gesis Leibniz Institute for the Social Sciences

# Potentials and Pitfalls of Social Media Data

*Indira Sen & Katrin Weller*

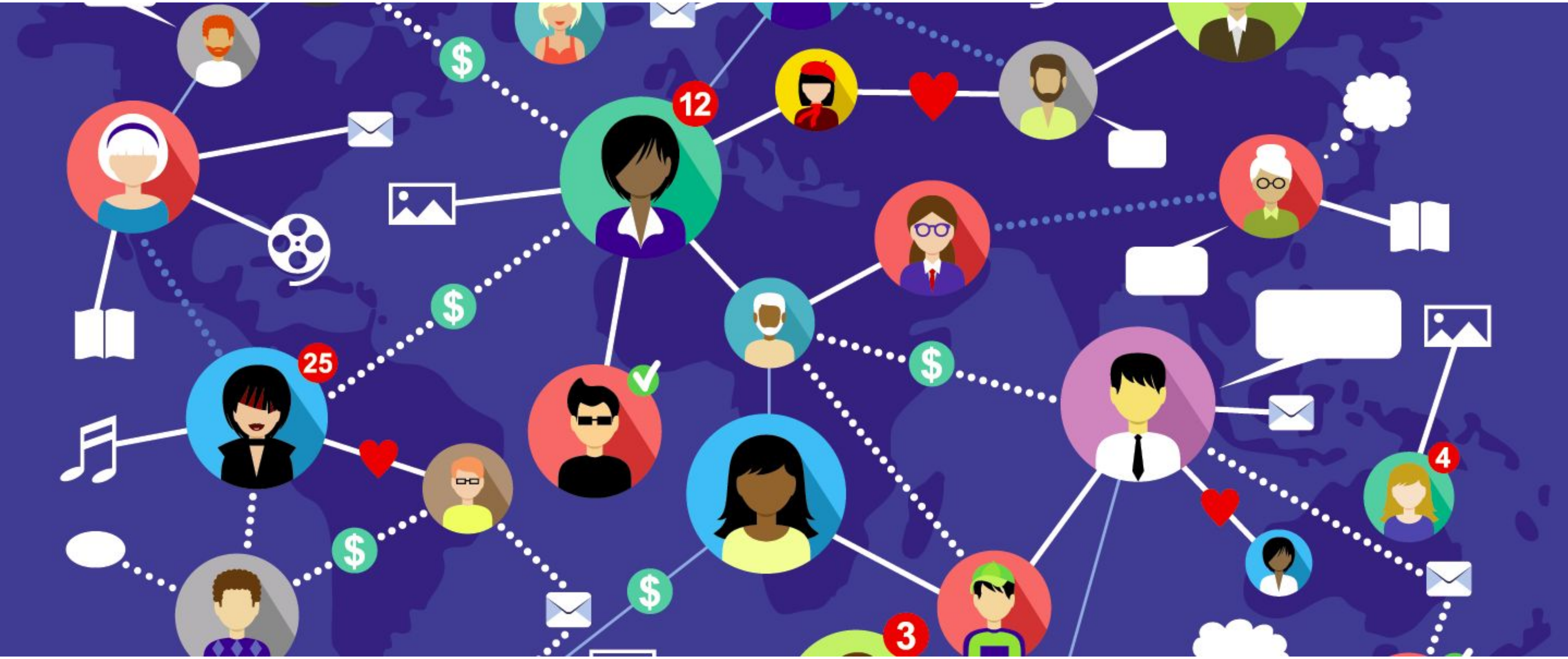
GESIS workshop - December 2022

Leibniz Association

# Agenda

- Session 1: Introduction to Research with Social Media Data (SMD)
- Session 2: SM Data Collection
- Session 3: SMD Preprocessing and Analysis
- Session 4: Potential Pitfalls of SMD
- Session 5: Identifying Pitfalls with help from surveys
- Session 6: Identifying Pitfalls in SMD
- Session 7: Mitigating Pitfalls
- Session 8: Documenting Pitfalls
- Session 9: Recap and Conclusions

