



Collecting digital trace data

Shuning Lu (Ph.D.)
Assistant Professor
Department of Communication
North Dakota State University

Email: shuning.lu@ndsu.edu
Twitter: [@shuning_lu](https://twitter.com/shuning_lu)



Today's agenda

Morning:

- Logistics 9:30-10:00
- Overview of digital trace data 10:00-10:30
- Data collection tutorial 10:40-11:30
- Grouping 11:30-noon

Noon:

- Lunch and guest speaker (Alvin Zhou)

Afternoon:

- Group activity & Report 13:30-16:00



What is digital trace data?

Social media

Geo-spatial data

News websites

Digital archives

Discussion: other kinds of data?



Strengths and weaknesses of digital trace data

Strengths: Always-on, unobtrusive, capture relationships

Weaknesses: proprietary, non-representativeness, drift, algorithmic confounding, unstructured, other bias

How to leverage the strengths?

How to alleviate the weakness?



Ways of collecting digital trace data

Downloading existing dataset

Screen/web scraping

API



Downloading existing dataset

<https://datasetsearch.research.google.com>

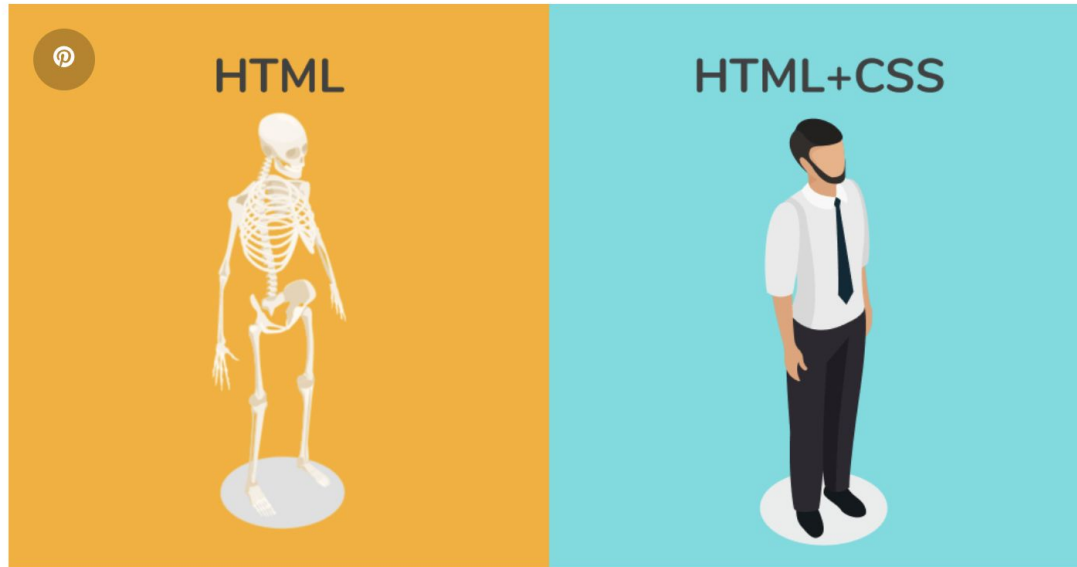
Macro-level data: World Bank, OECD, government, etc.

Traditional micro-level data: GSS, Pew Research, etc.

Public accessible platform data: Ad data, legal request data, special topics, etc.

Research data: GitHub, Open Science Framework, Harvard Dataverse, ICPSR, etc.

Screen/web scraping



- [Main page](#)
- [Contents](#)
- [Current events](#)
- [Random article](#)
- [About Wikipedia](#)
- [Contact us](#)
- [Donate](#)

Computational social science

For broader coverage of this topic, see [Quantitative social research](#).

