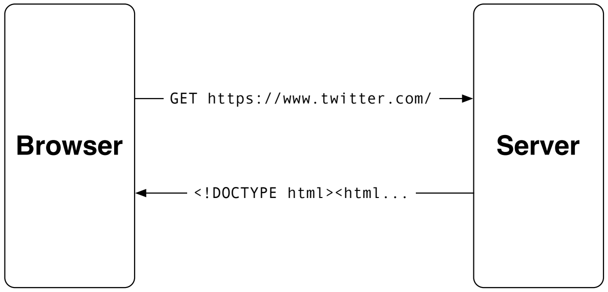
Homework

- Lab open M–F 8:00am–4:30pm
- On github.com
  - Join team Everyone
  - Class materials at http://github.com/Ling583/notes
  - Read *Bad Data Handbook* Chapter 4: “Bad Data Lurking in Plain Text”
  - Updated python in baddata4.ipynb
- On datacamp.com
  - Finish *Intro to Python for Data Science* and *Introduction to Shell for Data Science*
  - Do *Intermediate Python for Data Science*

Web APIs

- Online sources often make text available through an API (an interface that allows two programs to talk to each other)
- ‘RESTful’ APIs are built on top of the web’s native HTTP protocol



Web APIs

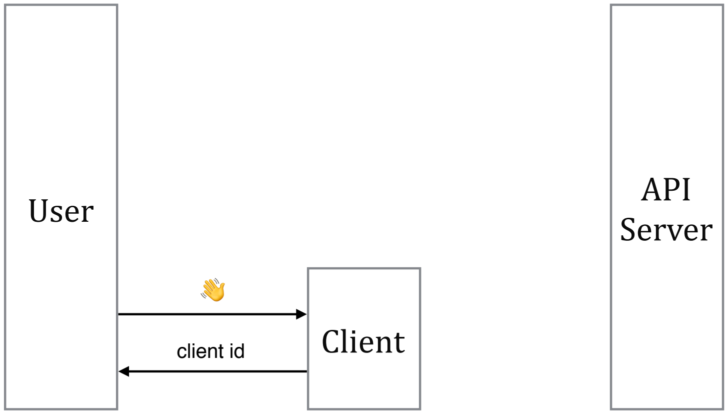
- Specialized URLs retrieve data (usually in XML or JSON)

```
GET https://api.twitter.com/1.1/search/tweets.json?
q=%23freebandnames&since_id=24012619984051000&max_id=2501261998405181
45&result_type=mixed&count=4
```

```
{“statuses”:
[{“coordinates":null,"favorited":false,"truncated":false,"created_at"
:"MonSep2403:35:21+00002012","id_str":"250075927172759552","entities"
:{"urls":[],"hashtags":[{"text":"freebandnames","indices":
[20,34]}],"user_mentions":
[]},"in_reply_to_user_id_str":null,"contributors":null,"text":"Aggres
sivePonytail#freebandnames","metadata":
{"iso_language_code":"en","result_type":"recent"},"retweet_count":
0,"in_reply_to_status_id_str":null,"id":
250075927172759552,"geo":null,"retweeted":false,"in_reply_to_user_id"
:null,"place":null,"user":
{"profile_sidebar_fill_color":"DDEEF6","profile_sidebar_border_color"
:"C0DEED","profile_background_tile":false,"name":"SeanCummings","prof
ile_image_url":"http://a0.twimg.com/profile_images/
2359746665/1v6zfgqo8g0d3mk7ii5s_normal.jpeg","created_at":"MonApr2606
:01:55+00002010","location":"LA,CA","follow_request_sent":null,"profi
le_link_color":"0084B4","is_translator":false,"id_str":"137238150","e
ntities":{"url":{"urls":[{"expanded_url":null,"url":"","indices":...
```

Web API

- Easy to handle HTTP requests in Python using the ‘requests’ module
- APIs define endpoints (URLs) and request types (GET, PUT, POST, DELETE, …)
- Authentication to control access to data
  - Control who as access to private data
  - Limit quantity of data the can be retrieved and/or the rate at which it can be retrieved
  - Most Web APIs support OAuth2 authentication