# Data collection


12
25
3
4

# 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

# 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.

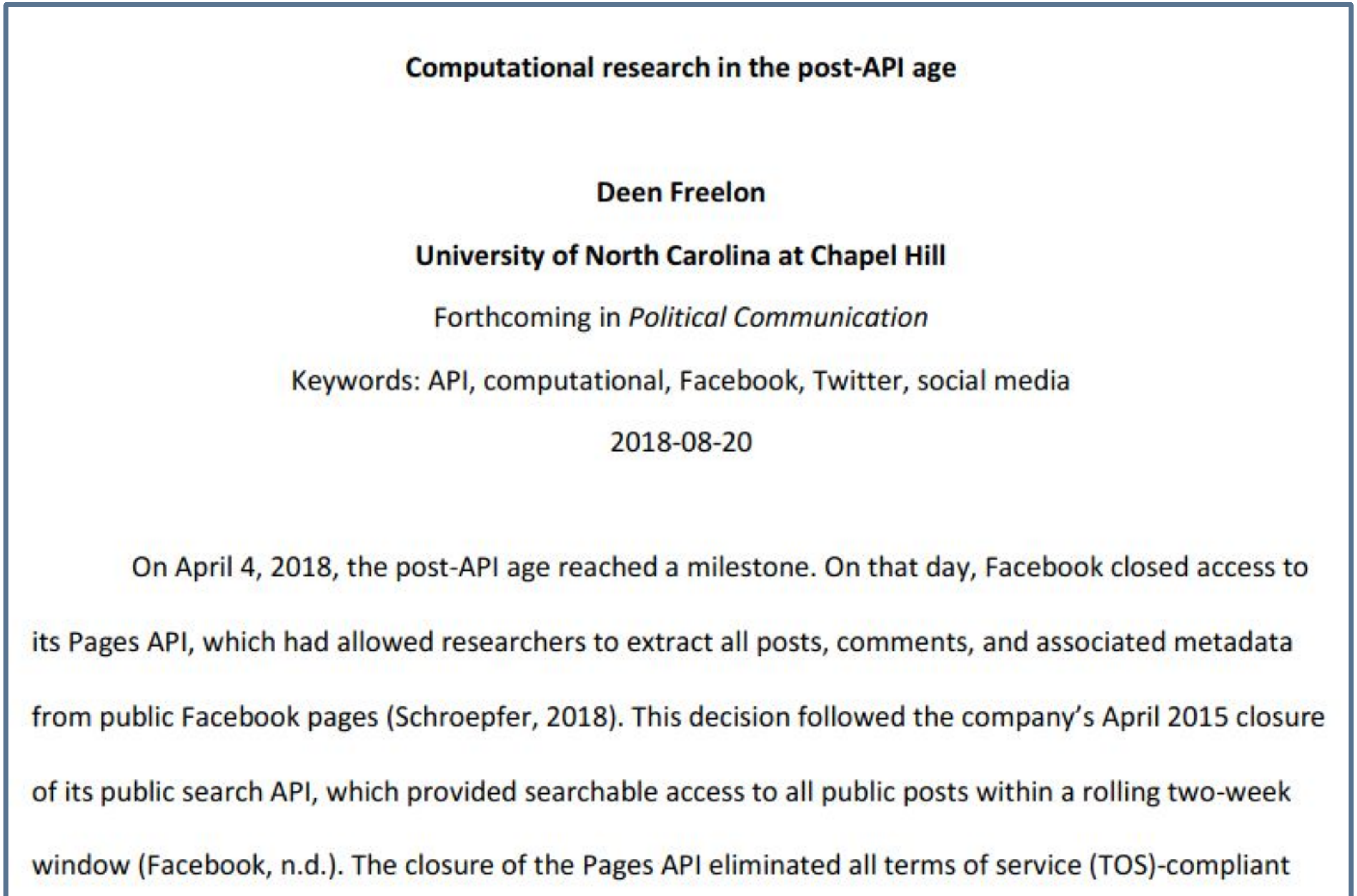
# More examples: APIs

- 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

- 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 it’s 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

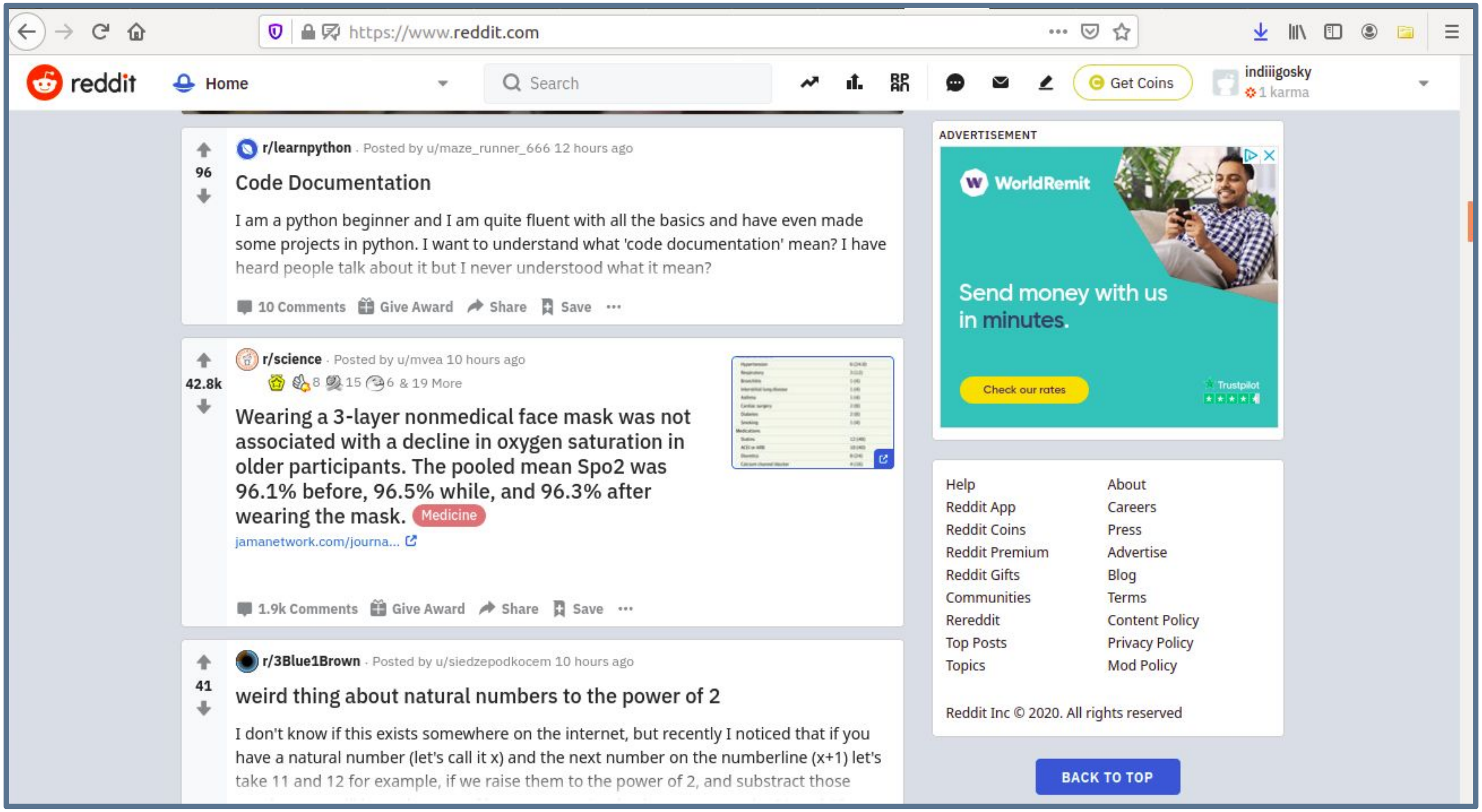
- What is Reddit?
  - Reddit: noun; a type of online community where users vote on content
- Created by Steve Huffman and Alexis Ohanian in June 2005
- Subreddit (an independent Reddit community within the greater Reddit community)

# 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



https://www.reddit.com

96

r/learnpython · Posted by u/maze_runner_666 12 hours ago

**Code Documentation**

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 · ...

42.8k

r/science · Posted by u/mvea 10 hours ago

8 · 15 · 6 & 19 More

**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

jamanetwork.com/journa...

1.9k Comments · Give Award · Share · Save · ...

