Python: reperire i link interni di un sito da una pagina web

In questo articolo vedremo come reperire i link interni di una pagina web con Python.

Possiamo utilizzare il modulo di terze parti Beautiful Soup in questo modo:

```
from bs4 import BeautifulSoup

def get_internal_links(page_html, host_name):
    soup = BeautifulSoup(page_html, 'html.parser')
    a = soup.find_all('a')
    links = []
    needle = f'https://{host_name}'
    for link in a:
        href = link.get('href', '')
        if href.startswith(needle):
            links.append(href)
    return links
```

Se l'attributo `href` di un elemento HTML a nella pagina contiene l'URL di base del sito, il suo valore viene aggiunto alla lista dei link interni restituito dalla funzione.