

Web scraping tools

Dr. Adel Ammar

Adapted from:

- ESTP course on Big Data Sources – Web, Social Media and Text Analytics, Olav ten Bosch, Statistics Netherlands
- Beautiful Soup: Build a Web Scraper With Python:
<https://realpython.com/beautiful-soup-web-scraper-python>

Outline

- Introduction
- Scraping tools
- Some scraping knowledge
- Web Scraping with BeautifulSoup in Python

Introduction (1)



- Web scraping is the **automated** process of **extracting data** from **websites**, often using scripts or programs.
- This technique enables users to gather large volumes of information efficiently, which can be utilized for analysis, research, or integration into applications.
- The incredible amount of data on the Internet is a rich resource for any field of research or personal interest. To effectively harvest that data, you'll need to become skilled at web scraping.

Introduction (2)



- The web scraping process involves carefully inspecting data sources, extracting raw HTML content, and parsing it to locate relevant information.
- There are many different tools for scraping available, which differ in their functionality and use.
- Tools and frameworks come and go, choose the one that fits the job.
- Scraping is not like typical IT. The life cycle (design, develop, test, maintain) is much shorter, it might not even be a cycle (one time use).

Introduction (3)

- Any tool is useless without some basic knowledge of web technology and internet experience, so we provide you some.
- At the end of this session we will do a simple scraping exercise with Python libraries for web scraping.

Introduction (4)

- We make a rough distinction between:
 - **Scraping**: the actual extraction of data / information from a web page
 - **Crawling**: following hyperlinks on the internet to traverse multiple pages and / or sites
 - **Search**: using (third party) search engines (such as Google) automatically to find information on the web
- Many tools offer a mix of these

Reasons for Web Scraping

- Organizations utilize web scraping to aggregate data, monitor market trends, gather competitive intelligence, or automate routine tasks such as data entry.
- This practice allows businesses to obtain insights that drive strategic decisions and operational efficiencies.

Challenges of Web Scraping

- Web scraping can face various challenges, including diverse website structures, frequent changes in page layouts, and legal issues concerning data usage.
- Unstable scripts are a realistic scenario, as many websites undergo active development.
 - The scrapers you'll build will probably require constant maintenance.
- Additionally, some sites implement anti-scraping measures to protect their content, limiting accessibility.

Scraping tools (1)



- iMacros (officially discontinued as of November 30, 2023):
 - Available for quite some years
 - Point and click (record and replay) as well as coding via API (Application Programming Interface)
 - Browser add-in as well as standalone program
 - Type of functionality: scrape, automate repetitive tasks on web pages (such as data extraction, form filling, and web testing).
 - Free and commercial version
 - Easy to start with

iMacros x
Recording
VERSION BUILD=9030808 RECORD...
TAB T=1
URL GOTO=http://www.ikea.com/...



PRODUCTEN ▾

RUIMTES ▾

WOONINSPIRATIE

DIT IS IKEA ▾



WOONKAMER: Bankstellen & fauteuils | Tv- & mediameubels | **Wandmeubels** | Salon- & bijzettafels | Verlichting | Textiel | Series

Dressoirs, buffetten & bijzettafels | Opbergsystemen woonkamer | **Boekenkasten** | Wandmeubelseries | Wandplanken | Kasten & vitrinekasten | Dozen & manden

☐ Online verkrijgbaar

Relevantie ▴ ▾



Prijs (EUR) | 5 | 807 | **Filter**

BOEKENKASTEN

Niet alleen voor boekenwormen

Droom je van een volledige boekenwand? Krijgen je tv en je familiekiekjes ook een plaatsje in de boekenkast? Hoe dan ook, onze multifunctionele boekenkasten zijn perfect voor alle spullen die je om je heen wil hebben. Maak een keuze uit de vele stijlen en maten.



