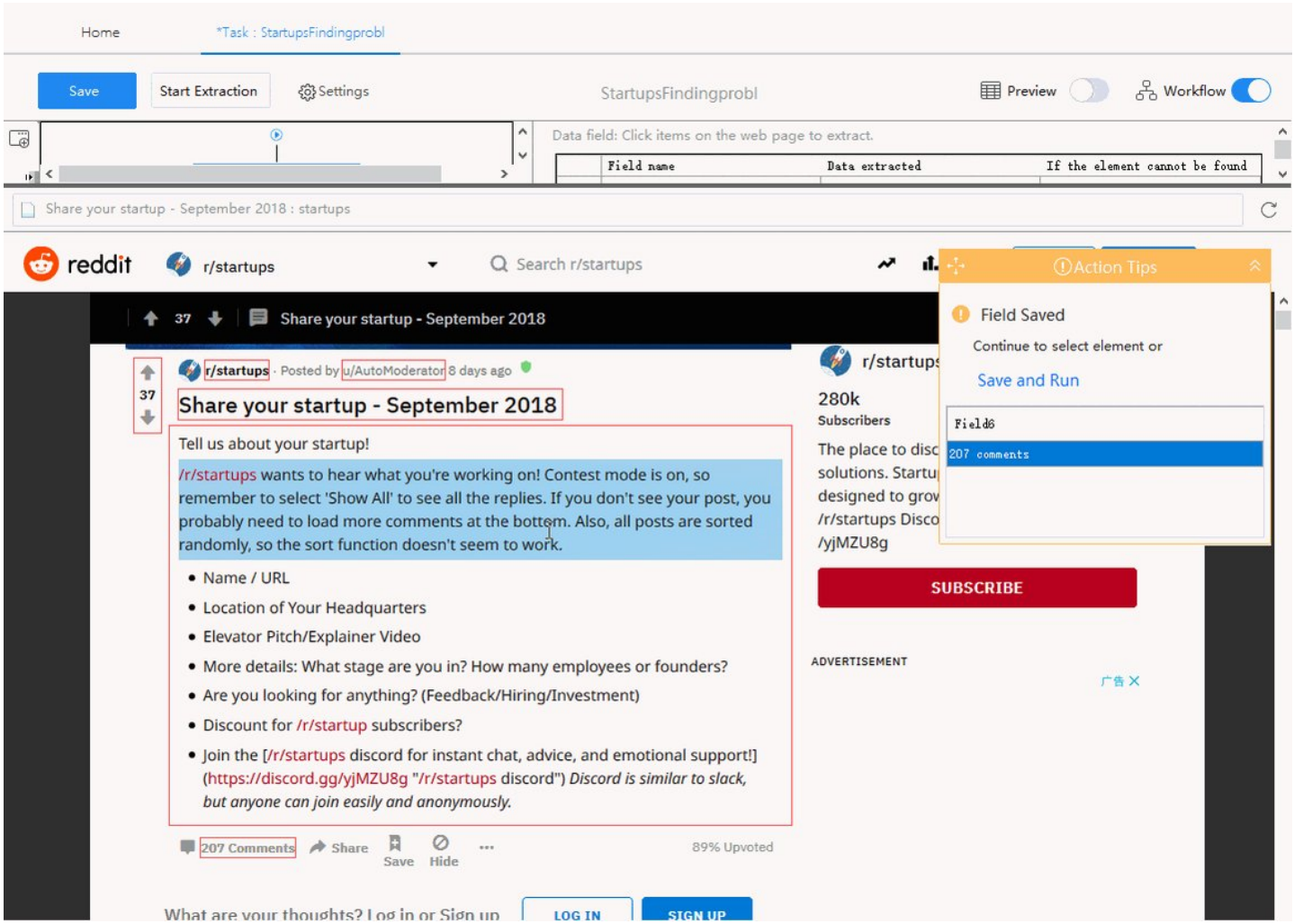
Twitter Thread by Philip Vollet

Philip Vollet
@philipvollet

Do you need social media data for your machine learning project?

