

Reminders

Complete the first three tutorials and submit completion reports by **11:59 p.m. 9/11.**

-  Announcements
-  COMM497DB Syllabus
-  Week 1: Submit your screen capture of completed tutorials along with documents of errors encountered
-  Document & share error messages received

Reminders

Report and share errors

-  Announcements
 -  COMM497DB Syllabus
 -  Week 1: Submit your screen capture of completed tutorials along with documents of errors encountered
 -  Document & share error messages received
-

Milestone check



- Tell your class partner what your R project is called, the filename of your R script, and the R project's working directory;
- Load the *rtweet* library
- Share with your group members any error/difficulty/frustration you've faced when working on tutorials

Document and share any error message & difficulty you have encountered on the public Google Docs

(https://docs.google.com/document/d/141qWy-ucwr5_5pSKAlySSfIZDRfJrEyeb8SoBnSPsAM/edit?usp=sharing)

Review

Code

↻ Start Over

▶ Run Code

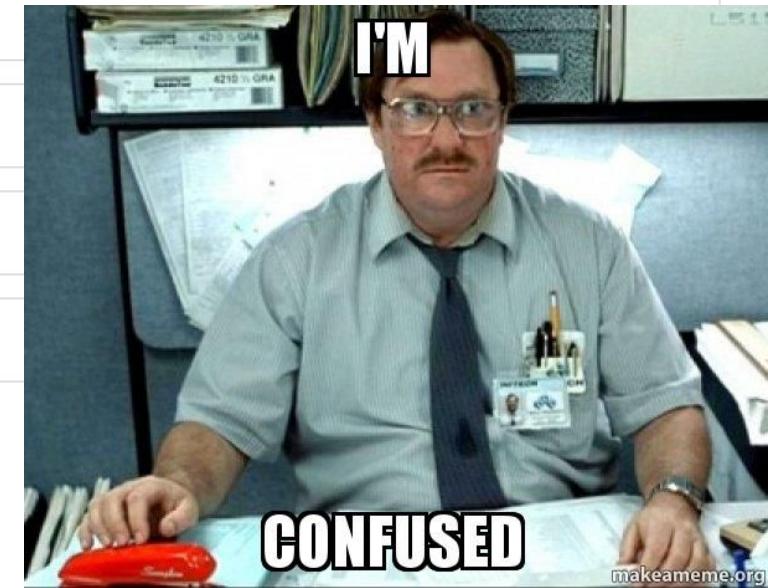
```
1 library(rtweet)
2 t <- stream_tweets("news", include_rts = FALSE, timeout = 6)
3 t
```

Streaming tweets for 6 seconds...

The stream disconnected prematurely. Reconnecting...

Reconnecting again...

NULL



Review

Why did I get this error?

Blame the “rate limiting” imposed by Twitter! Twitter caps how many data requests you can make within a 15 minute window.

Application Programming Interfaces (API)

Weiai Xu (Wayne), PhD

Assistant Professor

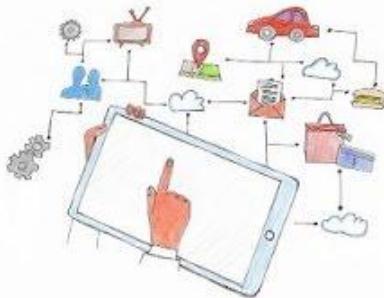
Department of Communication, UMass-Amherst

Email: weiaixu@umass.edu

curiositybits.cc

API according to YouTube videos

DEMYSTIFYING THE API



https://www.youtube.com/results?search_query=what+is+api

What is API anyway???

An API is like a doorman leading you to an internet platform's data treasures.

To access the data treasures, you need to **“swipe your key card”** and **“speak to”** the doorman in a language that he/she understands.



But first, let's talk about two conventional ways of collecting digital data

1. Web crawling (or called *screen-scraping*)
2. API

Use W2_API.R (available on Moodle)

Workflow in web crawling

1. Web crawling

Voting members by state [edit]

District	Representative	Party	Prior experience	Education	Assumed office	Residence	Born
Alabama 1	Bradley Byrne	Republican	Alabama Senate Alabama State Board of Education	Duke University (BA) University of Alabama (JD)	2014*	Fairhope	1955
Alabama 2	Martha Roby		Montgomery City Council	New York University (BM) Samford University (JD)			
Alabama 3	Mike Rogers		Calhoun County Commissioner Alabama House of Representatives	Jacksonville State University (BA, MPA) Birmingham School of Law (JD)	1 Alabama 1	Bradley Byrne	
Alabama 4	Robert Aderholt		Haleyville Municipal Judge	University of North Alabama Birmingham-Southern College (BA) Samford	2 Alabama 2	Martha Roby	



```
library(rvest)
library(tidyverse)
h <- read_html("https://en.wikipedia.org/wiki/Current_members_of_the_United_States_Congress")
reps <- h %>%
  html_node("table#votingmembers") %>%
  html_table(fill = TRUE)
```