Play Rec Manage

Record

Save Macro As

Stop

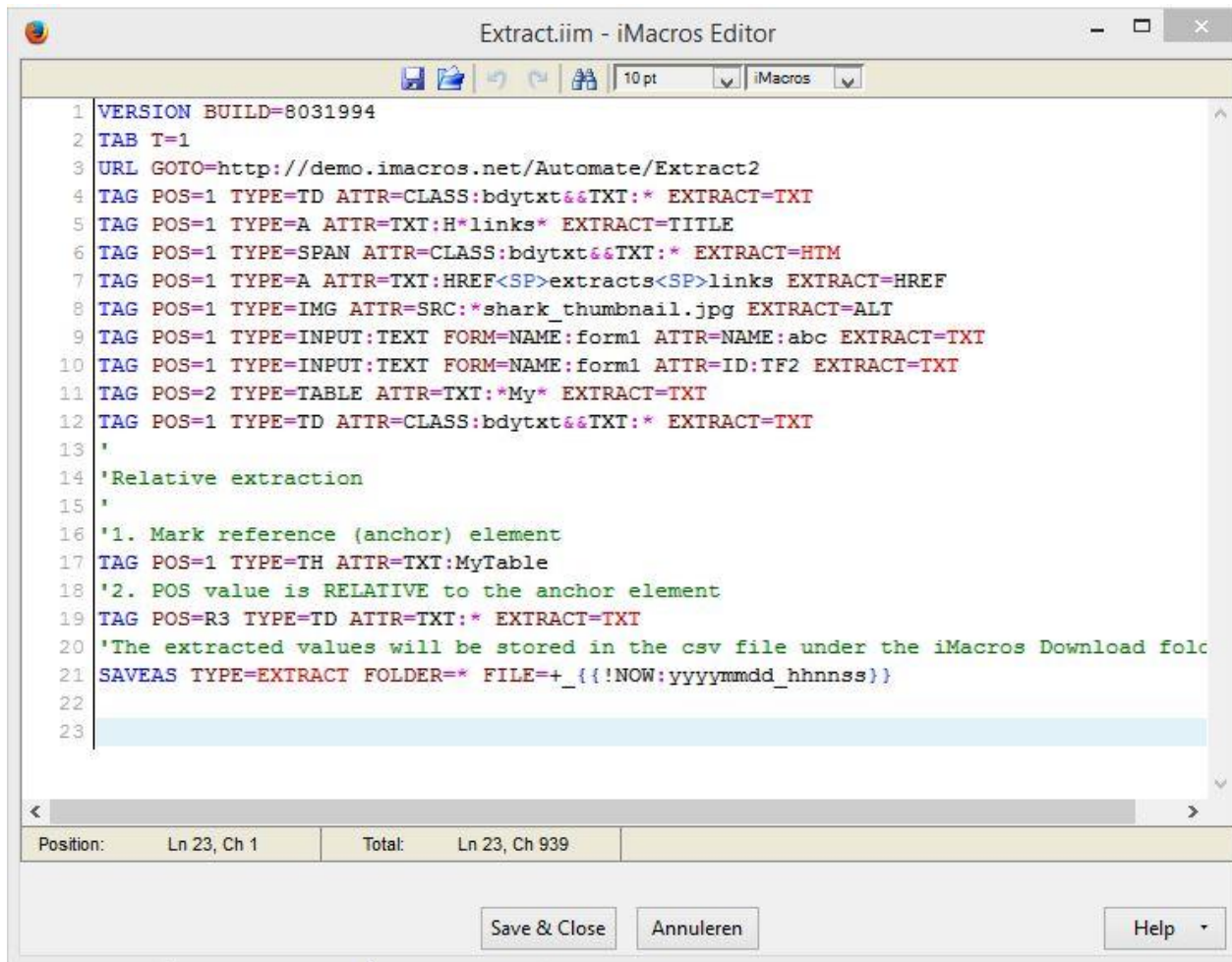
Record options

Save Page As

Take Screenshot

Del. Cache&Cookies

iMacros: generated code



```
1 VERSION BUILD=8031994
2 TAB T=1
3 URL GOTO=http://demo.imacros.net/Automate/Extract2
4 TAG POS=1 TYPE=ID ATTR=CLASS:bdytxt&&TXT:* EXTRACT=TXT
5 TAG POS=1 TYPE=A ATTR=TEXT:H*links* EXTRACT=TITLE
6 TAG POS=1 TYPE=SPAN ATTR=CLASS:bdytxt&&TXT:* EXTRACT=HTM
7 TAG POS=1 TYPE=A ATTR=TEXT:Href<SP>extracts<SP>links EXTRACT=HREF
8 TAG POS=1 TYPE=IMG ATTR=SRC:*shark_thumbnail.jpg EXTRACT=ALT
9 TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:form1 ATTR=NAME:abc EXTRACT=TEXT
10 TAG POS=1 TYPE=INPUT:TEXT FORM=NAME:form1 ATTR=ID:TF2 EXTRACT=TEXT
11 TAG POS=2 TYPE=TABLE ATTR=TEXT:*My* EXTRACT=TEXT
12 TAG POS=1 TYPE=ID ATTR=CLASS:bdytxt&&TXT:* EXTRACT=TEXT
13 '
14 'Relative extraction
15 '
16 '1. Mark reference (anchor) element
17 TAG POS=1 TYPE=TH ATTR=TEXT:MyTable
18 '2. POS value is RELATIVE to the anchor element
19 TAG POS=R3 TYPE=ID ATTR=TEXT:* EXTRACT=TEXT
20 'The extracted values will be stored in the csv file under the iMacros Download folder
21 SAVEAS TYPE=EXTRACT FOLDER=* FILE=+_{{!NOW:yyyymmdd_hhnnss}}
22
23
```

Position: Ln 23, Ch 1 Total: Ln 23, Ch 939

Save & Close Annuleren Help

The script language is specific to iMacros and allows users to interact with web elements using commands like TAG, EXTRACT, and SAVEAS.