- Twitter data?
- Reddit data?
- Facebook data?

Where to get it?



Reddit: Pushshift

Pushshift is a big-data storage and analytics project.

Most people know it for its copy of reddit comments and submissions.

https://t.co/ne0XqKIt9A

Reddit: Pushshift API

The https://t.co/HWZNWEvrxY Reddit API was designed and created by the /r/datasets mod team to help provide enhanced functionality and search capabilities for searching Reddit comments and submissions.

https://t.co/FkUE7R2jlb

Reddit: Pushshift file download

Note: The latest data for manual download is from April 2020

https://t.co/jBM4U71dnm

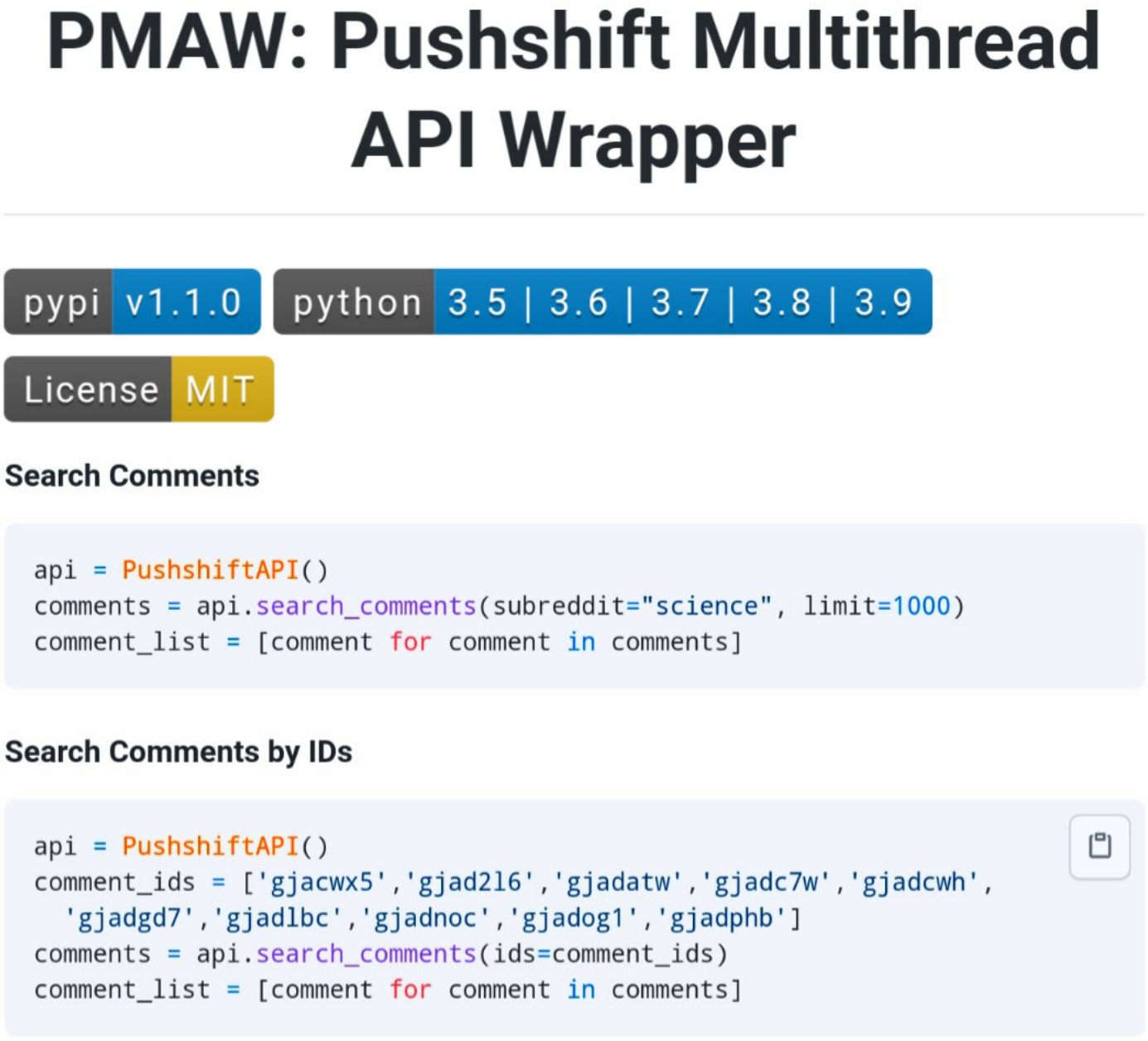
RC_2020-03-28.gz	Reddit Comments (JSON objects)	1,002,745,062	Aug 31 2020 11:13 PM
RC_2020-03-29.gz	Reddit Comments (JSON objects)	1,020,019,147	Aug 31 2020 11:13 PM
RC_2020-03-30.gz	Reddit Comments (JSON objects)	1,083,474,361	Aug 31 2020 11:13 PM
RC_2020-03-31.gz	Reddit Comments (JSON objects)	1,119,369,920	Aug 31 2020 11:13 PM
RC_2020-04-01.gz	Reddit Comments (JSON objects)	1,106,724,537	Aug 31 2020 11:15 PM
RC_2020-04-02.gz	Reddit Comments (JSON objects)	1,106,910,210	Aug 31 2020 11:15 PM
RC_2020-04-03.gz	Reddit Comments (JSON objects)	1,100,472,299	Aug 31 2020 11:15 PM
RC_2020-04-04.gz	Reddit Comments (JSON objects)	1,017,840,336	Aug 31 2020 11:15 PM
RC_2020-04-05.gz	Reddit Comments (JSON objects)	1,022,186,838	Aug 31 2020 11:16 PM