41

r/3Blue1Brown · Posted by u/siedzepodkocem 10 hours ago

**weird thing about natural numbers to the power of 2**

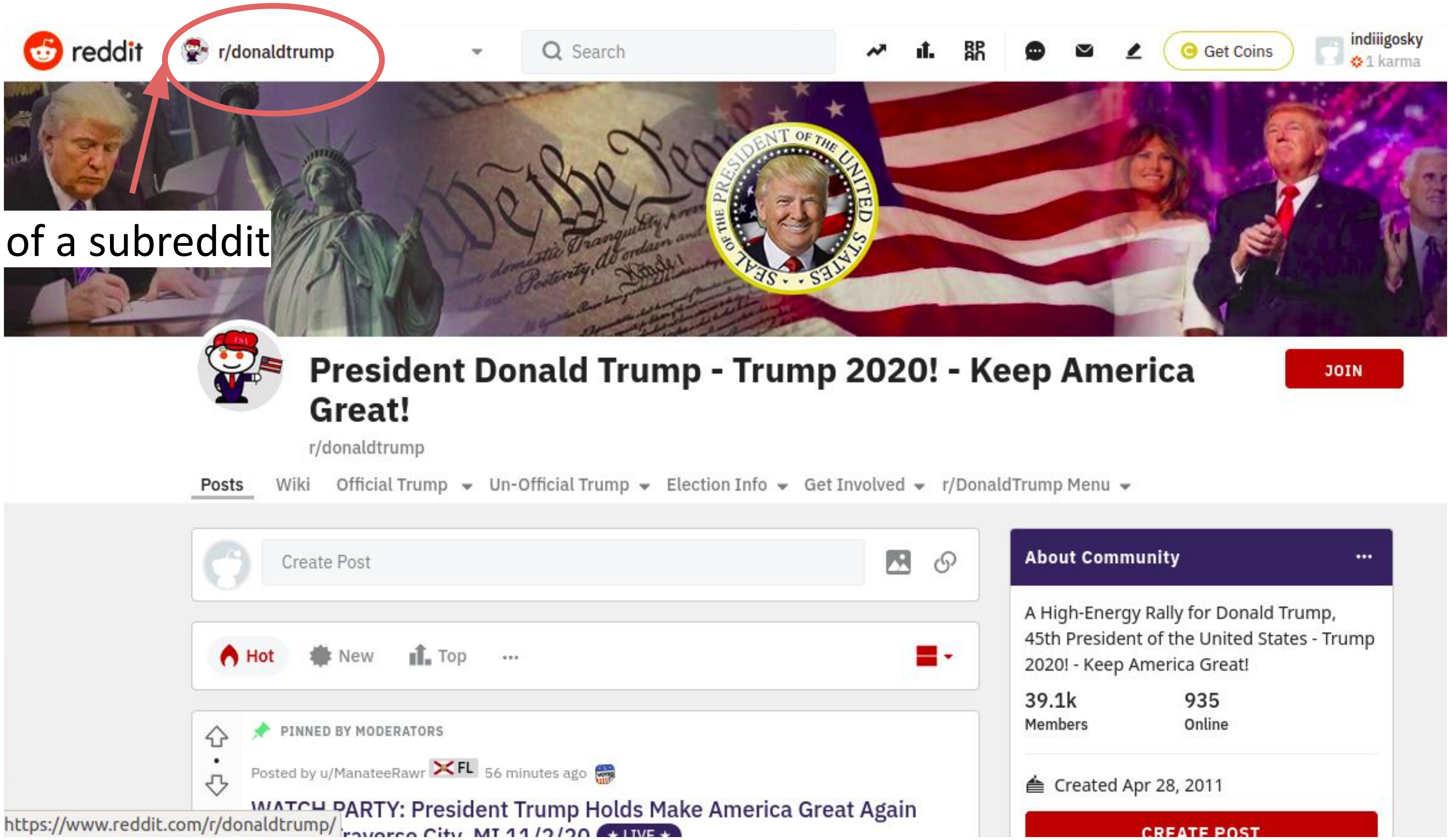
I don't know if this exists somewhere on the internet, but recently I noticed that if you have a natural number (let's call it x) and the next number on the numberline (x+1) let's take 11 and 12 for example, if we raise them to the power of 2, and substract those

| | |
|---|---|
| Help | About |
| Reddit App | Careers |
| Reddit Coins | Press |
| Reddit Premium | Advertise |
| Reddit Gifts | Blog |
| Communities | Terms |
| Rereddit | Content Policy |
| Top Posts | Privacy Policy |
| Topics | Mod Policy |

Reddit Inc © 2020. All rights reserved

BACK TO TOP

# Hands-on Data Collection: Reddit API



Example of a subreddit

reddit r/donaldtrump

Search

Get Coins

indiiigosky
1 karma

President Donald Trump - Trump 2020! - Keep America Great!

r/donaldtrump

JOIN

Posts Wiki Official Trump Un-Official Trump Election Info Get Involved r/DonaldTrump Menu

Create Post

Hot New Top ...

PINNED BY MODERATORS

Posted by u/ManateeRawr FL 56 minutes ago

WATCH PARTY: President Trump Holds Make America Great Again Traverse City, MI 11/2/20 LIVE

https://www.reddit.com/r/donaldtrump/

About Community

A High-Energy Rally for Donald Trump, 45th President of the United States - Trump 2020! - Keep America Great!

