Web scraping

Nombre:Stiv Quishpe

Fecha: 29/12/2024

link de GitHub al respositorio del codigo fuente

https://github.com/stiv001/web-scraping.git

Realizar las siguientes actividades

- Revisar qué es web scraping
- Realizar una prueba en python para dos librerías diferentes
- Realizar scraping de un sitio web de su elección

¿Qué es web scraping ?

Web scraping es una técnica utilizada para extraer datos de sitios web. Este proceso implica el envío de solicitudes HTTP para obtener el contenido de una página web y luego analizar ese contenido para extraer información específica.

Realizar una prueba en python para dos librerías diferentes

BeautifulSoup

BeautifulSoup es una biblioteca de Python utilizada para analizar documentos HTML y XML, permitiendo extraer datos de páginas web de manera sencilla. Facilita la navegación, búsqueda y modificación del árbol de análisis del documento, que se genera a partir del contenido HTML o XML. BeautifulSoup convierte el contenido de la página en un árbol de objetos Python, lo que permite acceder y manipular elementos específicos del documento mediante métodos intuitivos. Es especialmente útil para tareas de web scraping, donde se necesita extraer información estructurada de sitios web. BeautifulSoup es compatible con varios analizadores, como el analizador HTML nativo de Python y lxml, ofreciendo flexibilidad en el procesamiento de

documentos. Además, maneja automáticamente las diferencias en la estructura del HTML, lo que lo hace robusto frente a errores comunes en el marcado de las páginas web.

```
import requests
from bs4 import BeautifulSoup

url = 'https://dockerlabs.es/'
respuesta = requests.get(url)

if respuesta.status_code == 200:
    soup = BeautifulSoup(respuesta.text, 'html.parser')

    maquinas =soup.find_all ('div',onclick= True)


    for maquina in maquinas:
        onclick_text = maquina['onclick']
        nombre_maquina = onclick_text.split("'")[1]
        print ('El nombre de la maquina es:',nombre_maquina )

else:
    print('A ocurrido un error:',respuesta.status_code)
```

```
El nombre de la maquina es: Psycho
El nombre de la maquina es: Dance-Samba
El nombre de la maquina es: PequeÃ±as-Mentirosas
El nombre de la maquina es: Veneno
El nombre de la maquina es: Grandma
El nombre de la maquina es: Apolos
El nombre de la maquina es: Report
El nombre de la maquina es: Swiss
El nombre de la maquina es: WhereIsMyWebShell
El nombre de la maquina es: Inclusion
El nombre de la maquina es: Collections
El nombre de la maquina es: Trust
El nombre de la maquina es: Crackoff
```

```
El nombre de la maquina es: Predictable (Muy DifÃcil)
El nombre de la maquina es: BreakMySSH
El nombre de la maquina es: NodeClimb
El nombre de la maquina es: Library
El nombre de la maquina es: ConsoleLog
El nombre de la maquina es: HackZones
El nombre de la maquina es: PingPong
El nombre de la maquina es: FirstHacking
El nombre de la maquina es: Reverse
El nombre de la maquina es: MyBB
El nombre de la maquina es: Hidden
El nombre de la maquina es: 404-not-found
El nombre de la maquina es: 0xc0ffee
El nombre de la maquina es: DebugMe
El nombre de la maquina es: Stranger
El nombre de la maquina es: Stack
El nombre de la maquina es: Vendetta
El nombre de la maquina es: BuscaLove
El nombre de la maquina es: UserSearch
El nombre de la maquina es: Vulnerame
El nombre de la maquina es: chmod-4755
El nombre de la maquina es: Lfi.elf
El nombre de la maquina es: Candy
El nombre de la maquina es: Verdejo
El nombre de la maquina es: HedgeHog
El nombre de la maquina es: BorazuwarahCTF
El nombre de la maquina es: Sender
El nombre de la maquina es: Insecure
El nombre de la maquina es: Upload
El nombre de la maquina es: Domain
El nombre de la maquina es: DevTools
El nombre de la maquina es: Move
El nombre de la maquina es: Database
El nombre de la maquina es: Vacaciones
El nombre de la maquina es: Dark
El nombre de la maquina es: Rubiks
El nombre de la maquina es: Forgotten_Portal
El nombre de la maquina es: Force
El nombre de la maquina es: SecretJenkins
El nombre de la maquina es: HiddenCat
El nombre de la maquina es: Fileception
El nombre de la maquina es: HackMeDaddy
El nombre de la maquina es: Backend
```

```
El nombre de la maquina es: Vulnvault
El nombre de la maquina es: Allien
El nombre de la maquina es: HereBash
El nombre de la maquina es: Subversion (Beta)
El nombre de la maquina es: Reflection
El nombre de la maquina es: Memesploit
El nombre de la maquina es: WalkingCMS
El nombre de la maquina es: Mirame
El nombre de la maquina es: Rutas
El nombre de la maquina es: NorC
El nombre de la maquina es: ChatMe
El nombre de la maquina es: ShowTime
El nombre de la maquina es: Pinguinazo
El nombre de la maquina es: Balulero
El nombre de la maquina es: Inj3ct0rs
El nombre de la maquina es: ChocolateLovers
El nombre de la maquina es: Sites
El nombre de la maquina es: Redirection
El nombre de la maquina es: r00tless
El nombre de la maquina es: -Pn
El nombre de la maquina es: BruteShock
El nombre de la maquina es: Escolares
El nombre de la maquina es: AnonymousPingu
El nombre de la maquina es: File
El nombre de la maquina es: HackPenguin
El nombre de la maquina es: FindYourStyle
El nombre de la maquina es: Buffered
El nombre de la maquina es: Whoiam
El nombre de la maquina es: JenkHack
El nombre de la maquina es: Amor
El nombre de la maquina es: Darkweb
El nombre de la maquina es: Asucar
El nombre de la maquina es: Road to Olympus
El nombre de la maquina es: AguaDeMayo
El nombre de la maquina es: Master
El nombre de la maquina es: SkullNet
El nombre de la maquina es: Pressenter
El nombre de la maquina es: Los 40 Ladrones
El nombre de la maquina es: Picadilly
El nombre de la maquina es: Devil
El nombre de la maquina es: Winterfell
El nombre de la maquina es: Mapache2
El nombre de la maquina es: StrongJenkins
```

```
El nombre de la maquina es: HackTheHeaven
El nombre de la maquina es: Pntopntobarra
El nombre de la maquina es: PyRed
El nombre de la maquina es: Flow
El nombre de la maquina es: LittlePivoting
El nombre de la maquina es: DockHackLab
El nombre de la maquina es: Paradise
El nombre de la maquina es: SummerVibes
El nombre de la maquina es: Bashpariencias
El nombre de la maquina es: BigPivoting
El nombre de la maquina es: Corruptress
El nombre de la maquina es: Fooding
El nombre de la maquina es: Stellarjwt
El nombre de la maquina es: Queuemedic
El nombre de la maquina es: ChocolateFire
El nombre de la maquina es: DoubleTrouble
El nombre de la maquina es: Eclipse
El nombre de la maquina es: DockerLabs
El nombre de la maquina es: Wallet
El nombre de la maquina es: Cachopo
El nombre de la maquina es: Obsession
El nombre de la maquina es: DoubleFlow
El nombre de la maquina es: Elevator
El nombre de la maquina es: Cracker
```