District	Representative	Party	Prior experience	Education	Assumed office	
1 Alabama 1	Bradley Byrne	Republican	Alabama Senate Alabama State Board of Education	Duke University (BA) University of Alabama (JD)	2014*	
2 Alabama 2	Martha Roby	Republican	Montgomery City Council	New York University (BM) Samford University (JD)	2011	
3 Alabama 3	Mike Rogers	Republican	Calhoun County Commissioner Alabama House of Representatives	Jacksonville State University (BA, MPA) Birmingham School of Law (JD)	2003	
4 Alabama 4	Robert Aderholt	Republican	Haleyville Municipal Judge	Haleyville Municipal Judge	University of North Alabama Birmingham-Southern College (BA)	1997
5 Alabama 5	Mo Brooks	Republican	Alabama House of Representatives	Madison County Co...	Duke University (BA) University of Alabama (JD)	2011
6 Alabama 6	Gary Palmer	Republican	University of North Alabama		University of Alabama (BS)	2015
7 Alabama 7	Terri Sewell	Democratic	Birmingham-Southern College (BA)		Princeton University (BA) St Hilda's College, Oxford (M...	2011
8 Alaska at large	Don Young	Republican	Alaska Senate	Ship captain Mayor of Fort Yukon, Alaska	Yuba College California State University, Chico (BA)	1973*
9 Arizona 1	Tom O'Halleran	Democratic	Arizona 1	Senate	Lewis University DePaul University	2017
10 Arizona 2	Ann Kirkpatrick	Democratic	Arizona 2	U.S. House, Arizona House of Representatives	University of Arizona (BA, JD)	2019
11 Arizona 3	Raúl Grijalva	Democratic	Arizona 3	Pima County Board of Supervisors	University of Arizona (BA)	2003
12 Arizona 4	Paul Gosar	Republican	Arizona 4	President of the Northern Arizona Dental Society	Creighton University (BS, DDS)	2011
13 Arizona 5	Andy Biggs	Republican	Arizona 5	Arizona Senate	Brigham Young University (BA) University of Arizona (J...	2017
14 Arizona 6	David Schweikert	Republican	Arizona 6	Arizona House of Representatives	Arizona State University, Tempe (BS, MBA)	2011
15 Arizona 7	Ruben Gallego	Democratic	Arizona 7	Arizona Board of Education	Harvard University (BA)	2015
16 Arizona 8	Debbie Lesko	Republican	Arizona 8	Arizona House of Representatives	University of Wisconsin-Madison (BA)	2018*
17 Arizona 9	Greg Stanton	Democratic	Arizona 9	Arizona Senate President	Marquette University (BA) University of Michigan (JD)	2019
18 Arkansas 1	Rick Crawford	Republican	Arkansas 1	Broadcaster, businessman	Arkansas State University (BS)	2011
19 Arkansas 2	French Hill	Republican	Arkansas 2	Businessman	Vanderbilt University (BS)	

Workflow in web crawling

Partisan mix of the House by state [edit]

Partisan mix of the House by state

[show]

Voting members by state [edit]

District	Representative	Party	Prior experience	Education
Alabama 1	 Bradley Byrne		Alabama Senate Alabama State Board of Education	Duke University of Alabama (JL)
Alabama 2	 Martha Roby		Montgomery City Council	New York University (I) Samford University
Alabama 3	 Mike Rogers		Calhoun County Commissioner Alabama House	Jacksonville State University (I) MPA)

```
▶ <n2>...</n2>
▶ <div class="mw-collapsible mw-collapsed mw-made-collapsible" style="box-sizing:border-box;width:100%;font-size:95%;padding:4px;border:none;">...
▶ <h2>...</h2>
▼ <table class="wikitable sortable jquery-tablesorter" id="votingmembers">
  ▶ <thead>...</thead>
  ▶ <tbody>
    ▶ <tr>
      ▶ <td>...</td>
      ▶ <td nowrap="nowrap">...</td>
      <td rowspan="6" style="background-color:#E81B23">
      ▶ <td rowspan="6">Republican
      ▶ <td>...</td> == $0
      ▶ <td>
        <a href="/wiki/Duke_University" title="Duke University">Duke University</a>
        ▶ <span style="font-size:85%;">...</span>
      <br>
        <a href="/wiki/University_of_Alabama" title="University of Alabama">University of Alabama</a>
        ▶ <span style="font-size:85%;">...</span>
      </td>
      ▶ <td>2014*</td>
      ▶ <td>...</td>
      <td>1955
      ▶ </tr>
    ▶ <tr>...</tr>
```

Problems with web crawling

1. Based on extracting information from a webpage's HTML or XML codes. But each page has different formats and layouts. Thus, you need to tailor codes for each site;
2. Web crawling is like a **bot** automatically visiting and downloading data from websites. It could violate the “Terms of Service” of some websites. Most social media platforms simply do not allow web crawling.

