop(0)
        print("COMMENT BY "+str(comment.author))
        print("ID "+str(comment.id))
        print("PARENTID "+str(comment.parent_id))
        print("DATE "+str(comment.created_utc))
        print(comment.body.encode('utf-8'))
        # print(vars(comment))
        print("")
        comment_queue.extend(comment.replies)
```