

Social Media and Web APIs

Indira Sen



Types of data

- Texts
- Images
- Videos
- Mixed formats
- User profiles (with metadata)
- Connections I (friends, followers)
- Connections II (links/URLs)
- Connections/Actions (likes, favs, comments, downloads)

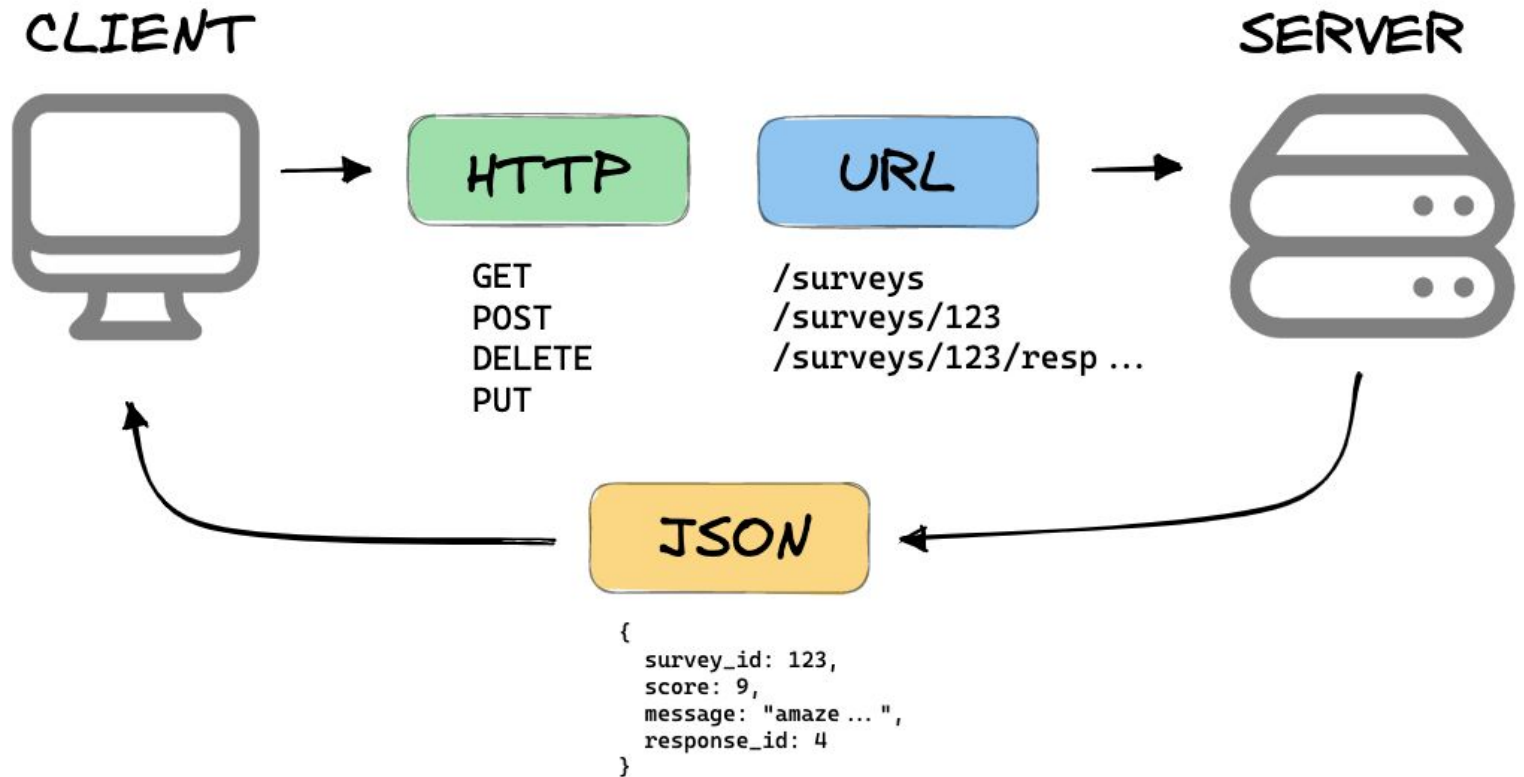
What are the various ways we can obtain social media data?