4. PROHIBITED USE OF THE SERVICES

You may not access or use, or attempt to access or use, the Services to take any action that could harm us or a third party. You may not use the Services in violation of applicable laws or in violation of our or any third party's intellectual property or other proprietary or legal rights. You further agree that you shall not attempt (or encourage or support anyone else's attempt) to circumvent, reverse engineer, decrypt, or otherwise alter or interfere with the Services, or any content thereof, or make any unauthorized use thereof. Without NYT's prior written consent, you shall not:

(i) access any part of the Services, Content, data or information you do not have permission or authorization to access or for which NYT has revoked your access;

(ii) use robots, spiders, scripts, service, software or any manual or automatic device, tool, or process designed to data mine or scrape the Content, data or information from the Services, or otherwise access or collect the Content, data or information from the Services using automated means;

<https://help.nytimes.com/hc/en-us/articles/115014893428-Terms-of-service>

Try web crawling in R

```
library(rvest)
library(tidyverse)
h <- read_html("https://en.wikipedia.org/wiki/Current_member

reps <- h %>%
  html_node("table#votingmembers") %>%
  html_table(fill = TRUE)
```

In `W2_API.R`

You would need to install two new libraries: *rvest* and *tidyverse*

‘Hacking into’ Facebook data?

Facebook has restricted its public API, so someone has developed this (in Python)

A Facebook crawler

scrapy crawler facebook crawl spider python scraper

74 commits 1 branch 0 releases 2 contributors Apache-2.0

Branch: master New pull request Find File Clone or download

rugantio Update README.md Latest commit bda7d6a on Jul 18

fbcrawl	Adding events crawler	2 months ago
.gitignore	fixed recursion on pages	7 months ago
LICENSE	Initial commit	last year
README.md	Update README.md	2 months ago
comments.png	docs for new spider	7 months ago
runner_facebook.sh	fix runner path	2 months ago
scrapy.cfg	final	last year
trump.png	final	last year

README.md

fbcrawl

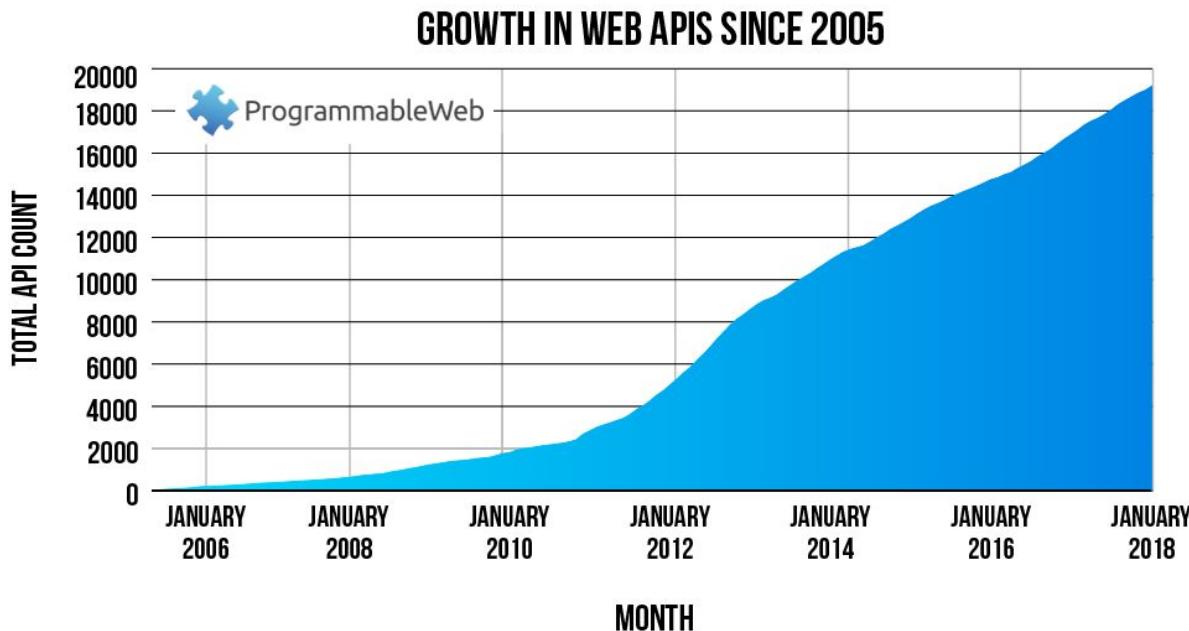
Fbcrawl is an advanced crawler for Facebook, written in python, based on the [Scrapy](#) framework.