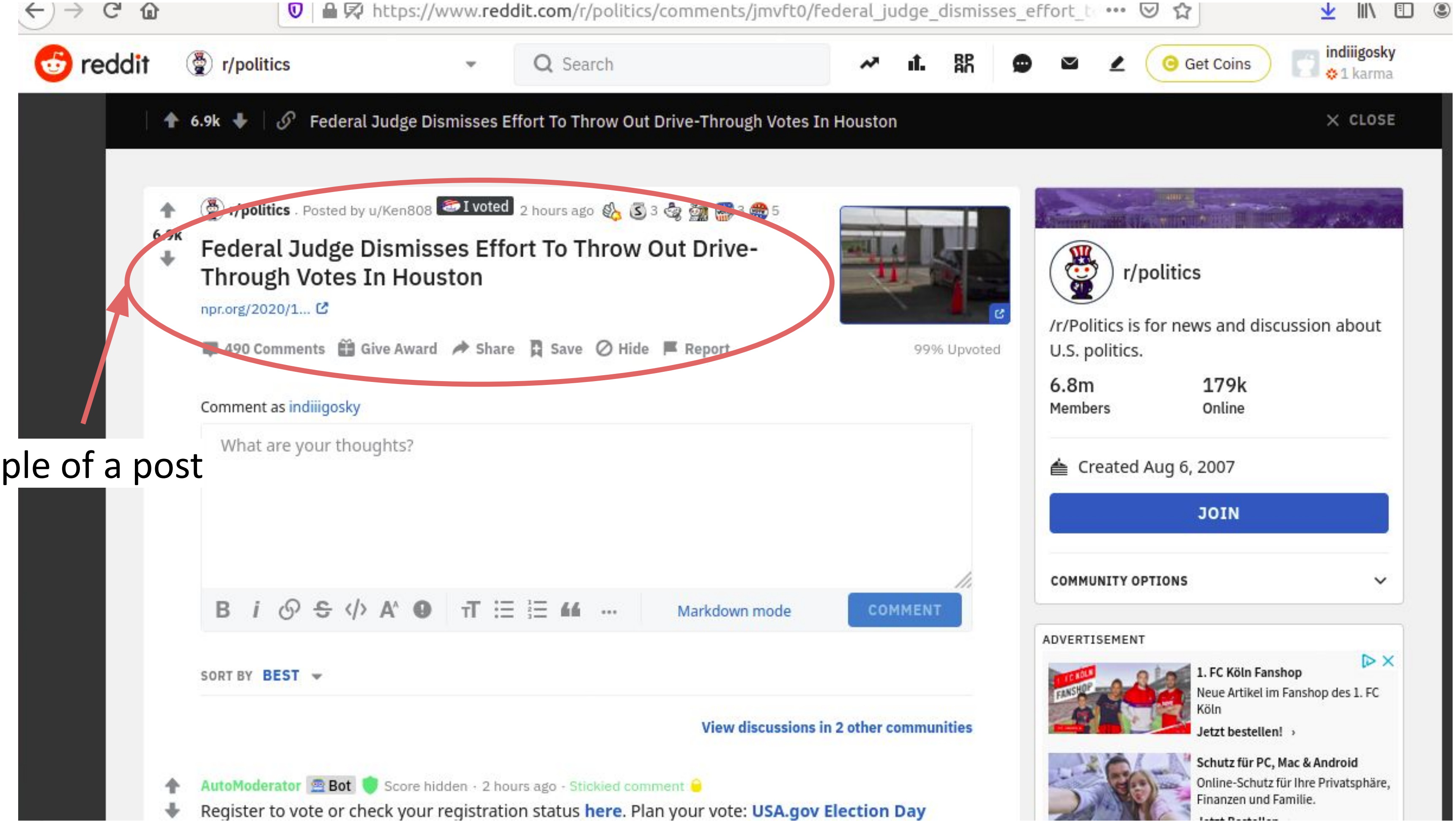
39.1k Members

935 Online

Created Apr 28, 2011

CREATE POST

# Hands-on Data Collection: Reddit API



Example of a post

https://www.reddit.com/r/politics/comments/jmvft0/federal_judge_dismisses_effort_t

Federal Judge Dismisses Effort To Throw Out Drive-Through Votes In Houston

r/politics · Posted by u/Ken808 I voted 2 hours ago 3 3 5

Federal Judge Dismisses Effort To Throw Out Drive-Through Votes In Houston

npr.org/2020/1...

Comment as indiiigosky

What are your thoughts?

View discussions in 2 other communities

AutoModerator Bot Score hidden · 2 hours ago · Stickied comment

Register to vote or check your registration status here. Plan your vote: USA.gov Election Day

r/politics

/r/Politics is for news and discussion about U.S. politics.

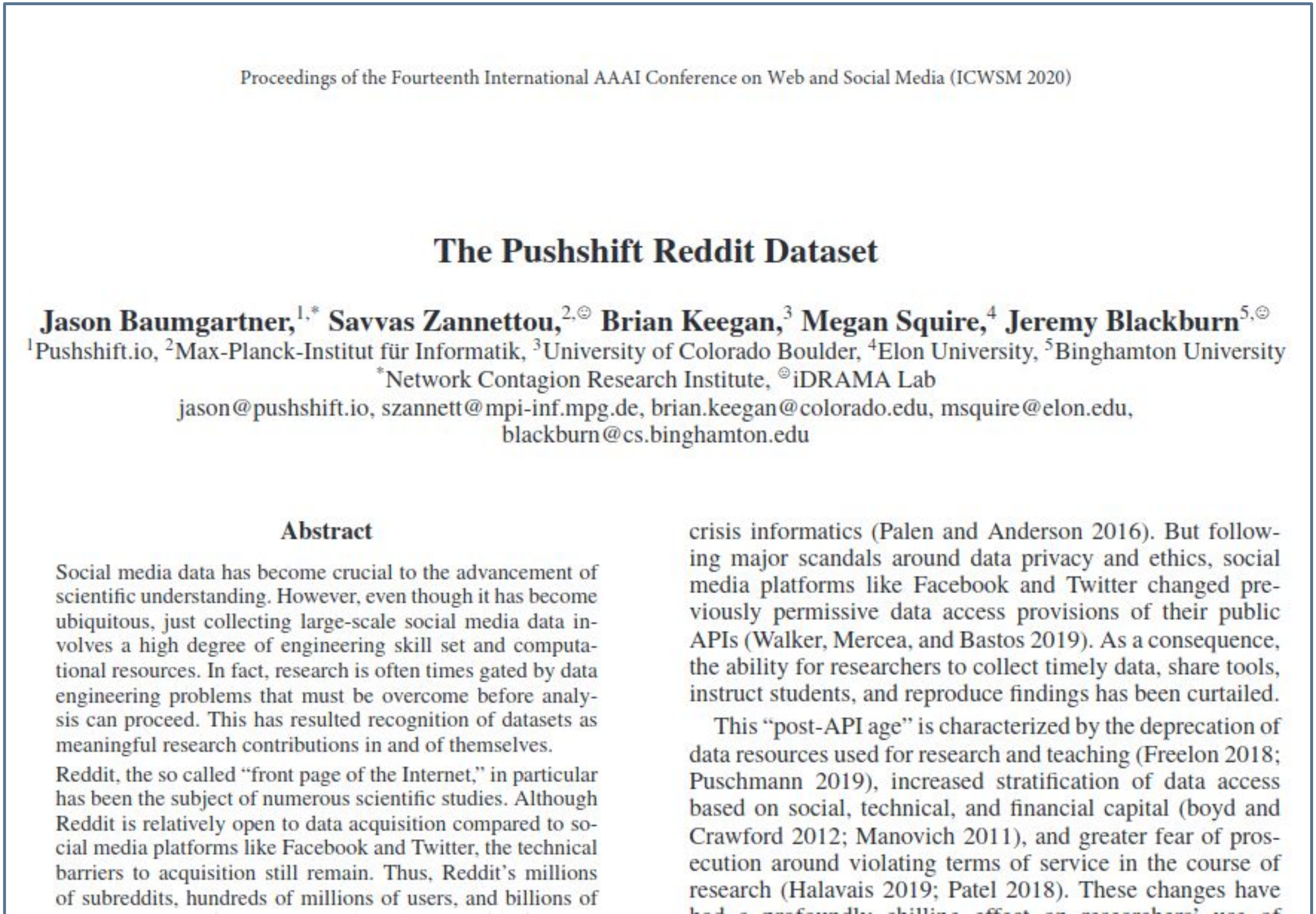
| 6.8m | 179k |
|---|---|
| Members | Online |