scrapy

Scrapy es un framework de código abierto en Python diseñado para la extracción de datos de sitios web, también conocido como web scraping. Permite a los desarrolladores definir cómo navegar y extraer información de las páginas web de manera eficiente y estructurada. Scrapy utiliza spiders, que son clases definidas por el usuario que especifican cómo realizar las solicitudes HTTP y cómo procesar las respuestas. Además, Scrapy maneja automáticamente la gestión de solicitudes, el seguimiento de enlaces y la descarga de páginas, lo que facilita la recolección de grandes cantidades de datos. También incluye herramientas para la manipulación de datos extraídos, como la limpieza y el almacenamiento en diversos formatos. Su arquitectura basada en eventos permite realizar múltiples solicitudes en paralelo, lo que mejora significativamente la velocidad de scraping. Scrapy es ampliamente utilizado en proyectos de minería de datos, investigación y desarrollo de motores de búsqueda, y en cualquier aplicación que requiera la recolección de datos de la web.

```python
import scrapy
from scrapy.crawler import CrawlerProcess

class TitleSpider(scrapy.Spider):
    name = "title_spider"
    start_urls = ["https://www.significados.com/programacion/"]

    def parse(self, response):
        # Extraer y mostrar el título de la página
        title = response.css("title::text").get()
        print(f"Título de la página: {title}")


process = CrawlerProcess()
process.crawl(TitleSpider)
process.start()
```

Realizar scraping de un sitio web de su elección

```python
import requests
from bs4 import BeautifulSoup

url = "https://www.significados.com/programacion/"

response = requests.get(url)

if response.status_code == 200:
    soup = BeautifulSoup(response.content, "html.parser")
    title = soup.find("title").get_text()
    print(f"Titulo de la pagina: {title}")
else:
    print(f"Error al acceder a la pagina: {response.status_code}")
```

```
2024-12-29 22:13:14 [urllib3.connectionpool] DEBUG: Starting new HTTPS connection (1): www.si
2024-12-29 22:13:14 [urllib3.connectionpool] DEBUG: https://www.significados.com:443 "GET /pr

Titulo de la pagina: Programación: Qué es, Concepto y Definición - Enciclopedia Significados
```