Computational social science is the academic sub-discipline concerned with computational approaches to analyze social phenomena. Fields include [computational economics](#), [computational sociology](#), [clinical informatics](#), and [digital media studies](#). It focuses on investigating social and behavioral relationships and interactions through so-

Contents [hide]

Inspect html/css codes



Screen/web scraping workflow

Installing/loading R package: Rvest, tidyverse, ggplot2

Choosing webpage

Extracting content

Basic cleaning

Exploratory data analysis

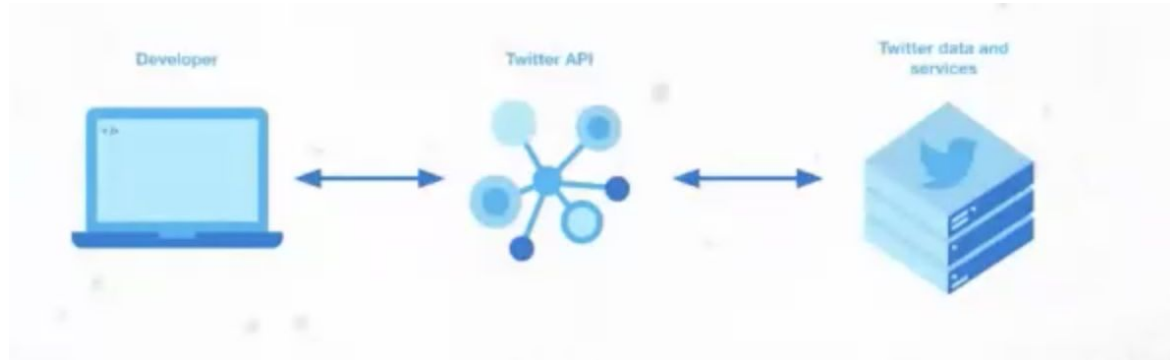


Tool time: web/screen scraping

Code along!

Using API

Application Programming Interface: customized requests of data from the server/owner





Available APIs

Reddit API

YouTube API

Twitter API

NYT API

Spotify API

...

<https://docs.google.com/spreadsheets/d/1ZEr3okdlb0zctmX0MZKo-gZKPsq5WGn1nJOxPV7al-Q/edit#gid=0>

Strengths and weaknesses of API

Strengths: efficiency, personalization, automation, integration, broad scope, structured

Weaknesses: availability, query limitation, affordability, maintenance



Social media companies like Reddit and Twitter are chasing the almighty dollar at the expense of its users and its own platforms. Credit: Avishkek Das/SOPA Images/LightRocket via Getty Images



API workflow

Read API documentation

*Sign up for a developer account

*Once approved, set up an App

*Save your credentials (keys and tokens)

*Set up user authentication information

Get your R environment set up

Start your first API request

Data wrangling

Reddit

- 100,000+ active communities
- Over 57M daily active uniques
- Over 50,000 daily active moderators (mods)
- More than 80% of the top 5,000 communities (by DAU) are open