Created Aug 6, 2007

# **Research Question:** What if we want to predict the 2024 US election with Reddit data?

# What is pushshift.io?

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



Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

# The Pushshift Reddit Dataset

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

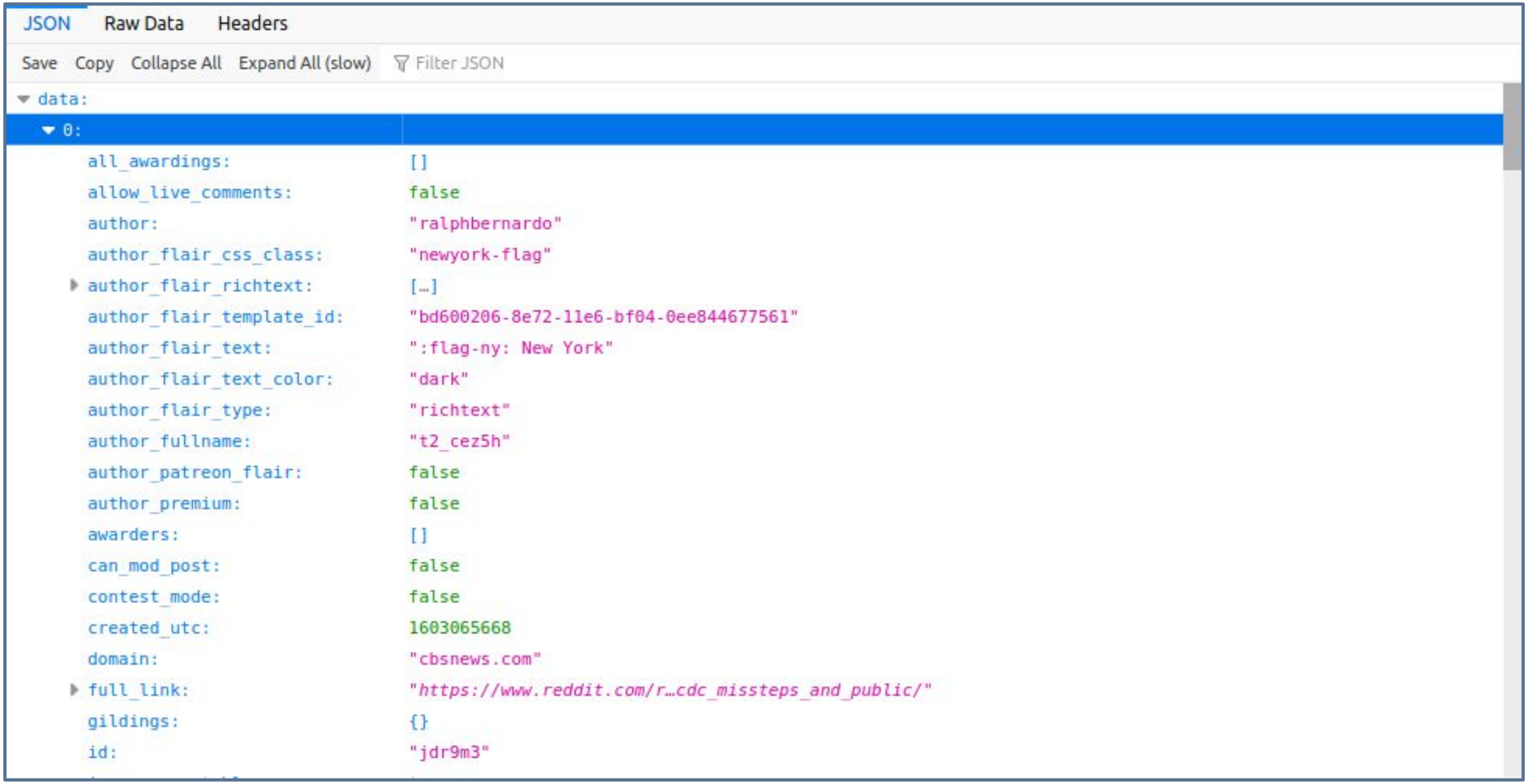
# Hands-on Data Collection: Pushshift.io

- 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

# Try this link in your browser:

**https://api.pushshift.io/reddit/search/submission/?title=&size=1000&after=1603065600&subreddit=politics**



```
▼ data:
  ▼ 0:
      all_awardings:               []
      allow_live_comments:         false
      author:                      "ralphbernardo"
      author_flair_css_class:      "newyork-flag"
    ▶ author_flair_richtext:       […]
      author_flair_template_id:    "bd600206-8e72-11e6-bf04-0ee844677561"
      author_flair_text:           ":flag-ny: New York"
      author_flair_text_color:     "dark"
      author_flair_type:           "richtext"
      author_fullname:             "t2_cez5h"
      author_patreon_flair:        false
      author_premium:              false
      awarders:                    []
      can_mod_post:                false
      contest_mode:                false
      created_utc:                 1603065668
      domain:                      "cbsnews.com"
    ▶ full_link:                   "https://www.reddit.com/r…cdc_missteps_and_public/"
      gildings:                    {}
      id:                          "jdr9m3"
```