- APIs (data collection) → [hands-on example](#)
- Tools for data collection (often based on APIs)
- Web scraping
- Reuse existing datasets (data archives)
- Data donations (crowdsourcing)
- [Official resellers (buy data)]
- Use 3rd party services (e.g. Pushsift, Crimson Hexagon)

APIs

- Application Programming Interface
- Typically not designed for research purposes (therefore not tailored to research needs)
- Several platforms offer APIs as access points
- Often you have to register, sometimes also accept terms of services

WHAT IS A REST API?



mannhowie.com

Example: Twitter

- Different access options for different purposes:
 - ~~Twitter Academic API:~~
<https://developer.twitter.com/en/products/twitter-api/academic-research>
 - ~~Twitter Developer:~~ <https://developer.twitter.com/>
 - ~~premium "Firehose" access. See also Twitter Enterprise:~~
<https://developer.twitter.com/en/enterprise>
 - Changes to APIs policies over time (functionalities, user agreements)
 - Limitations on volume and functions
- IMPORTANT: – Free APIs cover 7 days tweets; Premium APIs exist for 30-day search and full archive search.

More examples: APIs

- Facebook for Developers:
<https://developers.facebook.com/>
- Facebook Ads API:
<https://developers.facebook.com/docs/marketing-apis/>
- Instagram Developer:
<https://www.instagram.com/developer/>
- YouTube Developers:
<https://developers.google.com/youtube/>
- Weibo API:
<http://open.weibo.com/wiki/API%E6%96%87%E6%A1%A3/en>
- CrowdTangle: <https://www.crowdtangle.com/request>

More examples: APIs and Data

- Wikipedia:
<https://meta.wikimedia.org/wiki/Research:Data>
- Tiktok: <https://github.com/dfreelon/pyktok>
- 4chan: <https://github.com/4chan/4chan-API>
- Gab: <https://github.com/a-tal/gab>
- Github: <https://developer.github.com/v3/>
- Stackoverflow: <https://api.stackexchange.com/docs>
- Locating or Requesting Social Media Data
- ProgrammableWeb:
<https://www.programmableweb.com/>
- Precollected datasets:
 - <https://datasetsearch.research.google.com/>
 - <https://www.kaggle.com/datasets>

APIs are forever changing!

Computational research in the post-API age

Deen Freelon

University of North Carolina at Chapel Hill

Forthcoming in *Political Communication*

Keywords: API, computational, Facebook, Twitter, social media

2018-08-20

On April 4, 2018, the post-API age reached a milestone. On that day, Facebook closed access to its Pages API, which had allowed researchers to extract all posts, comments, and associated metadata from public Facebook pages (Schroepfer, 2018). This decision followed the company's April 2015 closure of its public search API, which provided searchable access to all public posts within a rolling two-week window (Facebook, n.d.). The closure of the Pages API eliminated all terms of service (TOS)-compliant

APIs are forever changing!

- For example, Facebook completely closed down many of its APIs and it is now very hard to get Facebook data besides CrowdTangle or FB Ads
- Twitter's API has undergone substantial changes over time
- Conclusion: Have to stay vigilant and continuously update our code to keep up with the APIs

Ephemerality / Social media as a moving target

- Evolution of platforms
- Evolution of user behaviour
- Changes in profile data (incl. user names)
- Ongoing conversations
- Data loss: deleted accounts / posts

hands-on examples

Hands-on Data Collection: Reddit API

- What is Reddit?
 - Reddit: noun; a type of online community where users vote on content
- Created by Steve Huffman and Alexis Ohanian in June 2005
- Subreddit (an independent Reddit community within the greater Reddit community)

Hands-on Data Collection: Reddit API

What is Reddit?

- Reddit is an open platform where the community can post anything
- Redditors vote on stories—up or down which leads to scores for posts and karma for users
- Anyone can post links or open discussions on Reddit
- Comments can be made on each post
- ‘Flairs’ are similar to tags / topics

Hands-on Data Collection: Reddit API

The screenshot shows the Reddit homepage in a web browser. The browser's address bar displays <https://www.reddit.com>. The Reddit logo and navigation links (Home, Search, etc.) are visible at the top. The main content area features three posts:

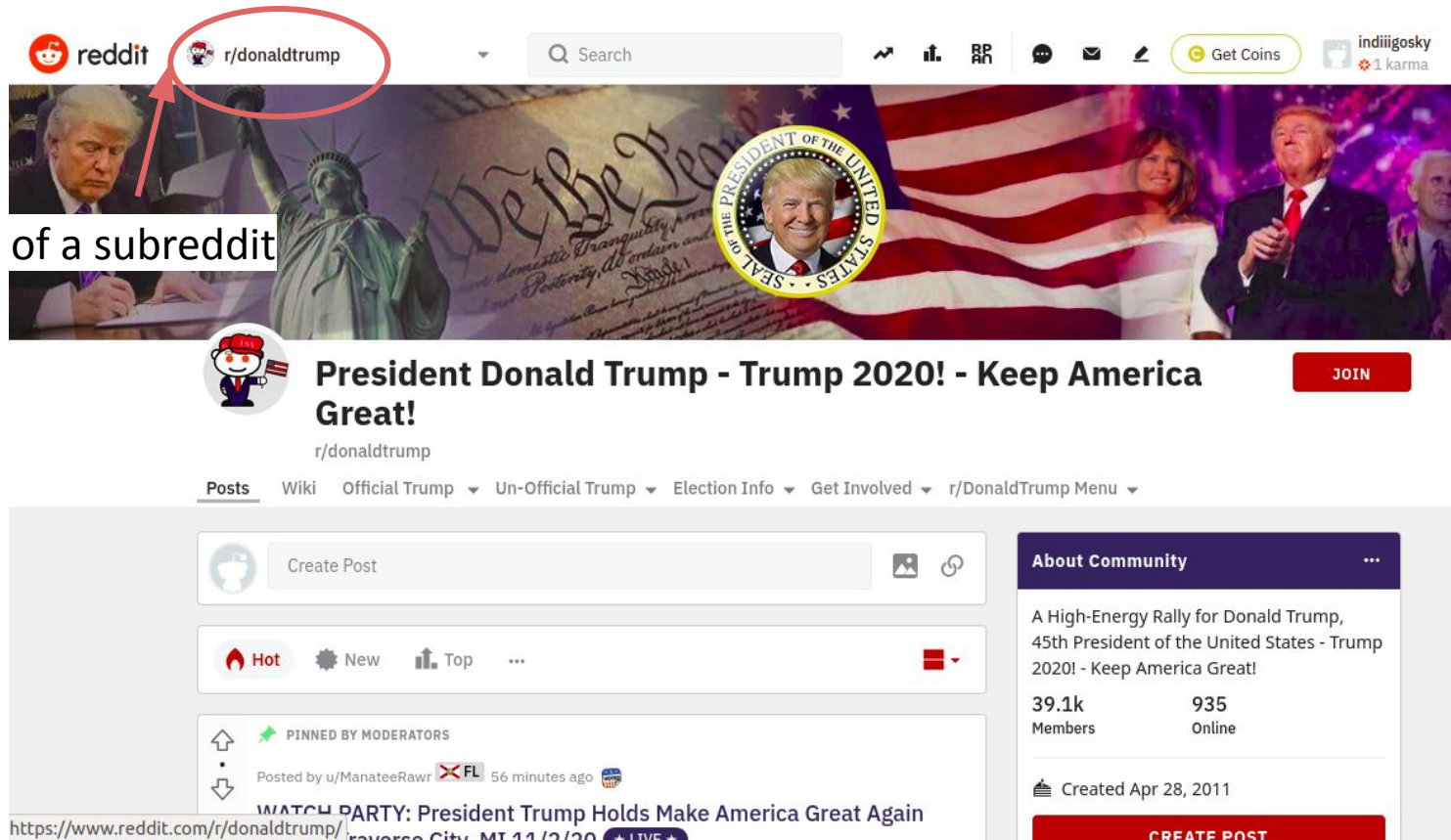
- Post 1:** **r/learnpython** - Posted by u/maze_runner_666 12 hours ago. Title: **Code Documentation**. Content: "I am a python beginner and I am quite fluent with all the basics and have even made some projects in python. I want to understand what 'code documentation' mean? I have heard people talk about it but I never understood what it mean?". 10 Comments, Give Award, Share, Save.
- Post 2:** **r/science** - Posted by u/mvea 10 hours ago. Title: **Wearing a 3-layer nonmedical face mask was not associated with a decline in oxygen saturation in older participants. The pooled mean Spo2 was 96.1% before, 96.5% while, and 96.3% after wearing the mask.** (Medicine). Source: jamanetwork.com/journa.... 1.9k Comments, Give Award, Share, Save. A table of data is visible in the post content.
- Post 3:** **r/3Blue1Brown** - Posted by u/siedzepodkocem 10 hours ago. Title: **weird thing about natural numbers to the power of 2**. Content: "I don't know if this exists somewhere on the internet, but recently I noticed that if you have a natural number (let's call it x) and the next number on the numberline (x+1) let's take 11 and 12 for example, if we raise them to the power of 2, and subtract those". 41 Comments, Give Award, Share, Save.

