basic-web-scraping

September 7, 2022

1 Web-scraping Example

Source: https://realpython.com/python-web-scraping-practical-introduction/

Web scraping is the process of collecting and parsing raw data from the Web, and the Python community has come up with some pretty powerful web scraping tools.

1.0.1 Getting the HTML from a URL

Collecting data from websites using an automated process is known as web scraping. Some websites explicitly forbid users from scraping their data with automated tools like the ones you'll create in this tutorial. Websites do this for two possible reasons:

- The site has a good reason to protect its data. For instance, Google Maps doesn't let you request too many results too quickly.
- Making many repeated requests to a website's server may use up bandwidth, slowing down the website for other users and potentially overloading the server such that the website stops responding entirely.

Important: Before using your Python skills for web scraping, you should always check your target website's acceptable use policy to see if accessing the website with automated tools is a violation of its terms of use. Legally, web scraping against the wishes of a website is very much a gray area.

Please be aware that the following techniques may be illegal when used on websites that prohibit web scraping.

[ ]:

```python
from urllib.request import urlopen
url = "https://en.wikipedia.org/wiki/Madagascar_grebe"

#urlopen() returns an HTTPResponse object:
page = urlopen(url)
print(page, type(page))

#returns a sequence of bytes
html_bytes = page.read()

#decode the bytes to a string using UTF-8:
html = html_bytes.decode("utf-8")

# print first 500 characters of HTML (more on cleaning this next module)
```

```
print(html[0:500])
```

```
<http.client.HTTPResponse object at 0x7fd1ef6d6e60> <class
'http.client.HTTPResponse'>
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>Madagascar grebe - Wikipedia</title>
<script>document.documentElement.className="client-js";RLCONF={"wgBreakFrames":f
alse,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefa
ultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","Ma
y","June","July","August","September","October","November","December"],"wgReques
tId":"fba8676c-6e65-415f-a012-c125ae015044
```

[ ]:

```
#find the index of certain words in the string
html.find("About")
```

[ ]:

```
51121
```