UNMAINTAINED

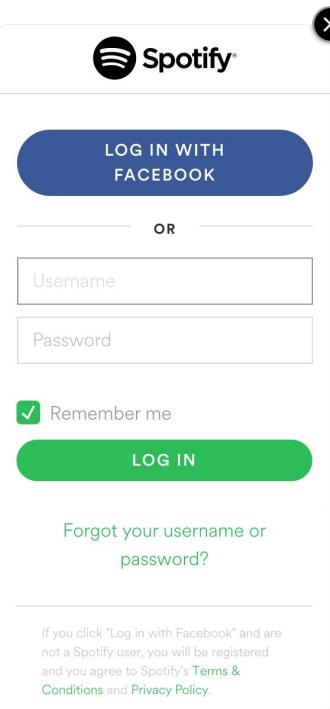
API

There are over 20,728 APIs out there. Most web platforms that you are familiar with have APIs for data sharing.

However, APIs also become increasingly restrictive.



APIs for cross-platform data sharing



Use case 1: You sign up Spotify using your Facebook account. Spotify connects to the Facebook API to retrieve your account information.

But...

Data-sharing between platforms could cause privacy concerns

Three screenshots illustrating data sharing. On the left, a Facebook sidebar menu shows links like News Feed, Messenger, Watch, Marketplace, and Explore. In the center, an Amazon sponsored ad for a NUU Mobile G3 phone is shown, featuring a 5-star rating and a 'Shop Now' button. On the right, another Amazon product listing for a Moto G6 phone is shown, also with a 5-star rating and a 'Shop Now' button.

APIs for data collection

Use case 2: Collect structured data from Twitter for analytics

API credentials

You need to “swipe your key card” to access API. For Twitter API, you need the following:

App name, API key, API secret key, Access token, Access token secret

Before making calls to Twitter’s API, you need to authenticate ourselves vis-a-vis Twitter’s API. This is done through the **create_token()** function in **rtweet** library (see the code example in the next slide).

```
#replace the following API credentials with the one posted on Moodle.  
mytoken <- create_token(  
  app = "APP1", #app name here  
  consumer_key = "DSD62iGWw16nMwaCSLkzfSQA", #consumer key here  
  consumer_secret = "p62iGWw16nMwaCSLkzfSQA", #consumer secret here  
  access_token = "153474365-XuWYfm1E423Ew6yuUM6Jfm7GMRHWJXzclWNPGCFmM", #access token here  
  access_secret = "tHq0Hq0xAqhaXHWlkXQ76HBQ7NVIX0rwvGRiH5cnsNE") #access secret here
```

This is JUST A CODE DEMO!

The API credentials shown above are NOT valid. The real API credentials are posted on Moodle (confidential account info, DO NOT SHARE)

Where do we get the API credentials

Developer Use cases Products Docs More Dashboard weiaiwayne ▾ 

Apps / COMM497DB_group1

App details **Keys and tokens** Permissions

Keys and tokens
Keys, secret keys and access tokens management.

Consumer API keys
 (API key)
 (API secret key)
[Regenerate](#)

Access token & access token secret
 (Access token)
 (Access token secret)

Read, write, and direct messages (Access level)
[Revoke](#) [Regenerate](#)

https://developer.twitter.com

How to obtain the Twitter API credentials

Prior to August, 2018, practically any Twitter user could create a Twitter app to obtain API credentials.

Due to the increasing public scrutiny over social media companies' practices in data protection, Twitter announced a big API update in 2018 that allows only **Twitter developer accounts** to obtain API credentials.

https://blog.twitter.com/developer/en_us/topics/tools/2018/new-developer-requirements-to-protect-our-platform.html

How to obtain the Twitter API credentials

You need to apply for a developer account. You will need to provide a cell-phone number and explain to Twitter what you intend to do with the app. It takes weeks or even months for Twitter to vet your application.

https://cbail.github.io/textasdata/apis/rmarkdown/Application_Programming_interfaces.html



Application under review.

Thanks! We've received your application and are reviewing it. We'll be in touch soon.