RC_2020-04-06.gz	Reddit Comments (JSON objects)	1,125,814,526	Aug 31 2020 11:16 PM
RC_2020-04-07.gz	Reddit Comments (JSON objects)	1,136,713,120	Aug 31 2020 11:16 PM
RC_2020-04-08.gz	Reddit Comments (JSON objects)	1,183,215,796	Aug 31 2020 11:16 PM
RC_2020-04-09.gz	Reddit Comments (JSON objects)	1,165,810,247	Aug 31 2020 11:17 PM
RC_2020-04-10.gz	Reddit Comments (JSON objects)	1,109,186,231	Aug 31 2020 11:17 PM
RC_2020-04-11.gz	Reddit Comments (JSON objects)	1,052,617,691	Aug 31 2020 11:18 PM
RC_2020-04-12.gz	Reddit Comments (JSON objects)	1,069,616,452	Aug 31 2020 11:40 PM
RC_2020-04-13.gz	Reddit Comments (JSON objects)	1,154,546,904	Aug 31 2020 11:41 PM
RC_2020-04-14.gz	Reddit Comments (JSON objects)	1,180,269,552	Aug 31 2020 11:41 PM
RC_2020-04-15.gz	Reddit Comments (JSON objects)	1,206,702,605	Aug 31 2020 11:41 PM
RC_2020-04-16.gz	Reddit Comments (JSON objects)	1,192,552,544	Sep 1 2020 10:08 AM
RC_2020-04-17.gz	Reddit Comments (JSON objects)	1,163,581,603	Sep 1 2020 10:09 AM
RC_2020-04-18.gz	Reddit Comments (JSON objects)	1,098,207,690	Sep 1 2020 10:09 AM

Reddit: PMAW: Pushshift Multithread API Wrapper

PMAW is an ultra minimalist wrapper for the Pushshift API which uses multithreading to retrieve Reddit comments and submissions.

If you pull data via Pushshift use PMAW, highly recommended!

