Web Scraping

https://github.com/kaopanboonyuen/GISTDA2023

Topics

- Part 1: Web Scraping
- Part 2: Twitter Scraping

Reference

1. https://scrape-it.cloud/blog/web-scraping-vs-api
2. https://thevatsalsaglani.medium.com/web-scraping-using-python-and-beautifulsoup-2e54e79415d6
3. https://www.scrapingbee.com/blog/python-web-scraping-beautiful-soup/
4. https://realpython.com/beautiful-soup-web-scraper-python/

Web Scraping vs API:
What's the Best Way to Extract Data?

https://scrape-it.cloud/blog/web-scraping-vs-api

What is Web Scraping?

Web scraping is a technique for automatically extracting target data from the Internet.

Scraping helps to take raw data in the form of HTML code from sites and convert it into a usable structured format.

When you try to extract any content from the Internet, it's called web scraping, even if you do it manually.

Web scrapers are used mainly by companies that want to gather information to understand their customers better, follow competitors, or do research.

For example, in the world of e-commerce, online retailers periodically analyze the publicly available pages of their competitors, scraping product titles and prices so they can adjust their pricing policies accordingly.

Explore the Website

https://realpython.github.io/fake-jobs/

Fake Python

Fake Jobs for Your Web Scraping Journey

Senior Python Developer
Payne, Roberts and Davis

Stewartbury, AA

2021-04-08

Learn | Apply

Energy engineer
Vasquez-Davidson

Christopherville, AA

2021-04-08

Learn | Apply

Legal executive
Jackson, Chambers and Levy

Port Ericaburgh, AA

2021-04-08

Learn | Apply

Fitness centre manager
Savage-Bradley

East Seanview, AP

2021-04-08

Learn | Apply

Explore the Website

https://realpython.github.io/fake-jobs/

Fake Python

Fake Jobs for Your Web Scraping Journey