We review applications to ensure compliance with our Terms of Service and Developer policies. [Learn more.](#)

You'll receive an email when the review is complete. While you wait, check out our [documentation](#), explore our [tutorials](#), or check out our [community forums](#).

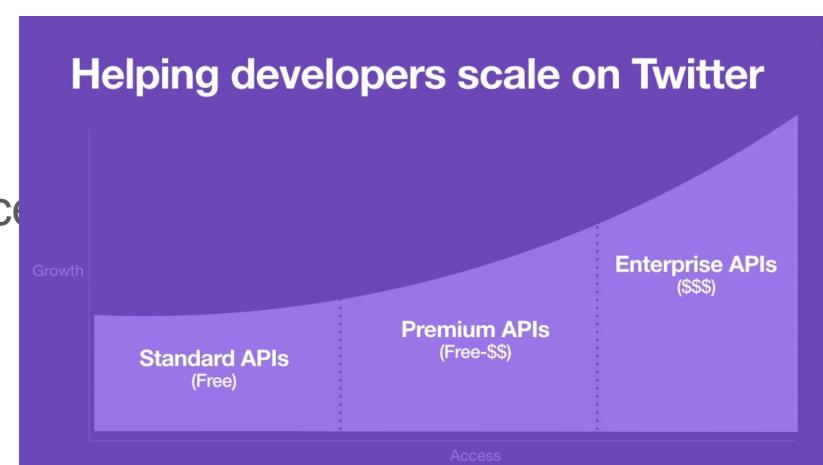
Rate limits in APIs

Internet is no longer the wild west when it comes to harvesting user data. Various web platforms impose strict rate limiting, restricting how much information an API user can collect within a period of time.

Rate limits for searching tweets:

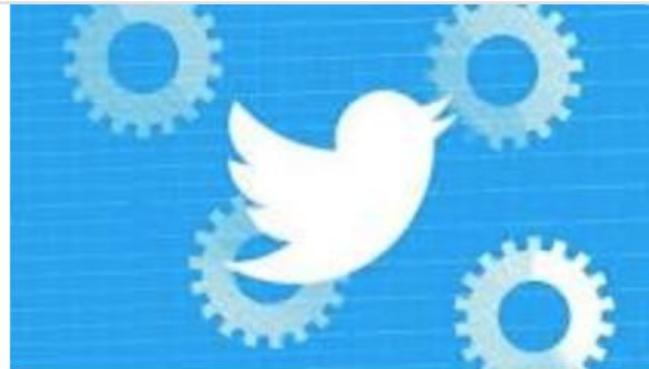
180 API calls within a 15-minute window (notice that this rule may change over time).

Want less restriction, **apply for the expensive Twitter Premium APIs!**



Data monetization

The pricing for the premium APIs ranges from \$149/month to \$2,499/month, based on the level of access needed. The first premium offering, the Search Tweets API, is launching today into public beta. This will give developers the ability to access the past 30 days of Twitter data. Nov 14, 2017



Twitter launches lower-cost subscription access to its data through ...
<https://techcrunch.com/.../twitter-launches-lower-cost-subscription-access-to-its-data-thro...>

Other limits in APIs

Historical tweets are NOT available through the free-tier Twitter API.

Want access? **Buy** data from Twitter or third-party data vendors (e.g., Gnip).

See the estimate provided by Sifter (a company in Amherst MA, no longer in business after the recent API change)

Hi Weiai Wayne,

The estimate has completed for Job: 20180323213156-7034

Rule Text: #netneutrality

Start Date: 12/01/2017

End Date: 12/20/2017

Estimated Activities: 5,881,000

The total cost of accepting this job is \$4917.50.

API will change like shifting sands

Web platforms are making APIs more restrictive for average developers/users. The access to and the power of data is now concentrated on platforms that own user data.

Twitter's recent API change has made some data analytics companies obsolete.



[home](#) [testimonials](#) [faq](#) [discovertext](#) [contact](#)

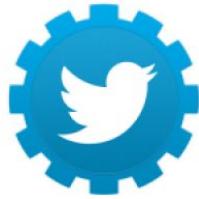
Unfortunately, this site has been decommissioned as of Sept. 30, 2018.

Thanks to our 6,970 users who created 16,128 free estimates from the complete, undeleted history of Twitter between 1/14/2014 and 9/29/2018. Please contact Twitter for approval of future academic or commercial use cases. If you can get an approved use case from Twitter, we can still help you work with the data inside [DiscoverText](#). [@DiscoverText](#) remains open and is still the **top-ranked text analysis platform on the Internet**.

All paid jobs prior to the decommissioning will still be honored.

API will change like shifting sands

Some big API changes that have occurred to social media platforms



Previously, every Twitter user could access Twitter's API for free, with some restrictions.