Scraping tools (2)



- Scrapy(scrapy.org):
 - **Python-based scraping and crawling framework**
 - **More IT oriented: coding skills required**
 - **Open source**
 - **Large user community**
 - **Used by some National Statistical Institutes for various scraping tasks**

Scrapy example

```
import scrapy

from tutorial.items import DmozItem

class DmozSpider(scrapy.Spider):
    name = "dmoz"
    allowed_domains = ["dmoz.org"]
    start_urls = [
        "http://www.dmoz.org/Computers/Programming/Languages/Python/",
    ]

    def parse(self, response):
        for href in response.css("ul.directory.dir-col > li > a::attr('href')"):
            url = response.urljoin(href.extract())
            yield scrapy.Request(url, callback=self.parse_dir_contents)

    def parse_dir_contents(self, response):
        for sel in response.xpath('//ul/li'):
            item = DmozItem()
            item['title'] = sel.xpath('a/text()').extract()
            item['link'] = sel.xpath('a/@href').extract()
            item['desc'] = sel.xpath('text()').extract()
            yield item
```



Scraping tools (3)

- Import.io:
 - **Point and click *and* coding**
 - **Fully web-based and hosted scraping**
 - **Type of functionality: scrape**
 - **Free and commercial licenses**

Extract web data the easy way

Drive data insight with the world's #1 web data platform.



Try it out

Request free trial

See some examples



geproefd en heeft nu een versing de zomer en haal de picknicksfeer naar binnen, door samen nog uren door te brengen aan de eetkamertafel.

[Bekijk eetkamertafels >](#)



NEW! Reduced price



RYET

Led-lamp E27 400 lumen,
globe

~~€ 6,99~~ € 3.99 / 2 st.



SKINANDE

Inbouwvaatwasser

~~€ 499,-~~ € 449,- /st.



KALLAX

Open kast, 3 kleuren

~~€ 149,-~~ € 129,- /st.



KULINARISK

Stoomoven, roestvrij staal

~~€ 699,-~~ € 599,- /st.



finally...
**Dinner in your own
garden!**

Add or manage URLs

Create a blank table

















Undo

Redo

Download CSV

Data view

Website view

#	Blockcenter link	Center image	Textupper value	Linkblock value	Textstrike price	Textbold price	Text value
1	RYET Led-lamp GU10 200 lumen...		RYET	Led-lamp GU10 200 lumen	€ 4.99	€ 3.99	/2 st.
2	VOLFGANG Stoel, verchroomd,...		VOLFGANG	Stoel, verchroomd, zwart	€ 69.95.-	€ 59.95.-	/st.
3	OTROLIG Inductiekookplaat € ...		OTROLIG	Inductiekookplaat	€ 499.-	€ 449.-	/st.
4	RÅSKOG Roltafel, beige € 49.9...		RÅSKOG	Roltafel, beige	€ 49.95	€ 39.95	/st.
5	RYET Led-lamp E27 400 lumen...		RYET	Led-lamp E27 400 lumen, globe	€ 6.99	€ 3.99	/2 st.
6	SKINANDE Inbouwvaatwasser' ...		SKINANDE	Inbouwvaatwasser	€ 499.-	€ 449.-	/st.
7	KALLAX Open kast, 3 kleuren €...		KALLAX	Open kast, 3 kleuren	€ 149.-	€ 129.-	/st.
8	KULINARISK Stoomoven, roestv...		KULINARISK	Stoomoven, roestvrij staal	€ 699.-	€ 599.-	/st.
9	RYET Led-lamp GU10 200 lumen...		RYET	Led-lamp GU10 200 lumen	€ 4.99	€ 3.99	/2 st.
10	VOLFGANG Stoel, verchroomd,...		VOLFGANG	Stoel, verchroomd, zwart	€ 69.95.-	€ 59.95.-	/st.
11	OTROLIG Inductiekookplaat € ...		OTROLIG	Inductiekookplaat	€ 499.-	€ 449.-	/st.
12	RÅSKOG Roltafel, beige € 49.9...		RÅSKOG	Roltafel, beige	€ 49.95	€ 39.95	/st.
13	RYET Led-lamp E27 400 lumen...		RYET	Led-lamp E27 400 lumen, globe	€ 6.99	€ 3.99	/2 st.
14	SKINANDE Inbouwvaatwasser' ...		SKINANDE	Inbouwvaatwasser	€ 499.-	€ 449.-	/st.
15	KALLAX Open kast, 3 kleuren €...		KALLAX	Open kast, 3 kleuren	€ 149.-	€ 129.-	/st.
16	KULINARISK Stoomoven, roestv...		KULINARISK	Stoomoven, roestvrij staal	€ 699.-	€ 599.-	/st.



Add column

Scraping tools (4): Selenium

- Selenium is a popular tool for automating web browsers. It allows users to write scripts that interact with web pages, mimicking user actions like clicking, form submission, and navigation.
- **Key Features:**
 - Supports multiple programming languages (Python, Java, JavaScript, etc.)
 - Works with different browsers (Chrome, Firefox, Safari)
 - Useful for web scraping, automated testing, and task automation

Scraping tools (4): Selenium

- **Advantages:** Handles dynamic content, JavaScript rendering, and more complex web interactions compared to static scraping tools.
- **Example Use:** Automating form submissions, scraping data from JavaScript-heavy pages.

Scraping tools (5)

- There are many more, such as:
 - **Nutch for crawling (Apache, java)**
 - **An extensive list is available on:**

<https://github.com/lorien/awesome-web-scraping>

- Scraping tools by Statistics Netherlands:
 - **CBS Robot Framework**
 - **CBS Robottool, a tool for detecting changes on websites**

Some scraping knowledge (1)

- HTTP: the communication protocol
- HTML: the language in which web pages are defined
- JS: javascript (code executing in the browser)
- CSS: style sheets, how web pages are styled. Important, but does not contain data.
- JPG, PNG, BMP: images, usually not interesting
- CSV / TXT / JSON / XML: data, interesting !!!