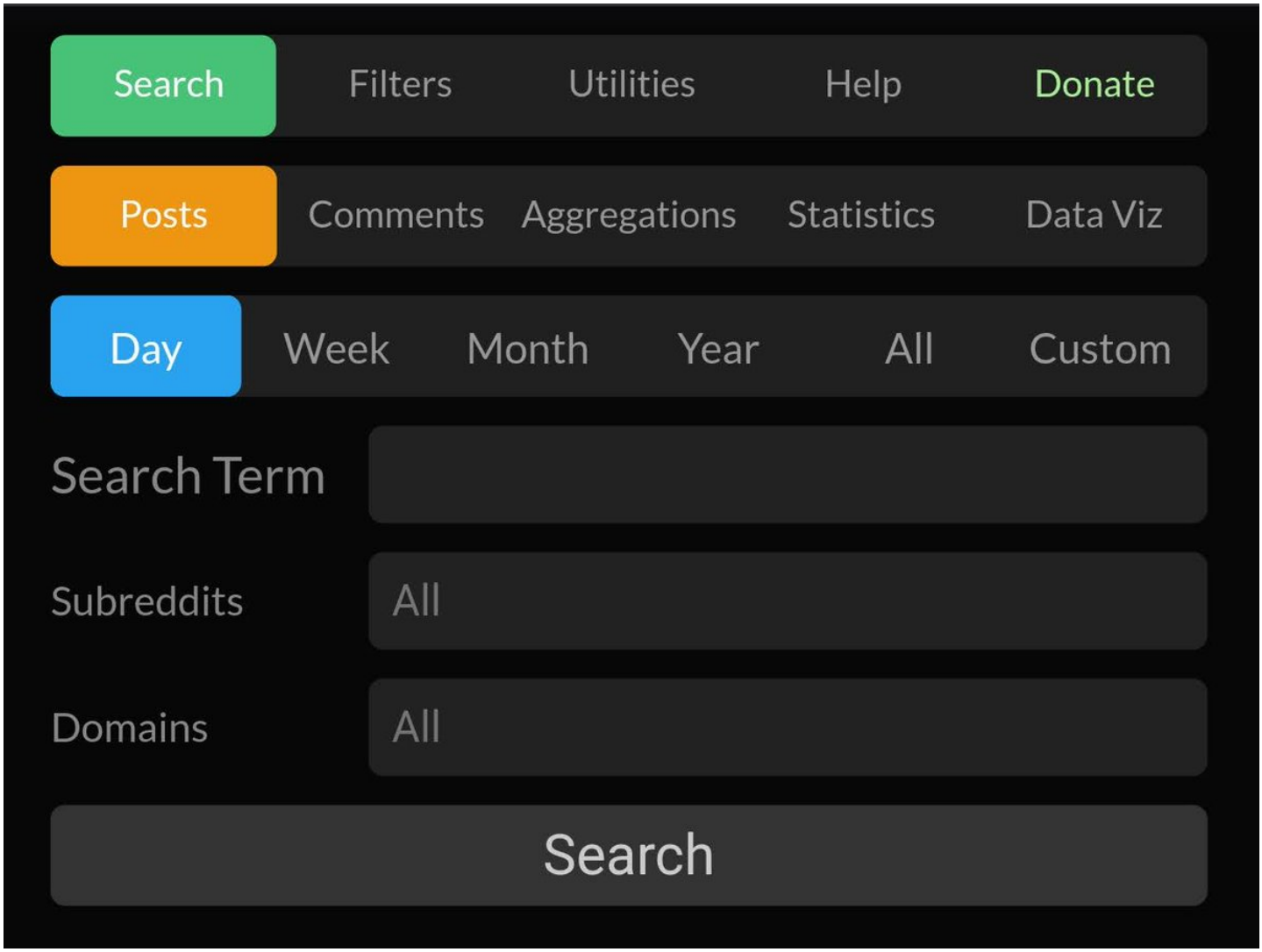
https://t.co/xSlaX3Di6T



Reddit: Redditsearch

Frontend which uses Pushshift for detail searches on subreddits or domain

https://t.co/8C37LM7aTx