Open only to Twitter developer accounts.



Previously, you could download a public Facebook page's posts and comments for free.

No access to public Facebook page data



Deen Freelon

@dfreelon

Following

So Facebook has shuttered API access without direct approval:

[newsroom.fb.com/news/2018/04/r ...](https://newsroom.fb.com/news/2018/04/restricting-data-access/)

Party's over, folks--time to start thinking about what happens if/when Twitter does the same.

“
These changes will better protect people’s information while still enabling developers to create useful experiences.”

An Update on Our Plans to Restrict Data Access on Facebook

Two weeks ago we promised to take a hard look at the information apps can use when you connect them to Facebook as well as other data practices. Today, we

newsroom.fb.com

10:04 PM - 5 Apr 2018

<https://newsroom.fb.com/news/2018/04/restricting-data-access/>

April 4, 2018

An Update on Our Plans to Restrict Data Access on Facebook

“

We believe these changes will better protect people’s information while still enabling developers to create useful experiences.

”

Pages API: Until today, any app could use the Pages API to read posts or comments from any Page. This let developers create tools for Page owners to help them do things like schedule posts and reply to comments or messages. But it also let apps access more data than necessary. We want to make sure Page information is only available to apps providing useful services to our community. So starting today, all future access to the Pages API will need to be approved by Facebook.

Discuss the questions with your class partner

- Why do web platforms provide APIs?
- Why do web platforms make APIs more restrictive?
- Having reviewed the current API rules and recent API changes, what's your take on the issue of equality in data protection and data access?



**SEE ONE,
DO ONE,
TEACH ONE.**

A graphic element consisting of a thin horizontal line with a small diamond shape in the center.

Try it yourself

Use **W_API.R** (it is on Moodle!)

W_API.R has two parts: one for doing the traditional web crawling (which you can safely skip), and the other for connecting to Twitter's API. Make sure you can connect to the API using the credential I've provided.

You will share the credentials. In order not to hit the rate limit, set a small n .

Required tutorials for this week

An interactive tutorial
for COMM 497DB

Weiwei Wayne Xu

Libraries/packages

Data frames

Connecting to the Twitter API

Collect tweets by keywords-hashtags

Collect Twitter user timeline

Collect Twitter user info

Make Wordclouds

Predict Ideology (in progress)

Collect Twitter user info

Collecting user information? That sounds creepy!

Not at all. We will conduct the data collection in strict compliance with Twitter's developer terms. In fact, just like on collecting tweets, Twitter makes it very limited as to what kind of user profile data are available through its API.

What can you do with the Twitter user data?

After running the code in this part of the tutorial, you will end up with a data frame containing a bunch of screen names and profile bios. You might ask: what can I do with it? You will be surprised by how much insights we can draw by just looking at these profiles. For example, we can use artificial intelligence to predict a user's ideology. We will, of course, save the predicted ideologies in a separate column.

Get followers and friends

Running the code below, we can get Nassim Nicholas Taleb's (@nntaleb) followers and friends. Nassim Nicholas Taleb is a well-known statistician and favorite author. He wrote the famous [*The Black Swan: The Impact of the Highly Improbable: With a new section on Fragility*](#). Interestingly, Taleb served as a [UMass-Amherst faculty](#) from January 2005 to January 2006.

Code



Start Over

Start Over

Wanna grab YouTube metadata? Try Google's API

Google Cloud Platform



New Project



You have 21 projects remaining in your quota. Request an increase or delete projects.

[Learn more](#)

[MANAGE QUOTAS](#)

Project Name *

COMM497DB

Project ID: comm497db. It cannot be changed later. [EDIT](#)

Location *

No organization

[BROWSE](#)

Parent organization or folder

[CREATE](#)

[CANCEL](#)



<https://console.cloud.google.com/apis/api>

Google now requires an app review to approve your API request

Wanna grab YouTube metadata? Try Google's API

Google Cloud Platform COMM497DB Search YOUTUBE

4 results

YouTube Data API v3 Google

The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...

YouTube Analytics API Google

Retrieves data from YouTube Analytics.

YouTube Ads Reach API Google

API Library

YouTube Reporting API Google

OBSOLETE

YouTube Data API v3 Google

The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...

ENABLE TRY THIS API ↗

Enable all YouTube APIs. Also enable Freebase API

Wanna grab YouTube metadata? Try Google's API

The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists, and channels.