Senior Python Developer
Payne, Roberts and Davis

Stewartbury, AA

2021-04-08

Learn | Apply

Energy engineer
Vasquez-Davidson

Christopherville, AA

2021-04-08

Learn | Apply

Legal executive
Jackson, Chambers and Levy

Port Ericaburgh, AA

2021-04-08

Learn | Apply

Fitness centre manager
Savage-Bradley

East Seanview, AP

2021-04-08

Learn | Apply

realpython.github.io/fake-jobs/

Fake Python

Fake Jobs for Your Web Scraping Journey

h2.title.is-5 240 × 22.5

Senior Python Developer

Payne, Roberts and Davis

Stewartbury, AA

2021-04-08

Elements Console Sources Network Performance Memory Application

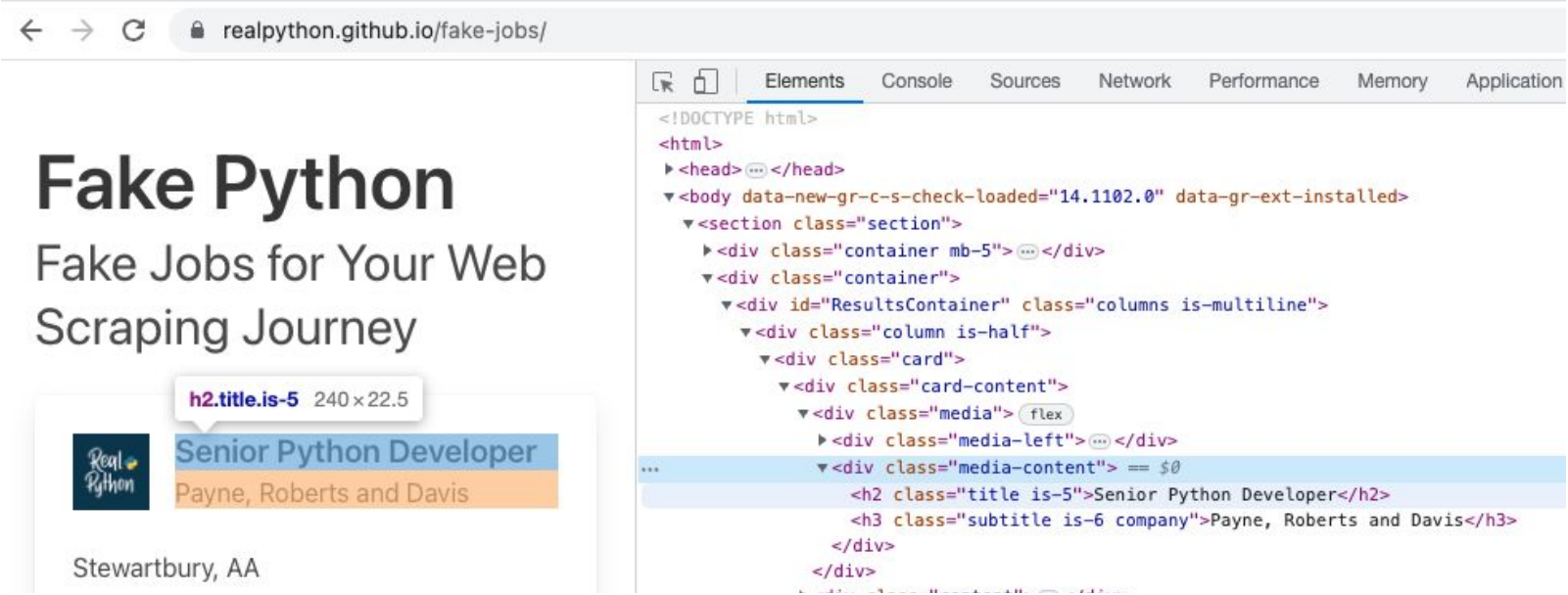
```
<!DOCTYPE html>
<html>
 ▸<head>…</head>
 ▾<body data-new-gr-c-s-check-loaded="14.1102.0" data-gr-ext-installed>
  ▾<section class="section">
   ▸<div class="container mb-5">…</div>
   ▾<div class="container">
    ▾<div id="ResultsContainer" class="columns is-multiline">
     ▾<div class="column is-half">
      ▾<div class="card">
       ▾<div class="card-content">
        ▾<div class="media">  flex
         ▸<div class="media-left">…</div>
…        ▾<div class="media-content"> == $0
           <h2 class="title is-5">Senior Python Developer</h2>
           <h3 class="subtitle is-6 company">Payne, Roberts and Davis</h3>
          </div>
         </div>
        ▸<div class="content">…</div>
        ▸<footer class="card-footer">…</footer>  flex
       </div>
      </div>
```

- **`class="title is-5"`** contains the title of the job posting.
- **`class="subtitle is-6 company"`** contains the name of the company that offers the position.
- **`class="location"`** contains the location where you'd be working.

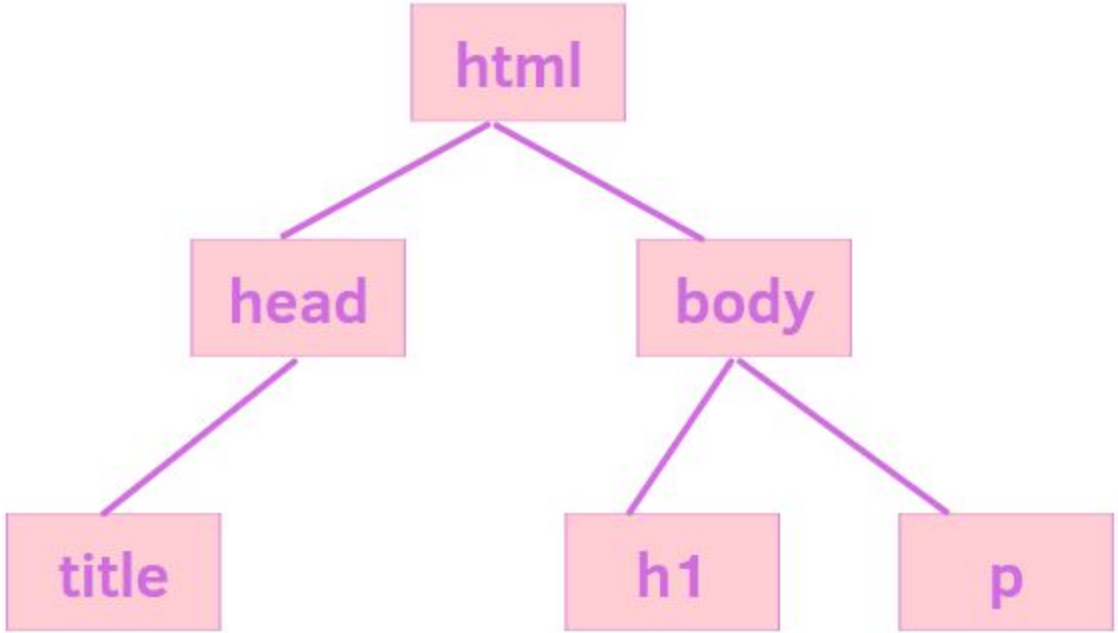
Christopherville, AA

… section.section div.container div#ResultsContainer.columns.is-multiline div.column.is-half div.