Some scraping knowledge (2)

- Before initiating a scrape, it's crucial to understand the structure of the target website.
- Use browser developer tools to inspect elements (Ctrl+Shift+I), locate data, and decipher URL components, ensuring a clear approach to extracting relevant information.

preprints.org/manuscript/202410.1204/v1

Version 1: Received: 14 October 2024 / Approved: 15 October 2024 / Online: 15 October 2024 (13:25:55 CEST)
Version 2: Received: 15 October 2024 / Approved: 16 October 2024 / Online: 16 October 2024 (10:52:58 CEST)

How to cite: Al-Batati, A. S.; Koubaa, A.; Abdelkader, M. ROS 2 Key Challenges and Advances: A Survey of ROS 2 Research, Libraries, and Applications. *Preprints* **2024**, 2024101204. <https://doi.org/10.20944/preprints202410.1204.v1> [\[CrossRef\]](#)

Abstract
This study presents a comprehensive systematic review that addresses the critical transition from ROS 1 to ROS 2, spotlighting the significant enhancements and the pressing need for a 2 detailed exploration of ROS 2 within the robotics community. Despite the extensive deployment 3 and adaptations of ROS in varied robotics applications, literature lacks a cohesive synthesis that 4 delineates the advancements, limitations, and broader impacts of ROS 2 compared to its predecessor, ROS 1. Our contribution bridges this gap by assembling the largest database of ROS-related research, 6 encompassing 7,498 articles, with a focused analysis in this survey on 431 ROS2-specific publications. 7 We categorize these into 13 articles that discuss and analyze core ROS 2 concepts, 8 articles that 9 propose frameworks or tools for ROS 2, and 10 articles utilizing ROS 2. Furthermore, we summarize 9 literature findings of ROS 2 challenges, advancements, and future direction in the fields of a.) security, 10 b.) real-time, c.) middleware, d.) embedded and distributed systems, e.) communication reliability 11 and QoS, and f.) multi-robot systems. The methodology involved meticulous data collection and 12 categorization from multiple databases, facilitating an in-depth online accessible resource. Results 13 underscore ROS2's enhancements in modularity, real-time capabilities, and security, extending its 14 applicability across various robotic platforms and industries. However, challenges in scalability and 15 reliability persist, signaling avenues for future enhancements. This review not only deepens the 16 understanding of ROS2's contributions but also charts a path for ongoing improvements in robotic 17 systems design.

Keywords
ROS; ROS 2; robotic operating system; modularity; real-time capabilities; security; multi-robot systems; literature review

Subject
Engineering, Control and Systems Engineering

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[Download PDF](#)

Comments (0)
We encourage comments and feedback from a broad range of readers. See [criteria for comments](#) and our [Diversity statement](#).

Leave a public comment

Send a private comment to the author(s)

All users must log in before leaving a comment

Related Articles

peer-review Articles

Feedback

Abstract

Keywords

Subject

Copyright:

Download PDF

Comments (0)

Related Articles

Abstract

Keywords

Subject

Copyright:

Download PDF

Comments (0)