MANAGE **TRY THIS API**

CREATE CREDENTIALS

Enable all YouTube APIs. Also enable Freebase API

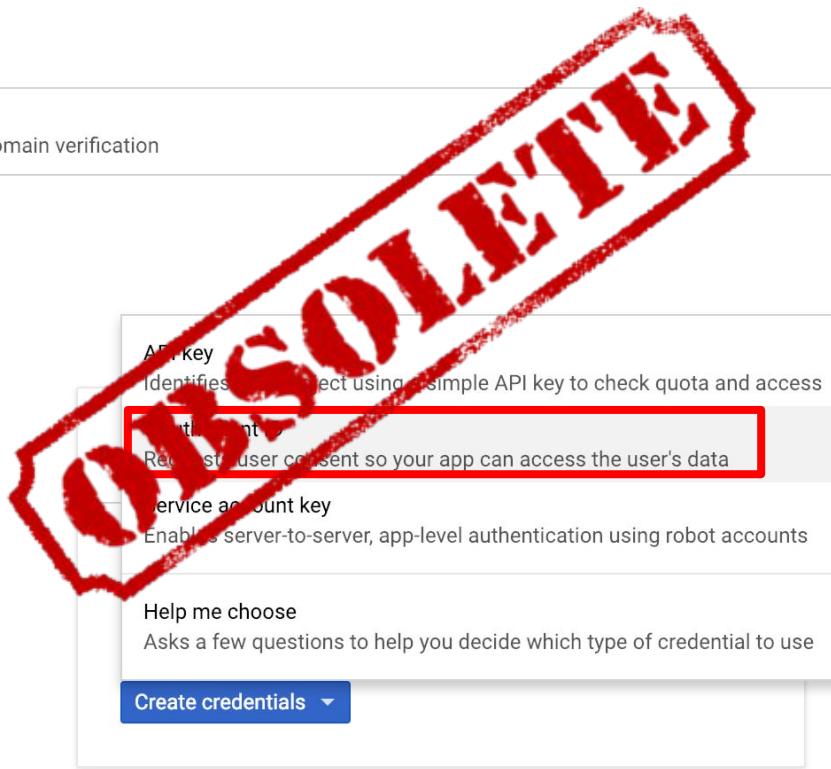
Wanna grab YouTube metadata? Try Google's API

Credentials

Credentials

OAuth consent screen

Domain verification



Wanna grab YouTube metadata? Try Google's API

[←](#) Create OAuth client ID

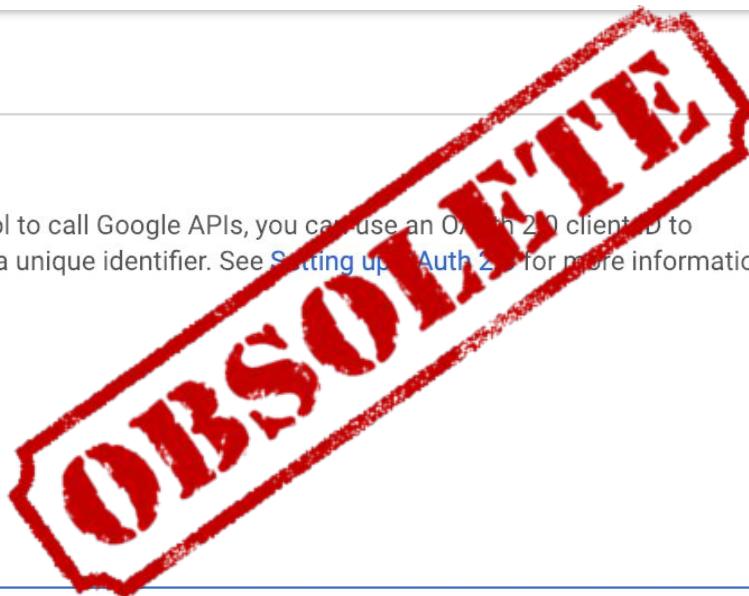
For applications that use the OAuth 2.0 protocol to call Google APIs, you can use an OAuth 2.0 client ID to generate an access token. The token contains a unique identifier. See [Setting up OAuth 2.0](#) for more information.

Application type

- Web application
- Android [Learn more](#)
- Chrome App [Learn more](#)
- iOS [Learn more](#)
- Other

Name [?](#)

app2



[Create](#)

[Cancel](#)

Wanna grab YouTube metadata? Try Google's API

Credentials

Credentials

OAuth consent screen

Domain verification

Create credentials

Create credentials t

OAuth 2.0 client

Name

app2

tuber1

OAuth client

The client ID and secret can always be accessed from Credentials in APIs & Services

i OAuth is limited to 100 [sensitive scope logins](#) until the [OAuth consent screen](#) is published. This may require a verification process that will take several days.

