PyLadies

Vienna 16.01.2021

Who?

International mentorship group with a focus on helping more women become active participants and leaders in the Python open-source community.

Our mission is to promote, educate and advance a diverse Python community through outreach, education, conferences, events and social gatherings.

Agenda for today

---

1. Web scraping with Python!
2. Tools and libraries for web scraping
3. How to solve common problems during web scraping.

Goals

- Learn basic of Web Scraping
- What can you do with it and what are the limitations
- Build your own scraper

What is Web Scraping

- Get information from websites, structured and unstructured
- Collection and parsing data
- Not every site is allowing scraping with automated tools and it is protecting content in various ways

Why is it useful?

- Internet is full of information and not every website/webservice has an accessible API
- Sometimes there is no better way to get the data
- But you should always respect website’s term of use and not create too large web traffic

Web scraping vs web crawling

- Web scraping
  - automated information extraction from web
  - Main goal is to get one or more specific information and store it
- Web crawling
  - bot, which “crawls” systematically through web
  - they are called spiders
  - for example get all the movies from USA and just store everything you know
  - Working in organized way - page indexing and keeping everything for later processing

How to do it? What do I need?

- Whole process can be splitted into two steps:
  - **Get (HTML) content of the desired page**
  - **Parse raw data into usable structure**
- It it useful to know a bit about HTML to identify page structure
- Python has awesome libraries to make it easy

First request

- using library requests

```
import requests
movie_url = "https://www.imdb.com/title/tt2467372/"
r = requests.get(url=movie_url)
r.status_code
r.text
r.json()
```

Requests library

- Awesome library for human beings
- In the request you can pass arguments, headers, set encoding
- Even open eg image content → `i = Image.open(BytesIO(r.content))`
- Support all standard methods - GET, POST, DELETE, PUT

Beautiful soup

- `from bs4 import BeautifulSoup`
- Not needed, you can use plain string operations

  → but will make it much easier!

- first, create a **soup** object from request text:
- `soup = BeautifulSoup(r.text, 'html.parser')`
- `soup.get_text()` - will get you whole text
- also check out: `soup.prettify()`

Beautiful soup 2

- quick start:
  - soup.title
  - soup.title.string
  - soup.find_all('a')
- many predefined methods:
  - soup.find()
  - soup.find_all()
  - soup.replace()
  - → combine them together
  - ratingValue = soup.find("span", {"itemprop" : "ratingValue"})

Automate - when we need a lot of data

- again - lot of libraries
- selenium for python
- install library - `pip install selenium`
- download and install **webdriver** for your browser
- import library

Automate

- Initiaze the driver:
  - DRIVER = 'chromedriver'
  - driver = webdriver.Chrome(DRIVER)
- Pass the url and define elements you want to find
- elements = driver.find_elements_by_class_name("fa-chevron-double-right")
- You can find element name in the “inspect” mode

Problems - and how to solve them

- one of the simplest protection from using robots - captcha.
  - Did you know? Completely Automated Public Turing test to tell Computers and Humans Apart
  - How to solve? Load image using PIL library and then using OCR get the text
  - If its “select images with.. “ type, you will need to perform image classification

Another problems

- site starts to refuse connection for excessive traffic

1. Rotate agents:
    - agent - A user agent is a string that a browser or application sends to each website you visit. A typical user agent string contains details like - the application type, operating system, software vendor, or software version of the requesting software user agent.
    - information is stored in header of the request

Agent rotation

- `headers = {"User-Agent":"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.97 Safari/537.36"}`
- `r = requests.get('http://httpbin.org/headers', headers=headers)`
- Why to rotate them? Target website then might not recognize that all requests are coming from single source

Proxies

---

2. Rotate ip addresses and use proxies

```
proxies = {
  'http': 'http://10.10.1.10:3128',
  'https': 'http://10.10.1.10:1080',
}
requests.get('http://example.org', proxies=proxies)
```

- free and not free ip addresses: https://free-proxy-list.net
- easy to “scrape” to update proxies in your list

Proxies

- does not help you if you need to log in
- also will not work if you are using Sessions - they already know who are you
- Do not use ip in sequence - suspicious easily
- If you are using library for scraping, they could be using library for preventing you -> be innovative :)

Information not stored as HTML

- Sometimes the page info is generated by javascript after some action or just shown as a image, etc…
- Not a simple HTML we know how to parse
- If it’s possible, you can use selenium to perform action and get the resulting HTML and parse that
- Or you will need to process image - completely different topic
- Not easy and not a general solution

Tools and packages

---

scrapy - very good package for scraping and crawling

lxml - great library for content reading

many paid tools, if you don’t want to build your own (sometimes it’s not worth it)

used here - **BeautifulSoup**, **requests**, **selenium**

Small project - extract some data from imdb

---

guidelines:

1. get content of website with your favourite movie
2. extract information from that movie - year, title, director, rating, …
3. rewrite everything as functions
4. get same information about every movie and automate it

hint: first get all the movie urls, then iterate over them using selenium and download content using soup and call your extraction function on soup object

PyLadies complete beginners course

Weekly 2h from 2.2. to 4.5. - registrations are now closed

If you want to join as a side mentor/coach, please let us know right now, right here or sent an email to pylades.vienna@gmail.com

for coaching - only elementary Python knowledge needed

Resources and materials general

- advent of code - adventofcode.com
- hackerrank - hackerrank.com
- Django Girls django tutorial
- https://www.practicepython.org
- Nice Python exercices at one place https://github.com/tystar86/python_exercises
- https://automatetheboringstuff.com
- https://diveintopython3.problemsolving.io

Next topics

Chatbot

Graphics

GUI

fill the form regarding your interests please :) → https://forms.gle/UtfgVGe6AhhRwx539

Thank you and see you next time

Coding session - **28.1.2021**

Next workshop - **13.2.2021** Data Analysis!