Related Articles

Elements

Console

Sources

Network

Performance

Memory

Application

Security

Styles

Computed

Layout

EventListeners

HTML

JS

CSS

JSON

XML

SVG

WebGL

WebAssembly

WebGPU

WebCodecs

WebMIDI

WebSerial

WebShare

WebStorage

WebUSB

WebVR

WebVox

WebXR

WebYJS

WebZ32

WebZ42

WebZ52

WebZ62

WebZ72

WebZ82

WebZ92

WebZ102

WebZ112

WebZ122

WebZ132

WebZ142

WebZ152

WebZ162

WebZ172

WebZ182

WebZ192

WebZ202

WebZ212

WebZ222

WebZ232

WebZ242

WebZ252

WebZ262

WebZ272

WebZ282

WebZ292

WebZ302

WebZ312

WebZ322

WebZ332

WebZ342

WebZ352

WebZ362

WebZ372

WebZ382

WebZ392

WebZ402

WebZ412

WebZ422

WebZ432

WebZ442

WebZ452

WebZ462

WebZ472

WebZ482

WebZ492

WebZ502

WebZ512

WebZ522

WebZ532

WebZ542

WebZ552

WebZ562

WebZ572

WebZ582

WebZ592

WebZ602

WebZ612

WebZ622

WebZ632

WebZ642

WebZ652

WebZ662

WebZ672

WebZ682

WebZ692

WebZ702

WebZ712

WebZ722

WebZ732

WebZ742

WebZ752

WebZ762

WebZ772

WebZ782

WebZ792

WebZ802

WebZ812

WebZ822

WebZ832

WebZ842

WebZ852

WebZ862

WebZ872

WebZ882

WebZ892

WebZ902

WebZ912

WebZ922

WebZ932

WebZ942

WebZ952

WebZ962

WebZ972

WebZ982

WebZ992

WebZ1002

WebZ1012

WebZ1022

WebZ1032

WebZ1042

WebZ1052

WebZ1062

WebZ1072

WebZ1082

WebZ1092

WebZ1102

WebZ1112

WebZ1122

WebZ1132

WebZ1142

WebZ1152

WebZ1162

WebZ1172

WebZ1182

WebZ1192

WebZ1202

WebZ1212

WebZ1222

WebZ1232

WebZ1242

WebZ1252

WebZ1262

WebZ1272

WebZ1282

WebZ1292

WebZ1302

WebZ1312

WebZ1322

WebZ1332

WebZ1342

WebZ1352

WebZ1362

WebZ1372

WebZ1382

WebZ1392

WebZ1402

WebZ1412

WebZ1422

WebZ1432

WebZ1442

WebZ1452

WebZ1462

WebZ1472

WebZ1482

WebZ1492

WebZ1502

WebZ1512

WebZ1522

WebZ1532

WebZ1542

WebZ1552

WebZ1562

WebZ1572

WebZ1582

WebZ1592

WebZ1602

WebZ1612

WebZ1622

WebZ1632

WebZ1642

WebZ1652

WebZ1662

WebZ1672

WebZ1682

WebZ1692

WebZ1702

WebZ1712

WebZ1722

WebZ1732

WebZ1742

WebZ1752

WebZ1762

WebZ1772

WebZ1782

WebZ1792

WebZ1802

WebZ1812

WebZ1822

WebZ1832

WebZ1842

WebZ1852

WebZ1862

WebZ1872

WebZ1882

WebZ1892

WebZ1902

WebZ1912

WebZ1922

WebZ1932

WebZ1942

WebZ1952

WebZ1962

WebZ1972

WebZ1982

WebZ1992

WebZ2002

WebZ2012

WebZ2022

WebZ2032

WebZ2042

WebZ2052

WebZ2062

WebZ2072

WebZ2082

WebZ2092

WebZ2102

WebZ2112

WebZ2122

WebZ2132

WebZ2142

WebZ2152

WebZ2162

WebZ2172

WebZ2182

WebZ2192

WebZ2202

WebZ2212

WebZ2222

WebZ2232

WebZ2242

WebZ2252

WebZ2262

WebZ2272

WebZ2282

WebZ2292

WebZ2302

WebZ2312

WebZ2322

WebZ2332

WebZ2342

WebZ2352

WebZ2362

WebZ2372

WebZ2382

WebZ2392

WebZ2402

WebZ2412

WebZ2422

WebZ2432

WebZ2442

WebZ2452

WebZ2462

WebZ2472

WebZ2482

WebZ2492

WebZ2502

WebZ2512

WebZ2522

WebZ2532

WebZ2542

WebZ2552

WebZ2562

WebZ2572

WebZ2582

WebZ2592

WebZ2602

WebZ2612

WebZ2622

WebZ2632

WebZ2642

WebZ2652

WebZ2662

WebZ2672

WebZ2682

WebZ2692

WebZ2702

WebZ2712

WebZ2722

WebZ2732

WebZ2742

WebZ2752

WebZ2762

WebZ2772

WebZ2782

WebZ2792

WebZ2802

WebZ2812

WebZ2822

WebZ2832

WebZ2842

WebZ2852

WebZ2862

WebZ2872

WebZ2882

WebZ2892

WebZ2902

WebZ2912

WebZ2922

WebZ2932

WebZ2942

WebZ2952

WebZ2962

WebZ2972

WebZ2982

WebZ2992

WebZ3002

WebZ3012

WebZ3022

WebZ3032

WebZ3042

WebZ3052

WebZ3062

WebZ3072

WebZ3082

WebZ3092

WebZ3102

WebZ3112

WebZ3122

WebZ3132

WebZ3142

WebZ3152

WebZ3162

WebZ3172

WebZ3182

WebZ3192

WebZ3202

WebZ3212

WebZ3222

WebZ3232

WebZ3242

WebZ3252

WebZ3262

WebZ3272

