Data Organization and Retrieval

Web Scraping

By
Saurabh G.
IIT Jammu
MTech Information Security

Agenda

+ Intro to Web Scraping
+ Pre-requisites
+ Example in detail
+ Demonstration
+ Discussion

What is Web Scraping?

+ Web scraping is the technique which allows us to extract data from a website.

+ Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications.

Why Web Scraping ?

Some website make their data available for downloading.

Some provide API's to access data

But often time the required data is not made available through the above convenient options.

Hence we do Web Scraping.

Steps of Web Scraping

01

Requesting the content of a specific URL from the server

02

Storing the content that is Returned

03

Identifying the elements of the page to locate the required data

04

Extract the data, reformat data, store the data.

Prerequisite

+ Python
+ File handling concepts of python
+ HTML, CSS
+ Beautiful Soup

Example in detail

Request.get()

Here you are constructing a request object which will be sent off to a server to request data. Second, a Response object is generated once Requests gets a response back from the server.

Example:

r = requests.get(url = URL, params = PARAMS)

r = requests.get('https://en.wikipedia.org/wiki/Monty_Python')

User Agent

- A browser's user agent string (UA) helps identify which browser is being used, what version, and on which operating system.

- Essentially, a user agent is a way for a browser to say “Hi, I’m Mozilla Firefox on Windows” or “Hi, I’m Safari on an iPhone” to a web server.

- Ex. Mozilla/5.0 (Linux; {Android Version}; {Build Tag etc.})
- AppleWebKit/{WebKit Rev} (KHTML, like Gecko)
- Chrome/{Chrome Rev} Mobile Safari/{WebKit Rev}

Response

```
import requests

# Making a get request
response = requests.get('https://api.github.com')

# printing request text
print(response.text)
```

{"current_user_url":"https://api.github.com/user","current_user_authorizations_html_url":"https://github.com/settings/connections/applications{/client_id}","authorizations_url":"https://api.github.com/authorizations","code_search_url":"https://api.github.com/search/code?q={query}{&page,per_page,sort,order}","commit_search_url":"https://api.github.com/search/commits?q={query}{&page,per_page,sort,order}","emails_url":"https://api.github.com/user/emails","emojis_url":"https://api.github.com/emojis","events_url":"https://api.github.com/events","feeds_url":"https://api.github.com/feeds","followers_url":"https://api.github.com/user/followers","following_url":"https://api.github.com/user/following{/target}","gists_url":"https://api.github.com/gists{/gist_id}","hub_url":"https://api.github.com/hub","issue_search_url":"https://api.github.com/search/issues?q={query}{&page,per_page,sort,order}","issues_url":"https://api.github.com/issues","keys_url":"https://api.github.com/user/keys","label_search_url":"https://api.github.com/search/labels?q={query}&repository_id={repository_id}{&page,per_page}","notifications_url":"https://api.github.com/notifications","organization_url":"https://api.github.com/orgs/{org}","organization_repositories_url":"https://api.github.com/orgs/{org}/repos{?type,page,per_page,sort}","organization_teams_url":"https://api.github.com/orgs/{org}/teams","public_gists_url":"https://api.github.com/gists/public","rate_limit_url":"https://api.github.com/rate_limit","repository_url":"https://api.github.com/repos/{owner}/{repo}","repository_search_url":"https://api.github.com/search/repositories?q={query}{&page,per_page,sort,order}","current_user_repositories_url":"https://api.github.com/user/repos{?type,page,per_page,sort}","starred_url":"https://api.github.com/user/starred{/owner}{/repo}","starred_gists_url":"https://api.github.com/gists/starred","user_url":"https://api.github.com/users/{user}","user_organizations_url":"https://api.github.com/user/orgs","user_repositories_url":"https://api.github.com/users/{user}/repos{?type,page,per_page,sort}","user_search_url":"https://api.github.com/search/users?q={query}{&page,per_page,sort,order}"}

Demonstration

- **NOTE: Demonstration of the code was performed during the lab.**
- **The code used of demonstration can be found in the below github repository:**
- https://github.com/danielspg/WebSraping1

Thank you !