Console What's New ×



- **`class="title is-5"`** contains the title of the job posting.
- **`class="subtitle is-6 company"`** contains the name of the company that offers the position.
- **`class="location"`** contains the location where you'd be working.


html
head
body
title
h1
p

WEB SCRAPING

PROS

Faster manual data collection

Ease of working with structured results

Data accuracy is higher than manual collection

Running on a schedule to get up-to-date data regularly

CONS

The need for regular maintenance

Requires specialized knowledge

It can be blocked when a large number of requests

The need to use proxies to avoid restrictions (geo-blocking, CAPTCHAs, etc)

Some difficulties with dynamic sites

What is API?

API stands for Application Programming Interface, which acts as an intermediary, allowing websites and software to communicate and exchange data and information.

To contact the API, you need to send it a request. The client must provide the URL and HTTP method to process the request correctly. You can add headers, body, and request parameters depending on the method.

Headers provide metadata about the request.

The body contains data such as fields for a new row in a database.

The API will process the request and send the response received from the web server.

Endpoints work in conjunction with API methods. Endpoints are specific URLs that the application uses to communicate with third-party services and its users.

API SCRAPING

PROS

Less resource-intensive, as unnecessary data is not loaded

Easy integration into applications for further data processing

The data is already structured

Bypasses an issue with dynamic page rendering

Faster than web scraping

CONS

Not all data can be obtained with one request

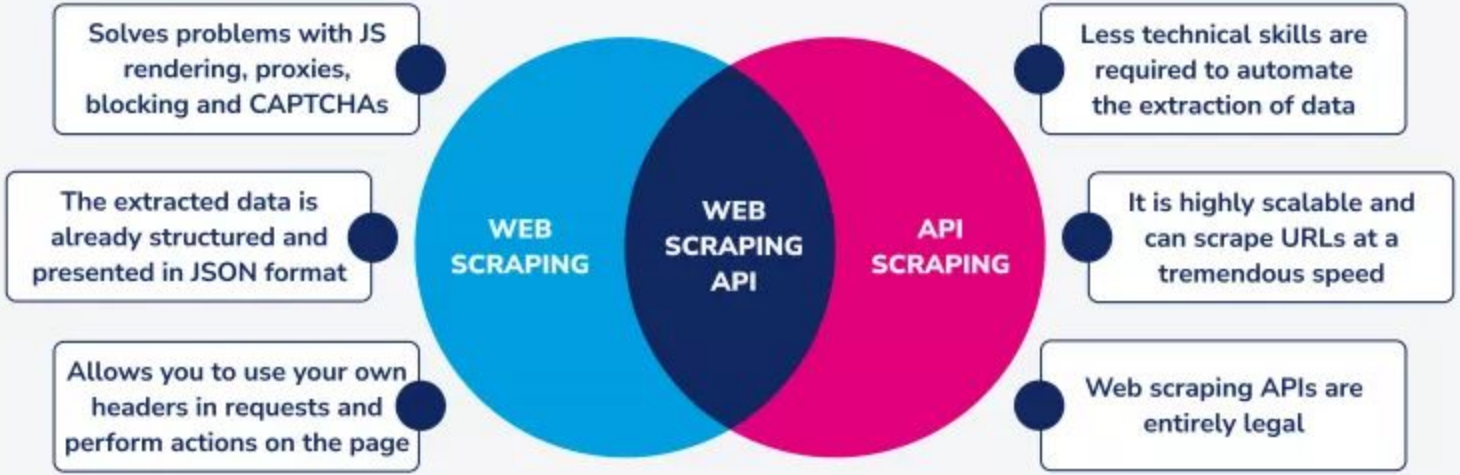
Not all sites have API endpoints

Limits on the number of requests from one IP and their frequency

APIs are generally limited to extracting data from a single website

WEB SCRAPING API: A COMBINATION OF THE TWO

Benefits Of Web Scraping API

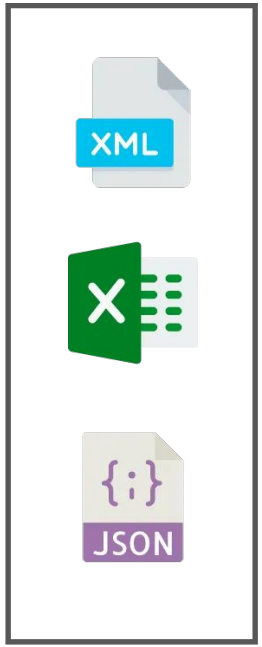


Web Scraping With Python Using BeautifulSoup

Target a website and define what data must be collected

Extract and analyze the source code

Build a scraper in Python



Structured data

https://medium.datadriveninvestor.com/introduction-to-scraping-in-python-with-beautifulsoup-and-requests-ab7b1c9bc113

What is web scraping, and why do I need it?

- Get recipes from your favorite cooking website or photos from a travel blog.
- Without an API, extracting the HTML, or scraping, might be the only way to get that content. I'm going to show you how to do this in Python.

Let’s take a look at the required Python libraries:

- The **request** library to make network requests
  - To scrape data from a website, we need to extract the content of the webpage.
  - Once the request is made to a website, the entire content of the webpage is available, and we can then evaluate the web content to extract data out from it. The content is made available in the form of plain text.

Python

```python
import requests

URL = "https://realpython.github.io/fake-jobs/"
page = requests.get(URL)

print(page.text)
```

Let’s take a look at the required Python libraries:

- The **html5lib** library for parsing HTML
  - Once the content is available, we need to specify the library that represents the parsing logic for the text available.
  - We’ll be using the html5lib library to parse the text content to HTML DOM-based representation.

Let’s take a look at the required Python libraries:

- The **beautifulsoup4** library for navigating the HTML tree structure
  - beautifulsoup4 takes the raw text content and parsing library as the input parameters.
  - In our example, we have exposed html5lib as a parsing library.
  - It can then be used to navigate and search for elements from the parsed HTML nodes.
  - It can pull data out from the HTML nodes and extract/search required nodes from HTML structure.

Python

```python
import requests
from bs4 import BeautifulSoup

URL = "https://realpython.github.io/fake-jobs/"
page = requests.get(URL)

soup = BeautifulSoup(page.content, "html.parser")
```

Beautiful Soup allows you to find that specific HTML element by its ID:

Python

```python
results = soup.find(id="ResultsContainer")
```

Python

```python
job_elements = results.find_all("div", class_="card-content")
```

Python

```python
for job_element in job_elements:
    title_element = job_element.find("h2", class_="title")
    company_element = job_element.find("h3", class_="company")
    location_element = job_element.find("p", class_="location")
    print(title_element)
    print(company_element)
    print(location_element)
    print()
```

Python

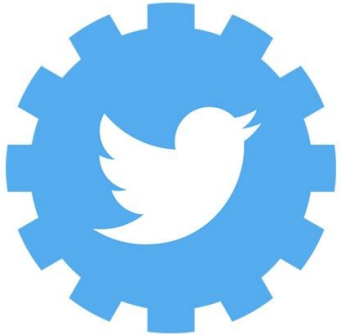
```python
for job_element in job_elements:
    title_element = job_element.find("h2", class_="title")
    company_element = job_element.find("h3", class_="company")
    location_element = job_element.find("p", class_="location")
    print(title_element.text.strip())
    print(company_element.text.strip())
    print(location_element.text.strip())
    print()
```

The results finally look much better:

Text

```text
Senior Python Developer
Payne, Roberts and Davis
Stewartbury, AA
```

Web Scraping Exercises:

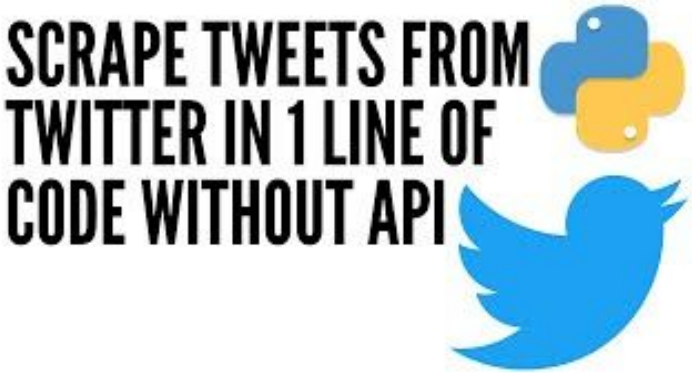


How to Scrape Millions of Tweets using SNSCRAPER

Snscrape

Snscrape is a scraper for social networking services (SNS).

