Collecting research data from Twitter’s public APIs

Kim Holmberg
Senior researcher, PhD
RUSE - Research Unit for the Sociology of Education
University of Turku, Finland
(e) kim.j.holmberg@utu.fi
(w3) http://kimholmberg.fi
@kholmber

Timothy D Bowman
Assistant Professor, PhD
School of Information Sciences
Wayne State University, USA
(e) timothy.d.bowman@wayne.edu
(w3) www.tdbowman.com
@timothydbowman
Developers

Tap into what's happening.

Publish and analyze Tweets, optimize ads, and create unique customer experiences.
Instant historical access to Tweets.

Understand the past to predict the future. Our Search APIs provide instant and complete access to the archive of public Twitter data, making it quick and easy to find the exact Tweets that matter to you. Whether you need a 7-day, 30-day, or 10 year look-back window, these endpoints provide the flexibility and access you need.

Explore endpoints

Example functionality

Standard: search/Tweets endpoint
Premium: Search Tweets API: 30-Days
Premium: Search Tweets API: Full-archive
Enterprise: 30-Day Search API
Enterprise: Full-Archive Search API
Filter

Complete coverage of Tweets as they happen.

From breaking news and entertainment to sports, politics, and everyday interests, Twitter is what’s happening in the world and what people are talking about right now. Our Filter APIs provide powerful filtering and coverage of the data you need, delivered in real-time. Get the data you need, the moment you need it.

Explore endpoints 

Example functionality

Filter Tweets API: statuses/filter
Enterprise: PowerTrack API
Snapshot of Tweets in real-time.

Sometimes a snapshot is all you need to understand what’s happening. Our Sample Tweets APIs return random samples of public Tweets as they happen. Discover unexpected trends and inform decisions in real-time.

Explore endpoints

Example functionality

Sample Tweets API: statuses/sample
Enterprise: Decahose
Conduct deep research backed by history.

The full archive of public Twitter data has over a decade of conversations around almost any topic you can imagine. Our batch process provides an economical way to filter significant amounts of data to find and analyze any public Tweet — back to the very first one in 2006.

Explore endpoints

Example functionality

Enterprise: Historical PowerTrack API
The Search API is rate limited to an $X$ amount of queries per a $Y$ minute window. According to multiple sources, the Streaming API (Sample API) appears to be limited to retrieve a maximum of 1% of the total amount of tweets published at that moment in time.
In a pilot study the two APIs were compared and tweets containing specific keywords were collected in August 24-25, 2015.

http://140dev.com/free-twitter-api-source-code-library/
The Streaming API outperformed the Search API for topics that generated moderate amounts of tweets and the differences between the APIs continues to increase for more popular topics.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Search API</th>
<th>Streaming API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>~ 1.5 M</td>
<td>~ 3.5 M</td>
</tr>
<tr>
<td>#election2016</td>
<td>7,610</td>
<td>8,709</td>
</tr>
</tbody>
</table>
Hi Kim,

Looking into the back end, the rule sets were actually switched on my email to you. Happy had 2.5 million matching tweets and the election has 5,000.

Sorry for that confusion!

Best,

Adam

<table>
<thead>
<tr>
<th></th>
<th>Search API</th>
<th>Streaming API</th>
<th>Twitter / GNIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>~ 1.5 M</td>
<td>~ 3.5 M</td>
<td>~ 2.5 M</td>
</tr>
<tr>
<td>#election2016</td>
<td>7,610</td>
<td>8,709</td>
<td>5,000</td>
</tr>
</tbody>
</table>
“Twitter’s worldwide network directly interconnects with over 3,000 unique networks in many datacenters worldwide. ...all in order to be as close as possible to the customer.”
Methods v2.0

1. Set up VPN Proxy Service
2. Create 6 Twitter Developer Apps
3. Set up 12 MySQL database tables
4. Write Python Scripts (one per location and Twitter API selection -> 12 Python scripts)
5. Set up CRON JOB on Ubuntu 16x Server to run all 12 Python scripts at same time
6. Let the tweets flow in...

VPN Location
- Brazil
- China
- Germany
- Great Britain
- Spain
- USA

Twitter API
# Add proxy server
class ProxyStream(tweepy.Stream):
    
    def new_session(self):
        super().new_session()
        self.session.proxies = {
            'https': SUPER_PROXY_URL,
            'http': SUPER_PROXY_URL
        }

CLIENT_ARGS = {
    'headers': {
        'User-Agent': 'Mozilla/5.0 (Windows NT 5.1; rv:31.0) Gecko/20100101 Firefox/31.0'
    },
    'proxies': {
        'http': SUPER_PROXY_URL,
        'https': SUPER_PROXY_URL,
    }
}
SEARCH_QUERY/FILTER = 'happy OR sad OR election2018 OR openscience OR avointiede'

9/16/18  15:50:10
9/16/18  20:39:35
4:49:25 time window

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>UNIQUE # in time frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>528,519</td>
</tr>
<tr>
<td>China</td>
<td>528,528</td>
</tr>
<tr>
<td>Germany</td>
<td>528,534</td>
</tr>
<tr>
<td>Spain</td>
<td>528,541</td>
</tr>
<tr>
<td>Great Britain</td>
<td>530,079</td>
</tr>
<tr>
<td>USA</td>
<td>528,535</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>UNIQUE # in time frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>322,701</td>
</tr>
<tr>
<td>China</td>
<td>324,097</td>
</tr>
<tr>
<td>Germany</td>
<td>0</td>
</tr>
<tr>
<td>Spain</td>
<td>323,992</td>
</tr>
<tr>
<td>Great Britain</td>
<td>323,997</td>
</tr>
<tr>
<td>USA</td>
<td>323,994</td>
</tr>
</tbody>
</table>
## Results v2.0

Compare tweet IDs across set of collected data in each category.

**STREAMING API**

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>UNIQUE TWEETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>2</td>
</tr>
<tr>
<td>China</td>
<td>8</td>
</tr>
<tr>
<td>Germany</td>
<td>12</td>
</tr>
<tr>
<td>Spain</td>
<td>1</td>
</tr>
<tr>
<td>Great Britain</td>
<td>1,539</td>
</tr>
<tr>
<td>USA</td>
<td>3</td>
</tr>
</tbody>
</table>

**REST API**

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>UNIQUE TWEETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil</td>
<td>50,440</td>
</tr>
<tr>
<td>China</td>
<td>304</td>
</tr>
<tr>
<td>Germany</td>
<td>0</td>
</tr>
<tr>
<td>Spain</td>
<td>82</td>
</tr>
<tr>
<td>Great Britain</td>
<td>1,255</td>
</tr>
<tr>
<td>USA</td>
<td>191</td>
</tr>
</tbody>
</table>
Version 3.0

- Partner with colleagues across the world with server and database access to run scripts and collect data
- Create/Collect additional Twitter developer apps
- Run each query separately
- Compare Results

- Please let us know if you would like to be part of our study!
Thank you for your attention

Kim Holmberg
kim.j.holmberg@utu.fi
http://kimholmberg.fi
@kholmber

Timothy D Bowman
timothy.d.bowman@wayne.edu
http://tdbowman.com
@timothydbowman