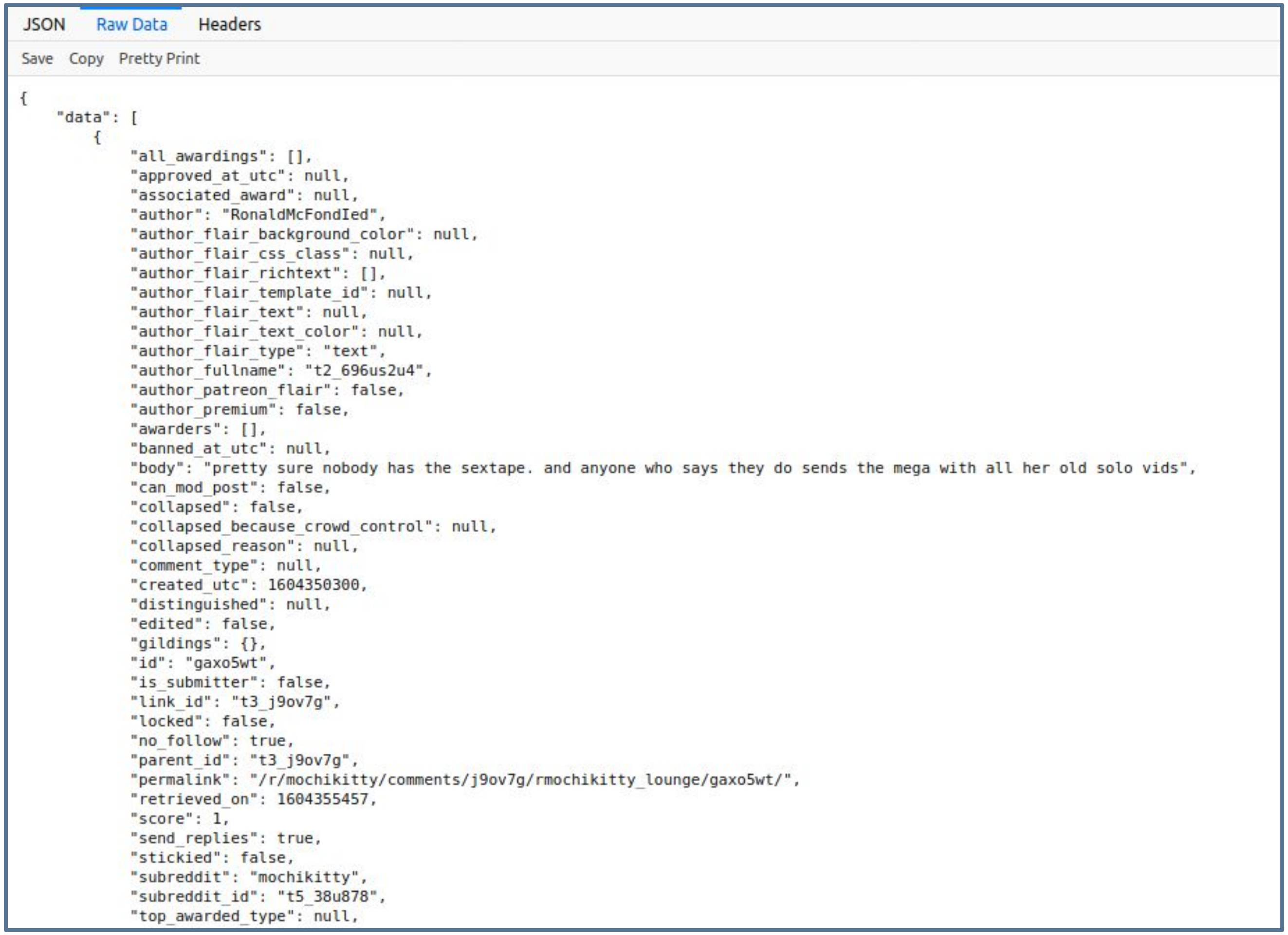
## Try this link in your browser:

[https://api.pushshift.io/reddit/search/submission/?title=&size=1000&after=1603065600&subreddit=politics](https://api.pushshift.io/reddit/search/submission/?title=&size=1000&after=1603065600&subreddit=politics)