It scrapes details like user profiles, hashtags, or searches and returns the discovered items, e.g. the relevant posts.



```python
# Set up the search query
search_term = "ลุงตู่"
since_date = "2022-01-01"
until_date = "2023-01-31"
#geocode = "13.736717,100.523186, 50km" # search within 50 km of bangkok

# Setting variables to be used below
maxTweets = 5000

# Creating list to append tweet data to
tweets_list = []

# create the search query
query = f"{search_term} since:{since_date} until:{until_date}"
```

```
tweets_df2
```

	Datetime	Tweet Id	Text	Username
0	2022-05-09 15:05:08+00:00	1523680474883043328	#ลุงตู่ #นายกลุงตู่ #covid19 https://t.co/YtyY...	TSJ2518
1	2022-05-09 14:43:12+00:00	1523674953354072065	"รักจริงๆ ไม่มีวันทิ้ง #ลุงตู่ ค่ะ"\nกำลังใจอย...	Kea_New
2	2022-05-09 14:32:10+00:00	1523672180499705856	ขอบคุณซ่อนะคะที่ช่วยโปรโมทผลงาน #ลุงตู่ น่ารั...	Kea_New
3	2022-05-09 14:17:22+00:00	1523668455160827905	เอาอีกแล้ว..อีตู่นะอีตู่😄😍😍\nแบบนี้..ชาวชลบุรี...	Kea_New
4	2022-05-09 13:53:31+00:00	1523662450226204672	@chanya_nath @Nakarin_KT แต่ประเทศไทยยังมีน้ำม...	TheAirgun
...	...	...	...	...
96	2022-05-02 02:31:19+00:00	1520954057317568512	เบาลงบ้าง อย่าชิงดีชิงเด่นกันเลย มาร่วมด้วยช่ว...	mmsamphant
97	2022-05-01 16:09:51+00:00	1520797657257889793	#ลุงตู่ #นายกลุงตู่ https://t.co/pPTLP6xgt8	TSJ2518
98	2022-05-01 15:44:56+00:00	1520791387272212480	@Pacifica_Kaz @vnomenon #ลุงตู่ พาพวกเราไปใช้ ...	causeiloveTH
99	2022-05-01 14:40:01+00:00	1520775051595108352	สู้ต่อไปนะคะ #ลุงตู่ ✌✌\nวันนี้มีลุงตู่ พรุ่...	Kea_New
100	2022-05-01 11:31:54+00:00	1520727708615217152	#ลุงตู่ #นายกลุงตู่ 01.05.2565 #วันแรงงานแห่งช...	TSJ2518

101 rows × 4 columns


ไทย
รวม
สร้างชาติ
พรรค
ประยุทธ์
อยู่
ที่
ประเทศ
66
สร้าง
ต้อง
ต่อ
ให้
ปี
มา
ทำ
และ
ทำต่อ
กับ
เป็น
ใน
นายกฯ
เลือก
ก็
นี้
ว่าๆ
ครับ
นายก
ทีม
ด้วย
2
จะ
112
ผม
เขต
ทุก
คน
แล้ว
2566
ไม่
แต่
ได้
ของ
ค่ะ
ไป
ขอ
save

```python
class Tweet(typing.NamedTuple, snscrape.base.Item):
    url: str
    date: datetime.datetime
    content: str
    renderedContent: str
    id: int
    username: str # Deprecated, use user['username'] instead
    user: 'User'
    outlinks: list
    outlinksss: str # Deprecated, use outlinks instead
    tcooutlinks: list
    tcooutlinksss: str # Deprecated, use tcooutlinks instead
    replyCount: int
    retweetCount: int
    likeCount: int
    quoteCount: int
    conversationId: int
    lang: str
    source: str
    media: typing.Optional[typing.List['Medium']] = None
    retweetedTweet: typing.Optional['Tweet'] = None
    quotedTweet: typing.Optional['Tweet'] = None
    mentionedUsers: typing.Optional[typing.List['User']] = None
```