WebZ3282

WebZ3292

WebZ3302

WebZ3312

WebZ3322

WebZ3332

WebZ3342

WebZ3352

WebZ3362

WebZ3372

WebZ3382

WebZ3392

WebZ3402

WebZ3412

WebZ3422

WebZ3432

WebZ3442

WebZ3452

WebZ3462

WebZ3472

WebZ3482

WebZ3492

WebZ3502

WebZ3512

WebZ3522

WebZ3532

WebZ3542

WebZ3552

WebZ3562

WebZ3572

WebZ3582

WebZ3592

WebZ3602

WebZ3612

WebZ3622

WebZ3632

WebZ3642

WebZ3652

WebZ3662

WebZ3672

WebZ3682

WebZ3692

WebZ3702

WebZ3712

WebZ3722

WebZ3732

WebZ3742

WebZ3752

WebZ3762

WebZ3772

WebZ3782

WebZ3792

WebZ3802

WebZ3812

WebZ3822

WebZ3832

WebZ3842

WebZ3852

WebZ3862

WebZ3872

WebZ3882

WebZ3892

WebZ3902

WebZ3912

WebZ3922

WebZ3932

WebZ3942

WebZ3952

WebZ3962

WebZ3972

WebZ3982

WebZ3992

WebZ4002

WebZ4012

WebZ4022

WebZ4032

WebZ4042

WebZ4052

WebZ4062

WebZ4072

WebZ4082

WebZ4092

WebZ4102

WebZ4112

WebZ4122

WebZ4132

WebZ4142

WebZ4152

WebZ4162

WebZ4172

WebZ4182

WebZ4192

WebZ4202

WebZ4212

WebZ4222

WebZ4232

WebZ4242

WebZ4252

WebZ4262

WebZ4272

WebZ4282

WebZ4292

WebZ4302

WebZ4312

WebZ4322

WebZ4332

WebZ4342

WebZ4352

WebZ4362

WebZ4372

WebZ4382

WebZ4392

WebZ4402

WebZ4412

WebZ4422

WebZ4432

WebZ4442

WebZ4452

WebZ4462

WebZ4472

WebZ4482

WebZ4492

WebZ4502

WebZ4512

WebZ4522

WebZ4532

WebZ4542

WebZ4552

WebZ4562

WebZ4572

WebZ4582

WebZ4592

WebZ4602

WebZ4612

WebZ4622

WebZ4632

WebZ4642

WebZ4652

WebZ4662

WebZ4672

WebZ4682

WebZ4692

WebZ4702

WebZ4712

WebZ4722

WebZ4732

WebZ4742

WebZ4752

WebZ4762

WebZ4772

WebZ4782

WebZ4792

WebZ4802

WebZ4812

WebZ4822

WebZ4832

WebZ4842

WebZ4852

WebZ4862

WebZ4872

WebZ4882

WebZ4892

WebZ4902

WebZ4912

WebZ4922

WebZ4932

WebZ4942

WebZ4952

WebZ4962

WebZ4972

WebZ4982

WebZ4992

WebZ5002

WebZ5012

WebZ5022

WebZ5032

WebZ5042

WebZ5052

WebZ5062

WebZ5072

WebZ5082

WebZ5092

WebZ5102

WebZ5112

WebZ5122

WebZ5132

WebZ5142

WebZ5152

WebZ5162

WebZ5172

WebZ5182

WebZ5192

WebZ5202

WebZ5212

WebZ5222

WebZ5232

WebZ5242

WebZ5252

WebZ5262

WebZ5272

WebZ5282

WebZ5292

WebZ5302

WebZ5312

WebZ5322

WebZ5332

WebZ5342

WebZ5352

WebZ5362

WebZ5372

WebZ5382

WebZ5392

WebZ5402

WebZ5412

WebZ5422

WebZ5432

WebZ5442

WebZ5452

WebZ5462

WebZ5472

WebZ5482

WebZ5492

WebZ5502

WebZ5512

WebZ5522

WebZ5532

WebZ5542

WebZ5552

WebZ5562

WebZ5572

WebZ5582

WebZ5592

WebZ5602

WebZ5612

WebZ5622

WebZ5632

WebZ5642

WebZ5652

WebZ5662

WebZ5672

WebZ5682

WebZ5692

WebZ5702

WebZ5712

WebZ5722

WebZ5732

WebZ5742

WebZ5752

WebZ5762

WebZ5772

WebZ5782

WebZ5792

WebZ5802

WebZ5812

WebZ5822

WebZ5832

WebZ5842

WebZ5852

WebZ5862

WebZ5872

WebZ5882

WebZ5892

WebZ5902

WebZ5912

WebZ5922

WebZ5932

WebZ5942

WebZ5952

WebZ5962

WebZ5972

WebZ5982

WebZ5992

WebZ6002

WebZ6012

WebZ6022

WebZ6032

WebZ6042

WebZ6052

WebZ6062

WebZ6072

WebZ6082

WebZ6092

WebZ6102

WebZ6112

WebZ6122

WebZ6132

WebZ6142

WebZ6152

WebZ6162

WebZ6172

WebZ6182

WebZ6192

WebZ6202

WebZ6212

WebZ6222

WebZ6232

WebZ6242

WebZ6252

WebZ6262

WebZ6272

WebZ6282

WebZ6292

WebZ6302

WebZ6312

WebZ6322

WebZ6332

WebZ6342

WebZ6352

WebZ6362

WebZ6372

WebZ6382

WebZ6392

WebZ6402

WebZ6412

WebZ6422

WebZ6432

WebZ6442

WebZ6452

WebZ6462

WebZ6472

WebZ6482

WebZ6492

WebZ6502

WebZ6512

WebZ6522

WebZ6532

WebZ6542

WebZ6552

WebZ6562

WebZ6572

WebZ6582

WebZ6592

WebZ6602

WebZ6612

WebZ6622

WebZ6632

WebZ6642

WebZ6652

WebZ6662

WebZ6672

WebZ6682

WebZ6692

WebZ6702

WebZ6712