*subreddit*

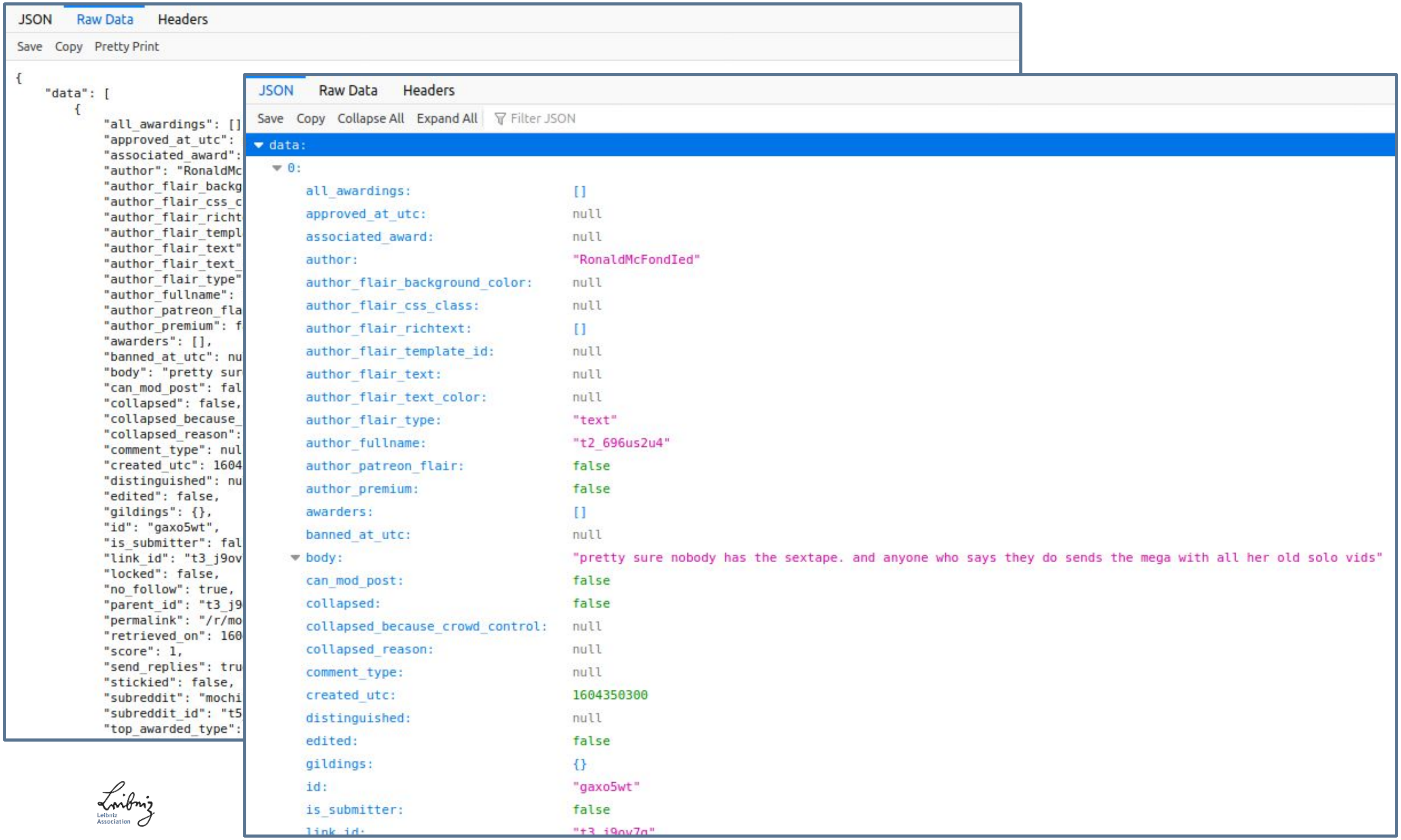
*endpoint for getting posts*

# What do we get from Pushshift.io?



```
{
    "data": [
        {
            "all_awardings": [],
            "approved_at_utc": null,
            "associated_award": null,
            "author": "RonaldMcFondIed",
            "author_flair_background_color": null,
            "author_flair_css_class": null,
            "author_flair_richtext": [],
            "author_flair_template_id": null,
            "author_flair_text": null,
            "author_flair_text_color": null,
            "author_flair_type": "text",
            "author_fullname": "t2_696us2u4",
            "author_patreon_flair": false,
            "author_premium": false,
            "awarders": [],
            "banned_at_utc": null,
            "body": "pretty sure nobody has the sextape. and anyone who says they do sends the mega with all her old solo vids",
            "can_mod_post": false,
            "collapsed": false,
            "collapsed_because_crowd_control": null,
            "collapsed_reason": null,
            "comment_type": null,
            "created_utc": 1604350300,
            "distinguished": null,
            "edited": false,
            "gildings": {},
            "id": "gaxo5wt",
            "is_submitter": false,
            "link_id": "t3_j9ov7g",
            "locked": false,
            "no_follow": true,
            "parent_id": "t3_j9ov7g",
            "permalink": "/r/mochikitty/comments/j9ov7g/rmochikitty_lounge/gaxo5wt/",
            "retrieved_on": 1604355457,
            "score": 1,
            "send_replies": true,
            "stickied": false,
            "subreddit": "mochikitty",
            "subreddit_id": "t5_38u878",
            "top_awarded_type": null,
```

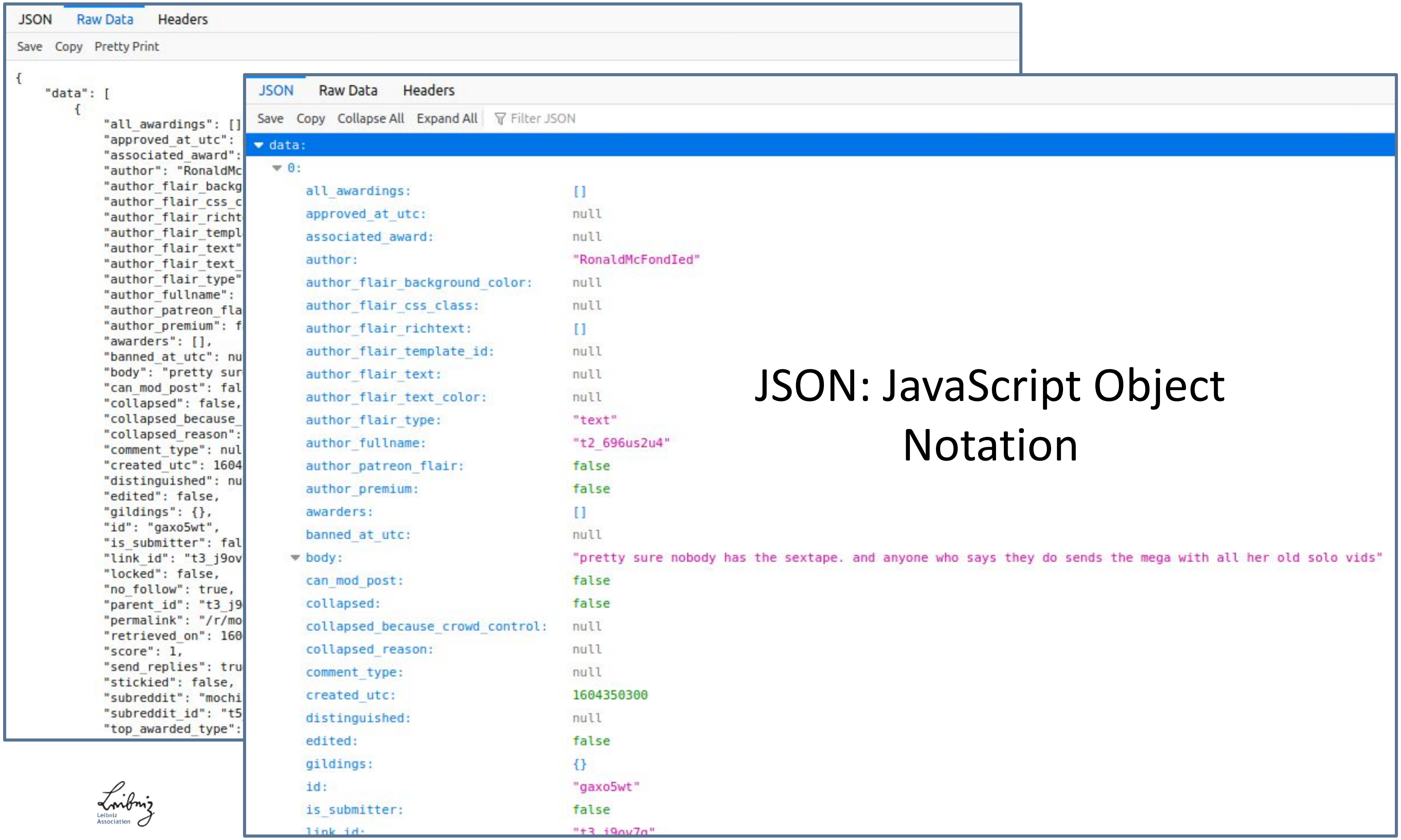
# What do we get from Pushshift.io?



```
{
    "data": [
        {
            "all_awardings": []
            "approved_at_utc":
            "associated_award":
            "author": "RonaldMc
            "author_flair_backg
            "author_flair_css_c
            "author_flair_richt
            "author_flair_templ
            "author_flair_text"
            "author_flair_text_
            "author_flair_type"
            "author_fullname":
            "author_patreon_fla
            "author_premium": f
            "awarders": [],
            "banned_at_utc": nu
            "body": "pretty sur
            "can_mod_post": fal
            "collapsed": false,
            "collapsed_because_
            "collapsed_reason":
            "comment_type": nul
            "created_utc": 1604
            "distinguished": nu
            "edited": false,
            "gildings": {},
            "id": "gaxo5wt",
            "is_submitter": fal
            "link_id": "t3_j9ov
            "locked": false,
            "no_follow": true,
            "parent_id": "t3_j9
            "permalink": "/r/mo
            "retrieved_on": 160
            "score": 1,
            "send_replies": tru
            "stickied": false,
            "subreddit": "mochi
            "subreddit_id": "t5
            "top_awarded_type":
```

```
▼ data:
  ▼ 0:
      all_awardings:                    []
      approved_at_utc:                  null
      associated_award:                 null
      author:                           "RonaldMcFondIed"
      author_flair_background_color:    null
      author_flair_css_class:           null
      author_flair_richtext:            []
      author_flair_template_id:         null
      author_flair_text:                null
      author_flair_text_color:          null
      author_flair_type:                "text"
      author_fullname:                  "t2_696us2u4"
      author_patreon_flair:             false
      author_premium:                   false
      awarders:                         []
      banned_at_utc:                    null
    ▼ body:                             "pretty sure nobody has the sextape. and anyone who says they do sends the mega with all her old solo vids"
      can_mod_post:                     false
      collapsed:                        false
      collapsed_because_crowd_control:  null
      collapsed_reason:                 null
      comment_type:                     null
      created_utc:                      1604350300
      distinguished:                    null
      edited:                           false
      gildings:                         {}
      id:                               "gaxo5wt"
      is_submitter:                     false
      link_id:                          "t3_j9ov7q"
```