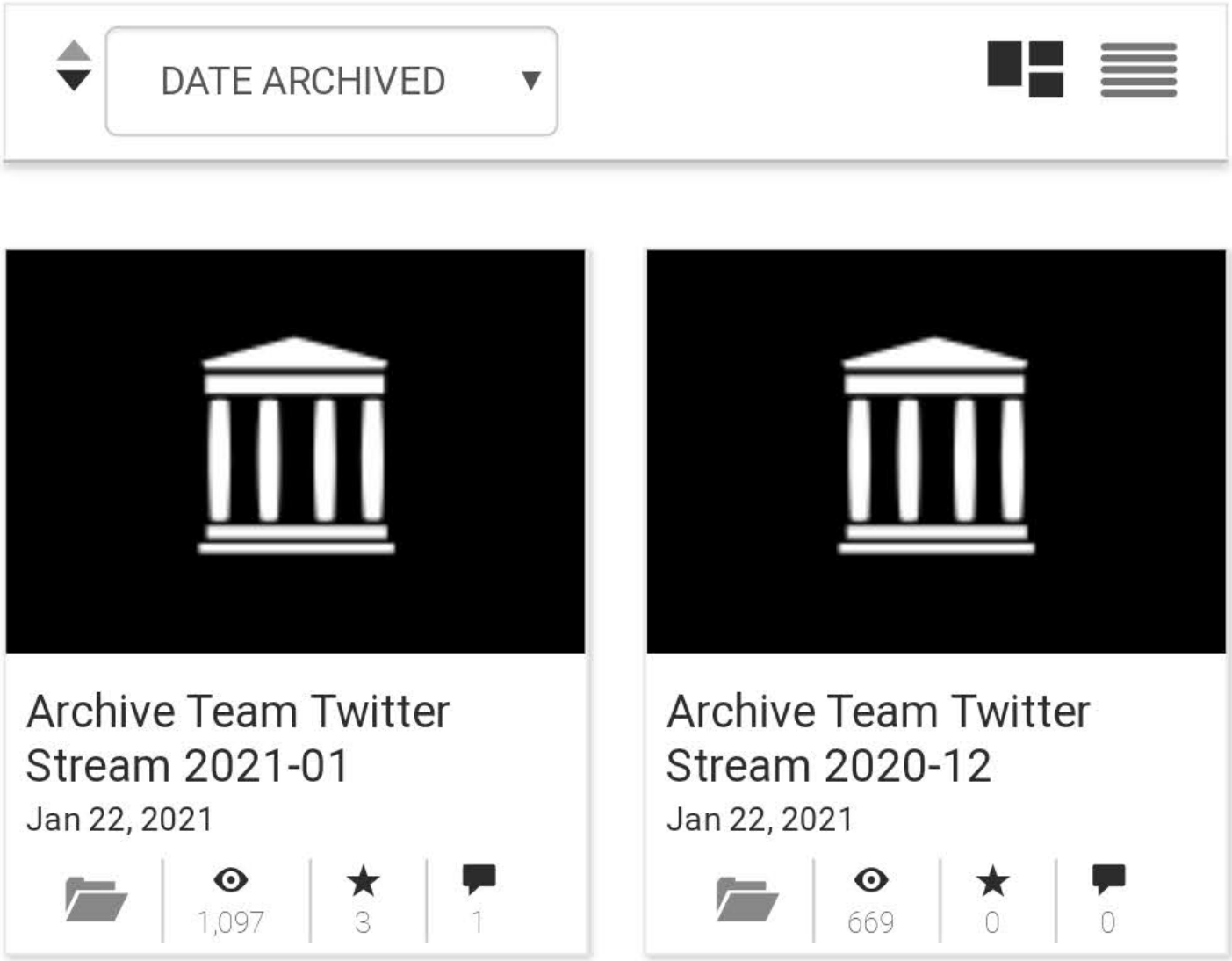


Twitter: Stream as download

The Internet Archive is a digital library of Internet sites and other cultural artifacts in digital form.