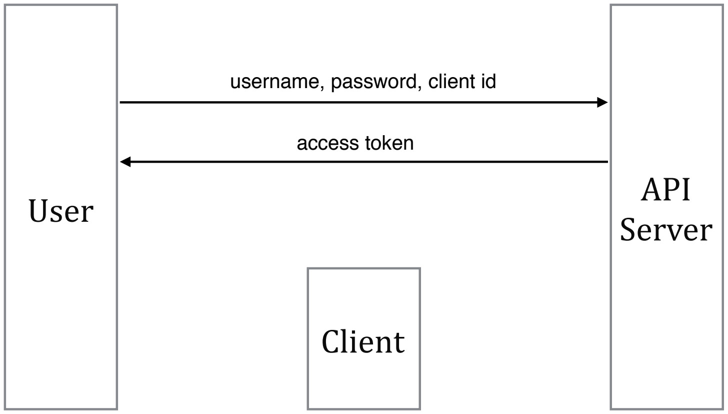
User

API
Server

👋

Client

client id



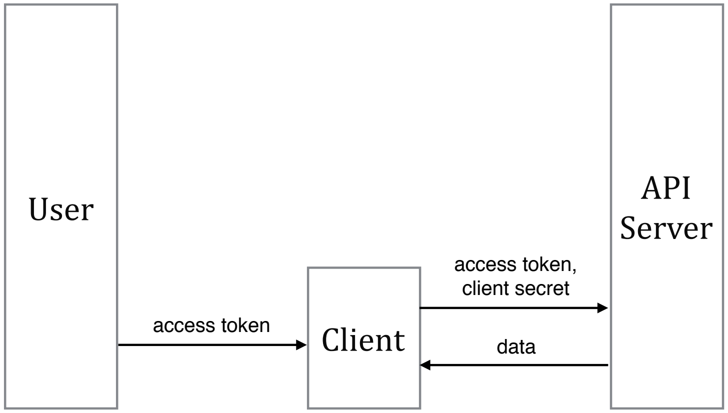
username, password, client id

access token

User

API
Server

Client



User

API
Server

access token,
client secret

access token

Client

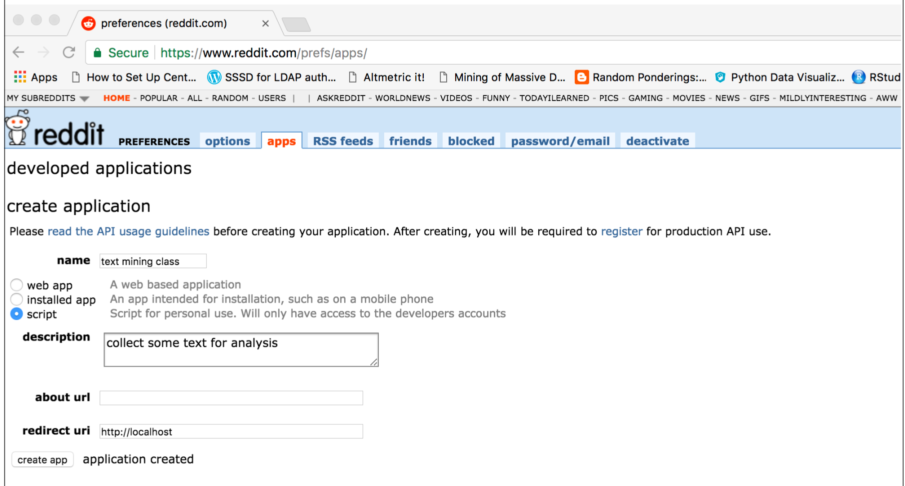
data

Web API

- Big-name web sites have site-specific Python libraries that wrap API calls in Python functions
- Most also handle OAuth2 authentication for you as well
- Quality of documentation varies
- A few deal with rate throttling, etc
- twython, praw

Text project

- The task for the first project is to collect and analyze text from Reddit comments
- More specifically: collect comment text and metadata from a specific subreddit and find keywords for specific users
- More more specifically:
  - register a reddit script application
  - write script application to download a sample of comments from a sub reddit
  - save comments in a pandas dataframe and group by author
  - calculate PMI for how often each word is employed by author



developed applications

create application

Please read the API usage guidelines before creating your application. After creating, you will be required to register for production API use.

name text mining class

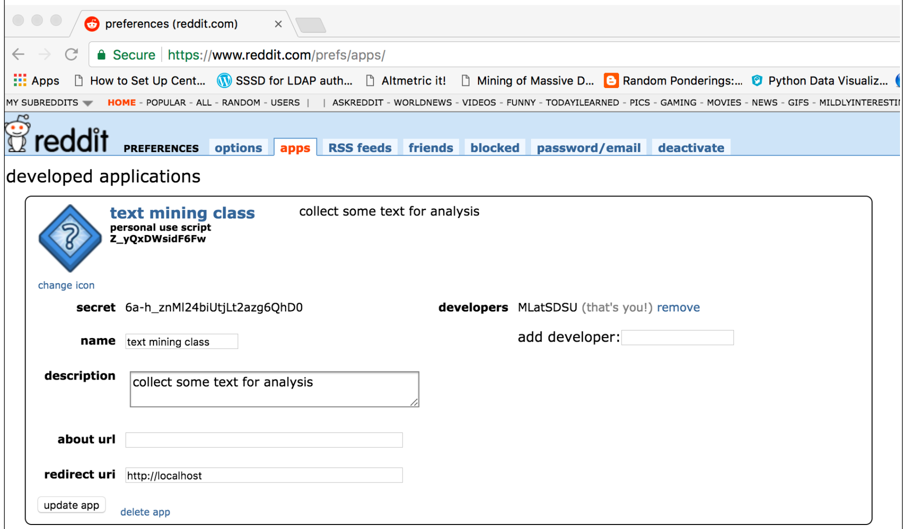
○ web app — A web based application
○ installed app — An app intended for installation, such as on a mobile phone
● script — Script for personal use. Will only have access to the developers accounts

description collect some text for analysis

about url

redirect uri http://localhost

create app application created



developed applications

text mining class collect some text for analysis
personal use script
Z_yQxDWsidF6Fw

change icon

secret 6a-h_znMl24biUtjLt2azg6QhD0

developers MLatSDSU (that's you!) remove

add developer:

name text mining class

description collect some text for analysis

about url

redirect uri http://localhost

update app delete app

Text project

- Data collection will take time (due to bandwidth limits and rate throttling), so . . .
  - Write program to download data
  - Collect a small development sample
  - Run program to download lots of data
  - Using development sample, write (and test) program to analyze data
  - When done with full data collection, apply finished analysis program