

Python: estrarre gli URL da una pagina web

In questo articolo vedremo come estrarre tutti gli URL dei link presenti in una pagina web con Python utilizzando il modulo BeautifulSoup.

Si tratta di fatto di individuare tutti gli elementi `a` ed ottenere il valore del loro attributo `href`.

```
from bs4 import BeautifulSoup

def get_page_urls(html):
    soup = BeautifulSoup(html, 'html.parser')
    links = soup.find_all('a')
    urls = []
    for link in links:
        href = link.get('href', '')
        if href.startswith('http'):
            urls.append(href)
    return urls
```

La nostra funzione restituirà una lista contenente gli URL dei link individuati.