WebZ6722

WebZ6732

WebZ6742

WebZ6752

WebZ6762

WebZ6772

WebZ6782

WebZ6792

WebZ6802

WebZ6812

WebZ6822

WebZ6832

WebZ6842

WebZ6852

WebZ6862

WebZ6872

WebZ6882

WebZ6892

WebZ6902

WebZ6912

WebZ6922

WebZ6932

WebZ6942

WebZ6952

WebZ6962

WebZ6972

WebZ6982

WebZ6992

WebZ7002

WebZ7012

WebZ7022

WebZ7032

WebZ7042

WebZ7052

WebZ7062

WebZ7072

WebZ7082

WebZ7092

WebZ7102

WebZ7112

WebZ7122

WebZ7132

WebZ7142

WebZ7152

WebZ7162

WebZ7172

WebZ7182

WebZ7192

WebZ7202

WebZ7212

WebZ7222

WebZ7232

WebZ7242

WebZ7252

WebZ7262

WebZ7272

WebZ7282

WebZ7292

WebZ7302

WebZ7312

WebZ7322

WebZ7332

WebZ7342

WebZ7352

WebZ7362

WebZ7372

WebZ7382

WebZ7392

WebZ7402

WebZ7412

WebZ7422

WebZ7432

WebZ7442

WebZ7452

WebZ7462

WebZ7472

WebZ7482

WebZ7492

WebZ7502

WebZ7512

WebZ7522

WebZ7532

WebZ7542

WebZ7552

WebZ7562

WebZ7572

WebZ7582

WebZ7592

WebZ7602

WebZ7612

WebZ7622

WebZ7632

WebZ7642

WebZ7652

WebZ7662

WebZ7672

WebZ7682

WebZ7692

WebZ7702

WebZ7712

WebZ7722

WebZ7732

WebZ7742

WebZ7752

WebZ7762

WebZ7772

WebZ7782

WebZ7792

WebZ7802

WebZ7812

WebZ7822

WebZ7832

WebZ7842

WebZ7852

WebZ7862

WebZ7872

WebZ7882

WebZ7892

WebZ7902

WebZ7912

WebZ7922

WebZ7932

WebZ7942

WebZ7952

WebZ7962

WebZ7972

WebZ7982

WebZ7992

WebZ8002

WebZ8012

WebZ8022

WebZ8032

WebZ8042

WebZ8052

WebZ8062

WebZ8072

WebZ8082

WebZ8092

WebZ8102

WebZ8112

WebZ8122

WebZ8132

WebZ8142

WebZ8152

WebZ8162

WebZ8172

WebZ8182

WebZ8192

WebZ8202

WebZ8212

WebZ8222

WebZ8232

WebZ8242

WebZ8252

WebZ8262

WebZ8272

WebZ8282

WebZ8292

WebZ8302

WebZ8312

WebZ8322

WebZ8332

WebZ8342

WebZ8352

WebZ8362

WebZ8372

WebZ8382

WebZ8392

WebZ8402

WebZ8412

WebZ8422

WebZ8432

WebZ8442

WebZ8452

WebZ8462

WebZ8472

WebZ8482

WebZ8492

WebZ8502

WebZ8512

WebZ8522

WebZ8532

WebZ8542

WebZ8552

WebZ8562

WebZ8572

WebZ8582

WebZ8592

WebZ8602

WebZ8612

WebZ8622

WebZ8632

WebZ8642

WebZ8652

WebZ8662

WebZ8672

WebZ8682

WebZ8692

WebZ8702

WebZ8712

WebZ8722

WebZ8732

WebZ8742

WebZ8752

WebZ8762

WebZ8772

WebZ8782

WebZ8792

WebZ8802

WebZ8812

WebZ8822

WebZ8832

WebZ8842

WebZ8852

WebZ8862

WebZ8872

WebZ8882

WebZ8892

WebZ8902

WebZ8912

WebZ8922

WebZ8932

WebZ8942

WebZ8952

WebZ8962

WebZ8972

WebZ8982

WebZ8992

WebZ9002

WebZ9012

WebZ9022

WebZ9032

WebZ9042

WebZ9052

WebZ9062

WebZ9072

WebZ9082

WebZ9092

WebZ9102

WebZ9112

WebZ9122

WebZ9132

WebZ9142

WebZ9152

WebZ9162

WebZ9172

WebZ9182

WebZ9192

WebZ9202

WebZ9212

WebZ9222

WebZ9232

WebZ9242

WebZ9252

WebZ9262

WebZ9272

WebZ9282

WebZ9292

WebZ9302

WebZ9312

WebZ9322

WebZ9332

WebZ9342

WebZ9352

WebZ9362

WebZ9372

WebZ9382

WebZ9392

WebZ9402

WebZ9412

WebZ9422

WebZ9432

WebZ9442

WebZ9452

WebZ9462

WebZ9472

WebZ9482

WebZ9492

WebZ9502

WebZ9512

WebZ9522

WebZ9532

WebZ9542

WebZ9552

WebZ9562

WebZ9572

WebZ9582

WebZ9592

WebZ9602

WebZ9612

WebZ9622

WebZ9632

WebZ9642

WebZ9652

WebZ9662

WebZ9672

WebZ9682

WebZ9692

WebZ9702

WebZ9712

WebZ9722

WebZ9732

WebZ9742

WebZ9752

WebZ9762

WebZ9772

WebZ9782

WebZ9792

WebZ9802

WebZ9812

WebZ9822

WebZ9832

WebZ9842

WebZ9852

WebZ9862

WebZ9872

WebZ9882

WebZ9892

WebZ9902

WebZ9912

WebZ9922

WebZ9932

WebZ9942

WebZ9952

WebZ9962

WebZ9972

WebZ9982

WebZ9992

WebZ10002

WebZ10012

WebZ10022

WebZ10032

WebZ10042

WebZ10052

WebZ10062

WebZ10072

WebZ10082

WebZ10092

WebZ10102

WebZ1011

Some scraping knowledge (3)