Here is your client ID

Here is your client secret

OK



Wanna grab YouTube metadata? Try Google's API

```
library(tuber)
```

```
#connect to YouTube's API. More at https://github.com/soodoku/tuber  
yt_oauth("enter Client ID here", "enter Client secret here", token = '')
```



Wanna grab YouTube metadata? Try Google's API

Sign in with Google

localhost:1410/?state=0zEo75EvSc&code=4/4ABz6TzHb6k4

Choose an account to continue to tuber

Weiai Xu
weiai.wayne.xu@gmail.com

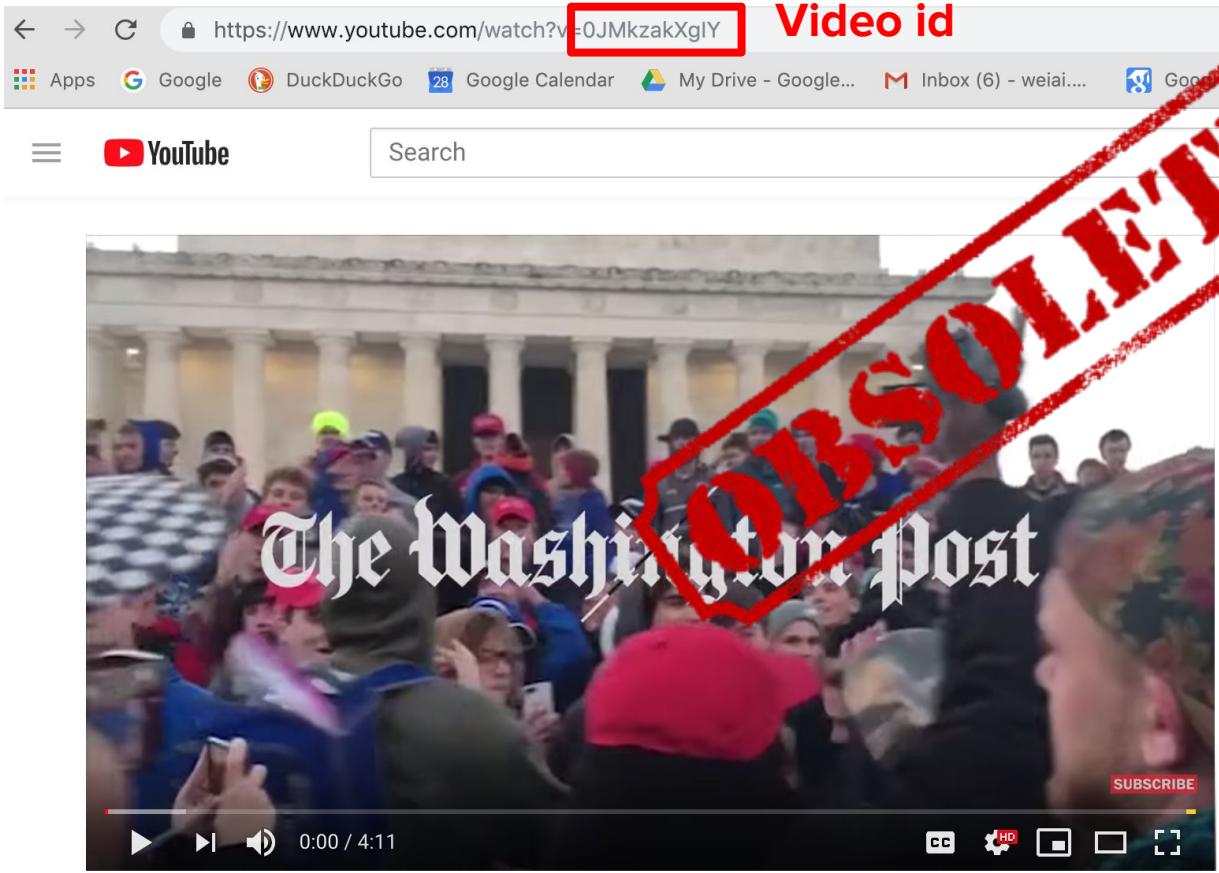
Use another account

OBSOLETE

Authentication complete. Please close this page and return to R.

Waiting for authentication in browser...
Press Esc/Ctrl + C to abort
Authentication complete.

Wanna grab YouTube metadata? Try Google's API



Try it yourself

Use **W_API.R** (it is on Moodle!).

The third part of the script is for using YouTube's API.

Milestone check



- Collect a bunch of tweets

Document and share any error message & difficulty you have encountered on the public Google Docs

(https://docs.google.com/document/d/141qWy-ucwr5_5pSKAllySSfIZDRfJrEyeb8SoBnSPsAM/edit?usp=sharing)