# What do we get from Pushshift.io?



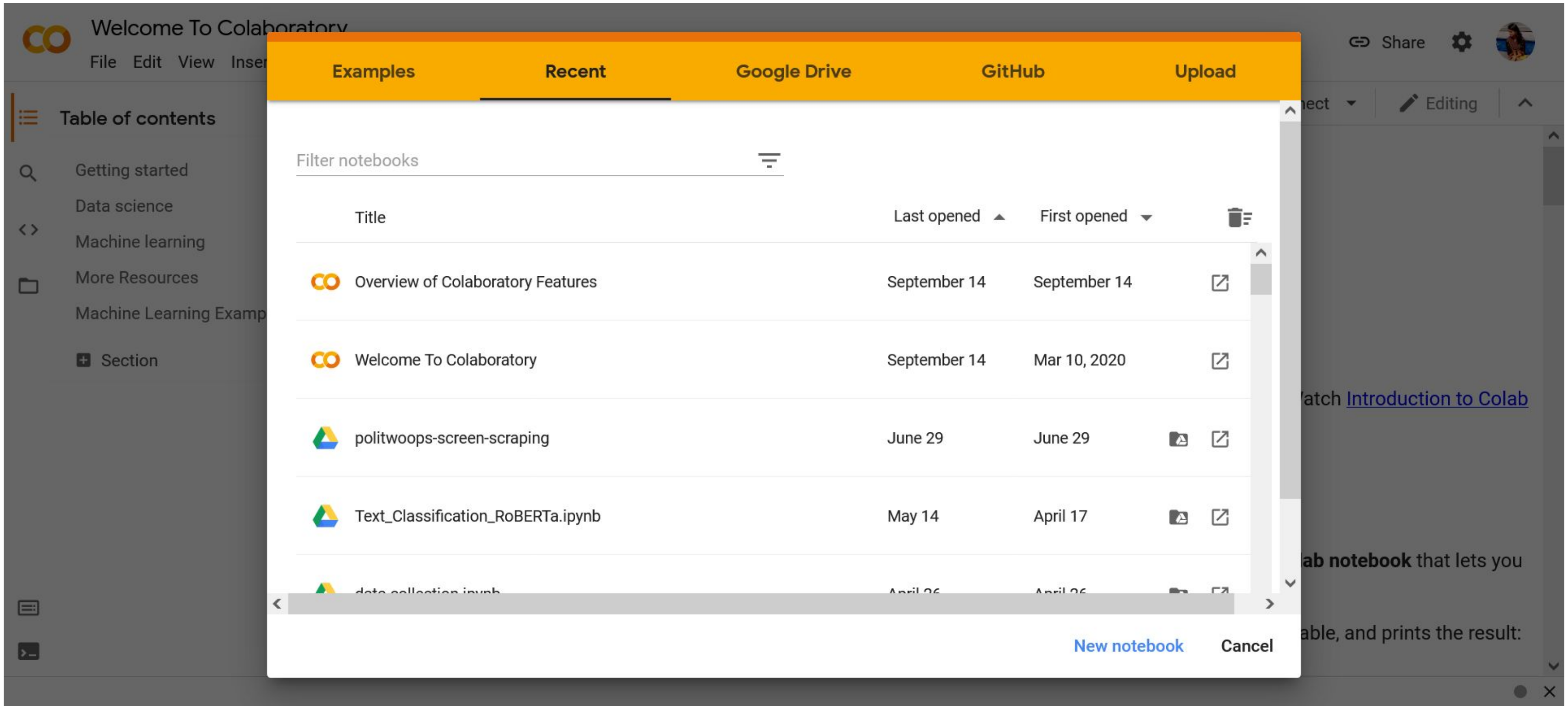
```
{
    "data": [
        {
            "all_awardings": []
            "approved_at_utc":
            "associated_award":
            "author": "RonaldMc
            "author_flair_backg
            "author_flair_css_c
            "author_flair_richt
            "author_flair_templ
            "author_flair_text"
            "author_flair_text_
            "author_flair_type"
            "author_fullname":
            "author_patreon_fla
            "author_premium": f
            "awarders": [],
            "banned_at_utc": nu
            "body": "pretty sur
            "can_mod_post": fal
            "collapsed": false,
            "collapsed_because_
            "collapsed_reason":
            "comment_type": nul
            "created_utc": 1604
            "distinguished": nu
            "edited": false,
            "gildings": {},
            "id": "gaxo5wt",
            "is_submitter": fal
            "link_id": "t3_j9ov
            "locked": false,
            "no_follow": true,
            "parent_id": "t3_j9
            "permalink": "/r/mo
            "retrieved_on": 160
            "score": 1,
            "send_replies": tru
            "stickied": false,
            "subreddit": "mochi
            "subreddit_id": "t5
            "top_awarded_type":
```

```
▼ data:
  ▼ 0:
      all_awardings:                        []
      approved_at_utc:                      null
      associated_award:                     null
      author:                               "RonaldMcFondIed"
      author_flair_background_color:        null
      author_flair_css_class:               null
      author_flair_richtext:                []
      author_flair_template_id:             null
      author_flair_text:                    null
      author_flair_text_color:              null
      author_flair_type:                    "text"
      author_fullname:                      "t2_696us2u4"
      author_patreon_flair:                 false
      author_premium:                       false
      awarders:                             []
      banned_at_utc:                        null
    ▼ body:                                 "pretty sure nobody has the sextape. and anyone who says they do sends the mega with all her old solo vids"
      can_mod_post:                         false
      collapsed:                            false
      collapsed_because_crowd_control:      null
      collapsed_reason:                     null
      comment_type:                         null
      created_utc:                          1604350300
      distinguished:                        null
      edited:                               false
      gildings:                             {}
      id:                                   "gaxo5wt"
      is_submitter:                         false
      link_id:                              "t3_j9ov7q"
```

JSON: JavaScript Object Notation

# Google Colab

- colab.research.google.com/



| Title | Last opened | First opened |
|---|---|---|
| Overview of Colaboratory Features | September 14 | September 14 |
| Welcome To Colaboratory | September 14 | Mar 10, 2020 |
| politwoops-screen-scraping | June 29 | June 29 |
| Text_Classification_RoBERTa.ipynb | May 14 | April 17 |
| data_collection.ipynb | April 26 | April 26 |

# 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