- Analyze a website:
 - **For example using firebug in Firefox or Web developer extensions in Chrome**
- Keep an eye on the format of a hyperlink:
 - **Fictitious example:**

`http://www.example.com/getdata?subject=books&display=label,price`



- **The parameters may be useful for scraping**

Query parameters in a URL consist of three main components:

1. **Start symbol:** A question mark (?) denotes the beginning of the query parameters.
2. **Information pairs:** Key-value pairs joined by an equal sign (key=value) hold the information.
3. **Separator:** Multiple query parameters are separated by an ampersand symbol (&).

Beautiful Soup: Web Scraping with Python

Why Use Python for Web Scraping?

- **Readability:** Python's syntax is clean and easy to understand.
- **Versatility:** It's a general-purpose language with a vast ecosystem of libraries.
- **Powerful Libraries:**
 - **requests:** Simplifies HTTP requests.
 - **Beautiful Soup:** Parses HTML and XML documents.

Overview of Python Libraries

- Python provides several robust libraries for web scraping.
- Two essential tools are **Beautiful Soup** for parsing HTML and the **Requests** library for handling HTTP requests.
- **Requests** simplifies the process of making HTTP requests, while **Beautiful Soup** excels at parsing and manipulating HTML data.
- Together, they form a robust ecosystem for extracting data from websites efficiently.

Introducing BeautifulSoup

- **Definition:** BeautifulSoup is a Python library for parsing HTML and XML documents.
- Parsers like BeautifulSoup help in interpreting the raw HTML content extracted.
- By targeting specific tags, class names, and IDs, users can isolate pertinent data and simplify the process of extracting relevant information from the markup.
- **Key Features:**
 - Navigates HTML structure: Easily finds elements by tags, IDs, classes, etc.
 - Extracts data: Retrieves text, attributes, and other information.
 - Handles malformed HTML: Robustly handles errors and inconsistencies.

Introducing Requests

- The ***Requests*** library simplifies the process of sending HTTP requests in Python.
- It allows users to make GET and POST requests, handle response data, and manage sessions seamlessly.
- This library eliminates the complexities of standard Python modules, streamlining the fetching of web content.

Setting Up the Environment

- **Install Necessary Libraries:**

```
pip install requests beautifulsoup4
```

- **Import Libraries:**

```
import requests  
from bs4 import BeautifulSoup
```

Fetching the Webpage

Use `requests.get()`:

```
url = "https://www.example.com"  
response = requests.get(url)
```

Replace "https://www.example.com" with the actual URL you want to scrape.

Check for Success:

```
if response.status_code == 200:  
    print("Page fetched successfully")  
else:  
    print("Error fetching page")
```

Creating a BeautifulSoup Object

- Parse the HTML:

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

- Explanation:
 - `response.content`: The HTML content of the fetched page.
 - `"html.parser"`: The parser to use for parsing the HTML.

Navigating the HTML Structure

- Inspect Element: Use your browser's developer tools to examine the HTML structure.
- Find Elements:
 - By ID: `soup.find(id="element_id")`
 - By Class: `soup.find_all(class_="element_class")`
 - By Tag: `soup.find_all("tag_name")`

Adjust the selectors (e.g., id, class) based on the target elements on the webpage.

Extracting Data

- Access Text: `element.text`
- Access Attributes: `element["attribute_name"]`
- Example:

```
title = soup.find("h1").text
price = soup.find("span", class_="price").text
print("Title:", title)
print("Price:", price)
```

Putting it All Together

```
import requests
from bs4 import BeautifulSoup

url = "https://www.example.com/product"
response = requests.get(url)

if response.status_code == 200:
    soup = BeautifulSoup(response.content, "html.parser")

    title = soup.find("h1").text
    price = soup.find("span", class_="price").text

    print("Title:", title)
    print("Price:", price)
else:
    print("Error fetching page")
```

What to Consider When Scraping

- Respect website guidelines for scraping.
- Rate Limiting: Avoid overwhelming servers with excessive requests.
- Legal and Ethical Considerations: Ensure your scraping practices comply with laws and regulations.

Alternatives to Web Scraping: APIs (1)

- Some websites provide Application Programming Interfaces (APIs) as a stable method for data access.
- APIs are designed for programmatic interactions, enabling users to request data in structured formats like JSON, minimizing parsing complexities associated with HTML.



Alternatives to Web Scraping: APIs (2)

- The front-end presentation of a site might change often, but such a change in the website's design doesn't affect its API structure.
- The structure of an API is usually more permanent, which means it's a more reliable source of the site's data.

Conclusion

- Web scraping is a versatile tool for automating data extraction from websites.
- There are many tools available for web scraping, each with different features (Scrapy, Selenium, Puppeteer,...)
- Beautiful Soup is a powerful library for web scraping.
- It simplifies the process of navigating and extracting data.
- Be mindful of ethical considerations and website guidelines.
- Encouragement: Continue exploring web scraping and experiment with different techniques.