Python Biella Group

Cos’è lo Scraping?

Estrarre dati da pagine web

- in formato strutturato
- quando i dati non sono forniti tramite API o con API limitate nel numero di richieste
- in modo (quasi sempre) anonimo
-

DISCLAIMER: non tutti i dati sono “gratuiti”, leggere sempre

- leggere sempre i termini di servizio
- sii responsabile
- non rubare

Use cases

Comparatori prezzi e-commerce

Dataset per ML

Informazioni Social Media

Dati meteorologici storici

....

Le 3 tecnologie standard nel Web

HTML

Definisce la struttura delle pagine Web

CSS

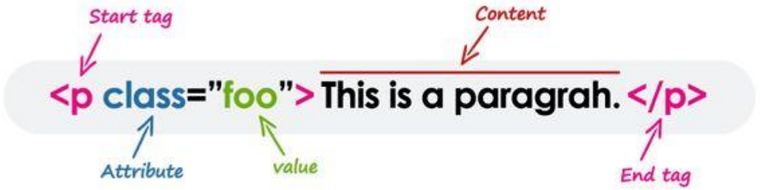
Un linguaggio di regole di stile che utilizziamo per applicare lo stile al nostro contenuto HTML, ad esempio, impostando i colori e i caratteri di sfondo e disponendo il nostro contenuto in più colonne.
All'interno della sintassi HTML, il tag di stile conterrà la sintassi per CSS

JavaScript

1) Linguaggio di scripting del browser Web
2) Modifica dinamicamente l'HTML (non possiamo recuperare dati da siti Web che modificano dinamicamente l'HTML originale. Beautifulsoup utilizza solo l'HTML originale)
3) Modifica dinamicamente i CSS
4) L'elemento HTML ha un tag di script per
    a) script interno, embeddato nella pagine
    c) script esterno

HTML

Sintassi di un elemento



Attributi

Gli attributi forniscono maggiori informazioni sul contenuto di un elemento

Esempio:

Google

L'attributo *href* indica che URL per l'elemento “Google” è https://www.google.com/ .

L'attributo *title* indica che “Google” è un motore di ricerca.

Attributi generici

Spesso cerchiamo di trovare un elemento specifico usando l'attributo **id** e **class**.
L'attributo id viene utilizzato per assegnare un nome o un identificatore univoco a un elemento all'interno di un documento.
Anche l'attributo class viene utilizzato per identificare gli elementi ma a differenza di id, l'attributo class non deve essere univoco nel documento.
Ciò significa che puoi applicare la stessa classe a più elementi in un documento.

Tools

Requests

Beautiful Soup
lxml

Selenium

Scrapy
Mechanize

Evita regex “puro” !

Python Biella Group

Scraping 101

Spiders
- Bot che scarica tutte le pagine

robots.txt
- File sul server che specifica i limiti di accesso per i bots

Python Biella Group

Regole in robots.txt

1. **Allow Full Access**

User-agent: *
Disallow:

Se lo trovi nel file robots.txt di un sito Web di cui stai tentando di eseguire la scansione, sei fortunato. Ciò significa che tutte le pagine del sito possono essere scansionate dai bot.

2. **Block All Access**

User-agent: *
Disallow: /

Dovresti stare alla larga da un sito con questo nel suo robots.txt. Afferma che nessuna parte del sito deve essere visitata utilizzando un crawler automatizzato e la violazione di ciò potrebbe comportare problemi legali.

3. **Partial Access**

User-agent: *
Disallow: /folder/

User-agent: *
Disallow: /file.html

Alcuni siti non consentono la scansione solo di sezioni o file particolari sul loro sito. In questi casi, dovresti indirizzare i tuoi robot a lasciare intatte le aree bloccate.

4. **Crawl Rate Limiting**

Crawl-delay: 11

Questo viene utilizzato per impedire ai crawler di colpire il sito troppo frequentemente. Poiché i frequenti accessi dei crawler potrebbero sottoporre il server a uno stress indesiderato e rallentare il sito per i visitatori umani, molti siti aggiungono questa riga nel proprio file robots. In questo caso, è possibile eseguire la scansione del sito con un ritardo di 11 secondi.

5. **Visit Time**

Visit-time: 0400-0845

Questo indica ai crawler le ore in cui è consentita la scansione. In questo esempio, è possibile eseguire la scansione del sito tra le 04:00 e le 08:45 UTC. I siti lo fanno per evitare il carico dei bot durante le ore di punta.

6. **Request Rate**

Request-rate: 1/10

Alcuni siti Web non intrattengono i robot che cercano di recuperare più pagine contemporaneamente. Il tasso di richiesta viene utilizzato per limitare questo comportamento. 1/10 come valore significa che il sito consente ai crawler di richiedere una pagina ogni 10 secondi.

Ostacoli

Javascripts
IP Blocking
Captchas
Login
Ad popups
CSS Sprites
Honeypots

I siti cambiano senza preavviso

Python Biella Group

La palestra degli scrapers

https://toscrape.com/

https://books.toscrape.com/
https://quotes.toscrape.com/