An advertisement for WorldRemit is displayed on the right side of the page, featuring a man using a smartphone and the text "Send money with us in minutes." and "Check our rates".

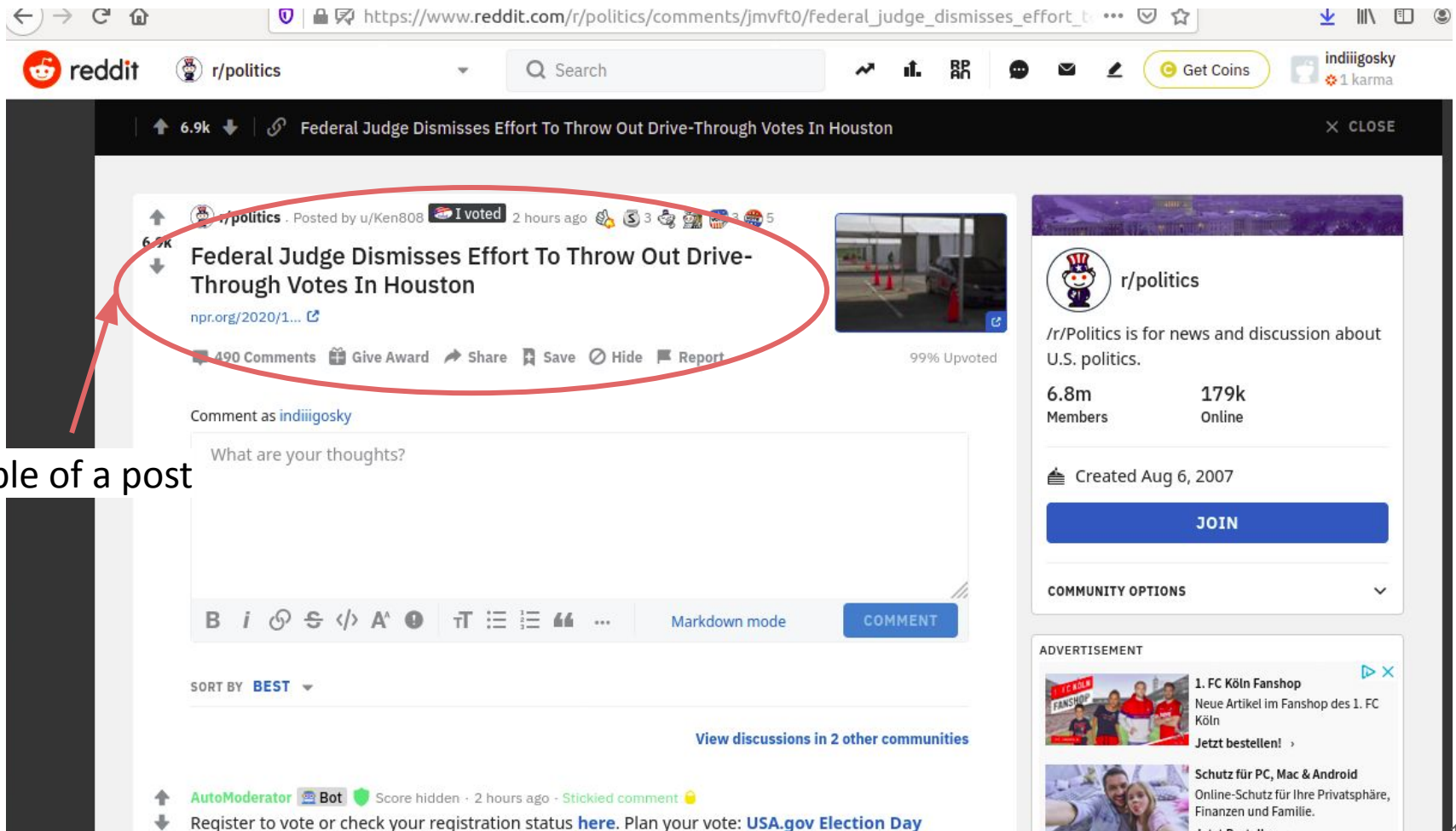
At the bottom right, there is a "BACK TO TOP" button.

Hands-on Data Collection: Reddit API

Example of a subreddit



Hands-on Data Collection: Reddit API



Example of a post

What is pushshift.io?

