



EBook Gratis

APRENDIZAJE web-scraping

Free unaffiliated eBook created from
Stack Overflow contributors.

#web-
scraping

Tabla de contenido

| | |
|--|----------|
| Acerca de | 1 |
| Capítulo 1: Comenzando con el raspado web | 2 |
| Observaciones..... | 2 |
| Examples..... | 2 |
| Web Scraping en Python (usando BeautifulSoup)..... | 2 |
| Creditos | 4 |

Acerca de

You can share this PDF with anyone you feel could benefit from it, downloaded the latest version from: [web-scraping](#)

It is an unofficial and free web-scraping ebook created for educational purposes. All the content is extracted from [Stack Overflow Documentation](#), which is written by many hardworking individuals at Stack Overflow. It is neither affiliated with Stack Overflow nor official web-scraping.

The content is released under Creative Commons BY-SA, and the list of contributors to each chapter are provided in the credits section at the end of this book. Images may be copyright of their respective owners unless otherwise specified. All trademarks and registered trademarks are the property of their respective company owners.

Use the content presented in this book at your own risk; it is not guaranteed to be correct nor accurate, please send your feedback and corrections to info@zzzprojects.com

Capítulo 1: Comenzando con el raspado web

Observaciones

Esta sección proporciona una descripción general de qué es el raspado web y por qué un desarrollador puede querer usarlo.

También debe mencionar cualquier tema importante dentro del rastreo web y vincular a los temas relacionados. Dado que la Documentación para el web scraping es nueva, es posible que deba crear versiones iniciales de esos temas relacionados.

Examples

Web Scraping en Python (usando BeautifulSoup)

Al realizar tareas de ciencia de datos, es común querer usar datos encontrados en Internet. Por lo general, podrá acceder a estos datos a través de una interfaz de programación de aplicaciones (API) o en otros formatos. Sin embargo, hay ocasiones en que solo se puede acceder a los datos que desea como parte de una página web. En casos como este, una técnica llamada raspado web entra en escena.

Para aplicar esta técnica para obtener datos de las páginas web, necesitamos tener conocimientos básicos sobre la estructura de la página web y las etiquetas utilizadas en el desarrollo de páginas web (es decir, `<html>`, ``, `<div>`, etc.). Si eres nuevo en el desarrollo web puedes aprenderlo [aquí](#).

Así que para comenzar con el desguace web, usaremos un [sitio web](#) simple. Usaremos el módulo de `requests` para obtener el contenido de la página web o el código fuente.

```
import requests
page = requests.get("http://dataquestio.github.io/web-scraping-pages/simple.html")
print (page.content) ## shows the source code
```

Ahora usaremos el módulo `bs4` para desechar el contenido y obtener los datos útiles.

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(page.content, 'html.parser')
print(soup.prettify()) ##shows source in html format
```

Puede encontrar las etiquetas requeridas utilizando la herramienta `inspect element` en su navegador. Ahora digamos que desea obtener todos los datos que están almacenados con la etiqueta `` Luego, puede encontrarla con el script

```
soup.find_all('li')
# you can also find all the list items with class='ABC'
# soup.find_all('p', class_='ABC')
# OR all elements with class='ABC'
```

```
# soup.find_all(class_="ABC")
# OR all the elements with class='ABC'
# soup.find_all(id="XYZ")
```

Luego puede obtener el texto en la etiqueta usando

```
for i in range(len(soup.find_all('li'))):
    print (soup.find_all('li')[i].get_text())
```

Todo el gui3n es peque1o y bastante simple.

```
import requests
from bs4 import BeautifulSoup

page = requests.get("http://dataquestio.github.io/web-scraping-pages/simple.html") #get the
page
soup = BeautifulSoup(page.content, 'html.parser') # parse according to html
soup.find_all('li') #find required tags

for i in range(len(soup.find_all('li'))):
    print (soup.find_all('li')[i].get_text())
```

Lea Comenzando con el raspado web en l3nea: <https://riptutorial.com/es/web-scraping/topic/7746/comenzando-con-el-raspado-web>

Creditos

| S. No | Capítulos | Contributors |
|-------|-------------------------------|--|
| 1 | Comenzando con el raspado web | Community , thepurpleowl |