Note: The last archived data is from January 2021

https://t.co/hywWgFtbjq



Twitter: Tweepy Twitter for Python!

An easy-to-use Python library for accessing the Twitter API.

Note: The downside is the API limitations of Twitter, so you need a lot of time.

https://t.co/dhc7x1lZ8U

Tweepy: Twitter for Python!



Installation

The easiest way to install the latest version from PyPI is by using pip:

```
pip install tweepy
```

Twitter: Script

Most twitter scraper are banned by Twitter or no longer work so here is a simple and unlimited twitter scraper with python and without authentication

Note: Headless mode no longer work and it uses Selenium to access Twitter

https://t.co/feZsbOFmJR

Facebook: Scrape Facebook public pages without an API key.

$ pip install facebook-scraper

https://t.co/Rh4s93P1YD

Facebook: Large Page-Page Network data

Nodes represent official Facebook pages while the links are mutual likes between sites.

Node features are extracted from the site descriptions that the page owners created to summarize the purpose of the site.

https://t.co/tLSvGA95Az

Social networks

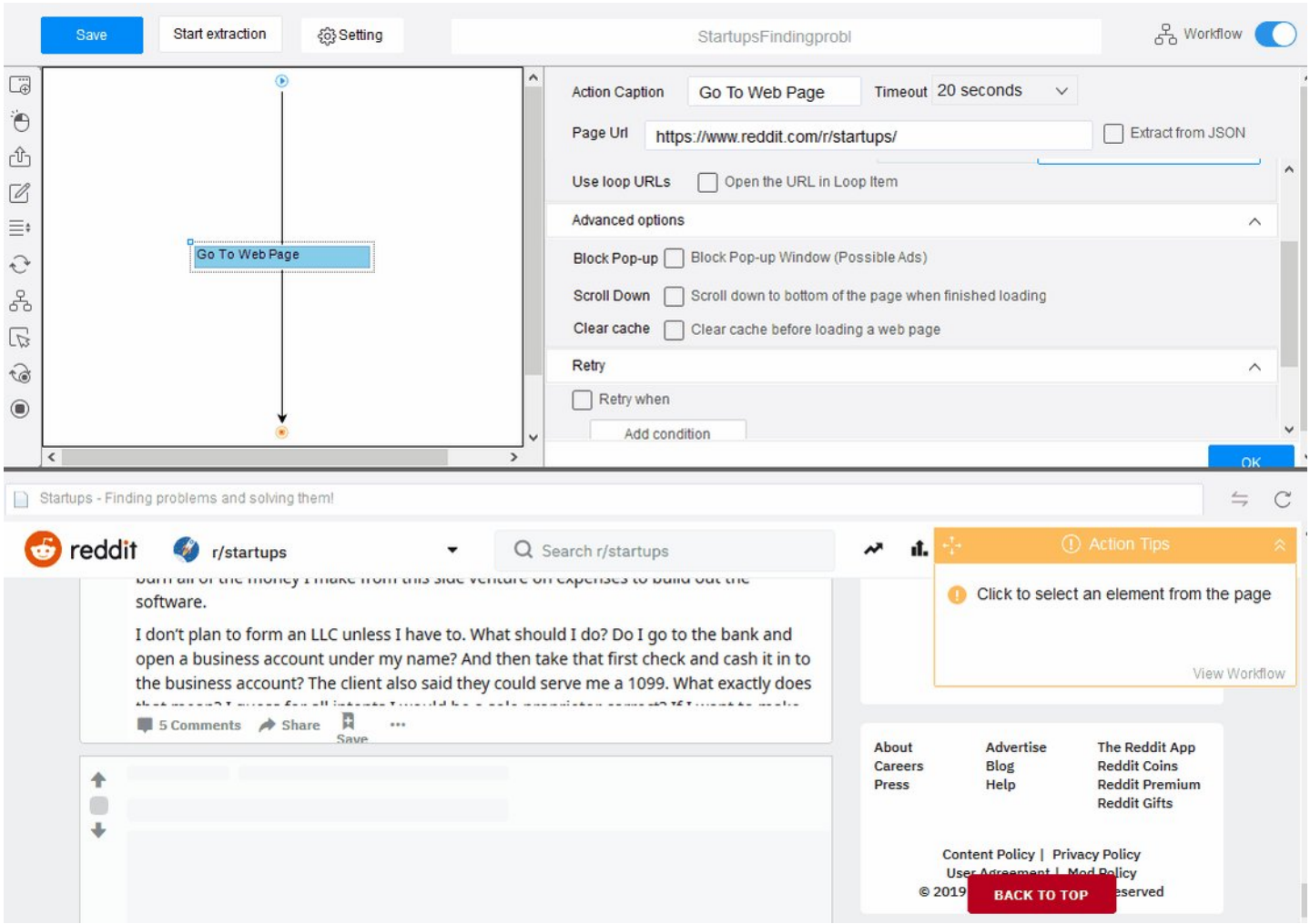
Name	Type	Nodes	Edges	Description
ego-Facebook	Undirected	4,039	88,234	Social circles from Facebook (anonymized)
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
ego-Twitter	Directed	81,306	1,768,149	Social circles from Twitter
soc-Epinions1	Directed	75,879	508,837	Who-trusts-whom network of Epinions.com
soc-LiveJournal1	Directed	4,847,571	68,993,773	LiveJournal online social network
soc-Pokec	Directed	1,632,803	30,622,564	Pokec online social network
soc-Slashdot0811	Directed	77,360	905,468	Slashdot social network from November 2008