- Built by Jason Baumgartner, stores many different sources of data
- Especially famous for Reddit data
- <https://pushshift.io/>

The Pushshift Reddit Dataset

Jason Baumgartner,^{1,*} Savvas Zannettou,^{2,⊙} Brian Keegan,³ Megan Squire,⁴ Jeremy Blackburn^{5,⊙}
¹Pushshift.io, ²Max-Planck-Institut für Informatik, ³University of Colorado Boulder, ⁴Elon University, ⁵Binghamton University

^{*}Network Contagion Research Institute, [⊙]iDRAMA Lab

jason@pushshift.io, szannett@mpi-inf.mpg.de, brian.keegan@colorado.edu, msquire@elon.edu,
blackburn@cs.binghamton.edu

Abstract

Social media data has become crucial to the advancement of scientific understanding. However, even though it has become ubiquitous, just collecting large-scale social media data involves a high degree of engineering skill set and computational resources. In fact, research is often times gated by data engineering problems that must be overcome before analysis can proceed. This has resulted in recognition of datasets as meaningful research contributions in and of themselves.

Reddit, the so called “front page of the Internet,” in particular has been the subject of numerous scientific studies. Although Reddit is relatively open to data acquisition compared to social media platforms like Facebook and Twitter, the technical barriers to acquisition still remain. Thus, Reddit’s millions of subreddits, hundreds of millions of users, and billions of

crisis informatics (Palen and Anderson 2016). But following major scandals around data privacy and ethics, social media platforms like Facebook and Twitter changed previously permissive data access provisions of their public APIs (Walker, Mercea, and Bastos 2019). As a consequence, the ability for researchers to collect timely data, share tools, instruct students, and reproduce findings has been curtailed.

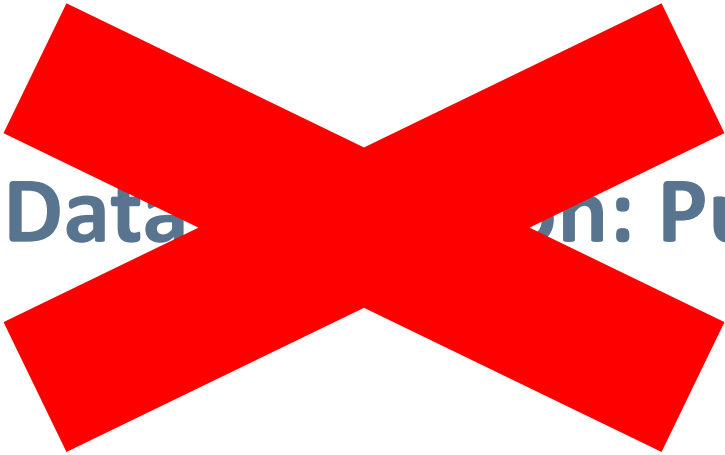
This “post-API age” is characterized by the deprecation of data resources used for research and teaching (Freelon 2018; Puschmann 2019), increased stratification of data access based on social, technical, and financial capital (boyd and Crawford 2012; Manovich 2011), and greater fear of prosecution around violating terms of service in the course of research (Halavais 2019; Patel 2018). These changes have had a profoundly chilling effect on researchers’ use of

Hands-on Data Collection: Pushshift.io

- What kind of data can we get?
 - submissions, comments, user histories
- What can't we get?
 - deleted content, views
- API Endpoints:
 - `/reddit/search/comment`
 - `/reddit/search/submission`

Hands-on Data Collection: Pushshift.io

Hands-on Data Science: Pushshift.io



WHERE ARE THEY NOW? —

Exploring Reddit's third-party app environment 7 months after the APIcalypse

Apollo dev: "I don't believe Reddit's leadership... cares about developers anymore."

SCHARON HARDING - 2/1/2024, 1:30 PM

Hands-on Data Collection: Reddit API

Alternative ways of getting Reddit data:

- Official Reddit API: <https://www.reddit.com/dev/api/>
- Google BigQuery: <https://cloud.google.com/bigquery>
- scraping Reddit data with GBQ:
 - https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments
 - <https://towardsdatascience.com/scrape-reddit-data-using-python-and-google-bigquery-44180b579892>

More on Social Media data collection and data quality:

- Nicola Osborne, Univ. of Edinburgh: Working with Social Media Data: Ethics & good practice around collecting, using and storing data
<https://www.slideshare.net/suchprettyeyes/working-with-social-media-data-ethics-good-practice-around-collecting-using-and-storing-data>
- Roberto Ulloa, GESIS: Introduction to Online Data Acquisition
<https://www.youtube.com/watch?v=inUvEFLG5EA>