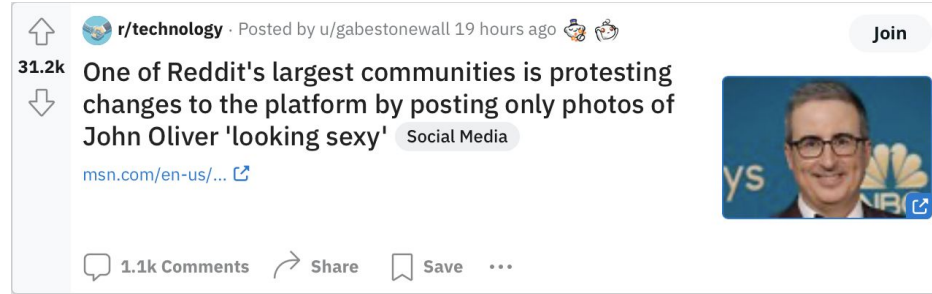
Post

Subreddit

User

Comment

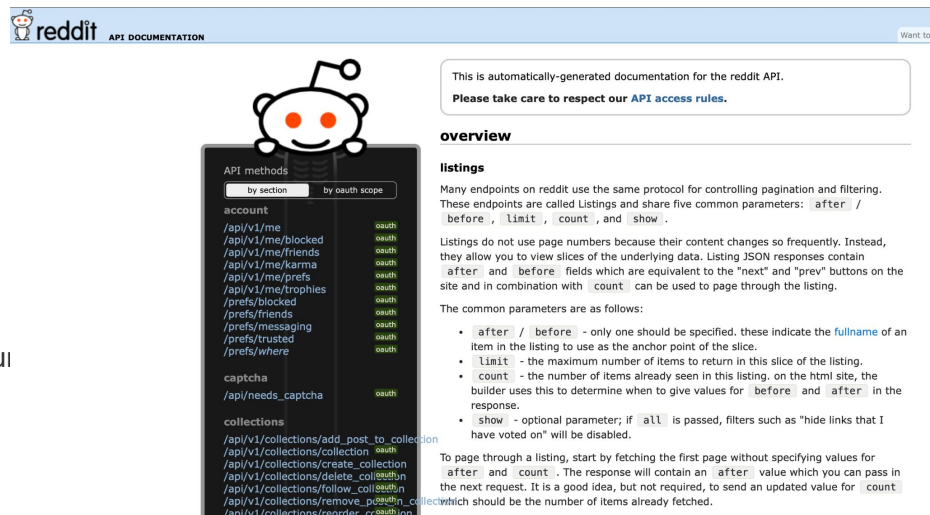
Upvotes



Reddit API

The Reddit API allows programmatic control of virtually every function user can perform on the site.

- gather and analyze data for academic research
- gather and analyze data for brand management
- monitor and moderate content on the platform
- build recommendation systems based on trends and user interests
- build bots for automated tasks (posting)
- ...



reddit API DOCUMENTATION

This is automatically-generated documentation for the reddit API.
Please take care to respect our [API access rules](#).

overview

listings

Many endpoints on reddit use the same protocol for controlling pagination and filtering. These endpoints are called Listings and share five common parameters: `after` / `before`, `limit`, `count`, and `show`.

Listings do not use page numbers because their content changes so frequently. Instead, they allow you to view slices of the underlying data. Listing JSON responses contain `after` and `before` fields which are equivalent to the "next" and "prev" buttons on the site and in combination with `count` can be used to page through the listing.

The common parameters are as follows:

- `after` / `before` - only one should be specified, these indicate the `fullname` of an item in the listing to use as the anchor point of the slice.
- `limit` - the maximum number of items to return in this slice of the listing.
- `count` - the number of items already seen in this listing, on the html site, the builder uses this to determine when to give values for `before` and `after` in the response.
- `show` - optional parameter; if `all` is passed, filters such as "hide links that I have voted on" will be disabled.

To page through a listing, start by fetching the first page without specifying values for `after` and `count`. The response will contain an `after` value which you can pass in the next request. It is a good idea, but not required, to send an updated value for `count` which should be the number of items already fetched.



R package for interacting with Reddit API

RedditExtractor



Summary

Reddit Extractor is an R package for extracting data out of Reddit. It allows you to:

1. find subreddits based on a search query
2. find a user and their Reddit history
3. find URLs to threads of interest and retrieve comments out of these threads



Tool time: Reddit API

Code along!



Group activity

Summary

An open-ended group exercise to gain practice with collecting digital trace data (screen-scraping or APIs), formulating related research questions, and extending upon the data in research design and/or presentation.

Activity

- Phase One (1:30 - 2:00pm):
 - Work together to identify a topic that digital trace data can help describe or be used to address.
 - Identify a relevant sampling frame. e.g., if your topic is on politics, the sampling frame might be a list of elected officials.
 - Discuss the strengths/weaknesses of the data for addressing your research topic.
- Phase Two (2:00 - 3:30pm):
 - Collect the data!
 - Develop a hybrid research design that allows you to combine digital trace data with some other type of data that will generate more information and address some of the weaknesses you identified with the data
 - Discuss the strengths/limitations of what you have completed.
- Come back together as a large group and discuss projects at the end of the day.