soc-Slashdot0922	Directed	82,168	948,464	Slashdot social network from February 2009
wiki-Vote	Directed	7,115	103,689	Wikipedia who-votes-on-whom network
wiki-RfA	Directed, Signed	10,835	159,388	Wikipedia Requests for Adminship (with text)
gemsec-Deezer	Undirected	143,884	846,915	Gemsec Deezer dataset
gemsec-Facebook	Undirected	134,833	1,380,293	Gemsec Facebook dataset
soc-RedditHyperlinks	Directed, Signed, Temporal, Attributed	55,863	858,490	Hyperlinks between subreddits on Reddit
soc-sign-bitcoin-otc	Weighted, Signed, Directed, Temporal	5,881	35,592	Bitcoin OTC web of trust network
soc-sign-bitcoin-alpha	Weighted, Signed, Directed, Temporal	3,783	24,186	Bitcoin Alpha web of trust network
comm-f2f-Resistance	Weighted, Directed, Temporal	451	3,126,993	Dynamic face-to-face interaction network between group of people
musae-twitch	Undirected	34,118	429,113	Social networks of Twitch users.
musae-facebook	Undirected	22,470	171,002	Facebook page-page network with page names.
act-mooc	Bipartite, Directed, Attributed, Temporal	7,143	411,749	Student actions on a MOOC platform, with student drop-out binary labels.
musae-github	Undirected	37,700	289,003	Social network of Github developers.
feather-deezer-social	Undirected	28,281	92,752	Social network of Deezer users from Europe.
feather-lastfm-social	Undirected	7,624	27,806	Social network of LastFM users from Asia.

Octoparse: Easy Web Scraping for Anyone

Everything you need to automate your web scraping.

Note: It's a paid service.

https://t.co/f0bBLpkxSZ



Spread the open source love!

If you know an amazing project drop me message @philipvollet

we need this edit function. my inner zen isn't balanced every time i spot a typo