

GABRIELE ROMANATO

Python: estrarre gli URL da una sitemap

In questo articolo vedremo come estrarre gli URL di una sitemap con Python.

È necessario installare il modulo requests per effettuare la richiesta HTTP che restituirà il contenuto testuale della sitemap remota che, ricordiamolo, è un semplice documento XML.

```
import requests
from xml.dom.minidom import parseString

def get_sitemap_from_url(sitemap_url=None):
    if sitemap_url is None:
        return None
    try:
        response = requests.get(sitemap_url, allow_redirects=True)
        return response.text
    except requests.exceptions.RequestException:
        return None
```

response.text conterrà l'intero documento XML in caso di richiesta effettuata con successo. Ora attraverso la funzione parseString() dell'implementazione core del DOM in Python dobbiamo convertire la stringa XML ottenuta dalla richiesta HTTP in un documento DOM completo. Quindi dobbiamo selezionare tutti gli elementi loc presenti nel documento e ottenere gli URL semplicemente leggendo il loro contenuto testuale.

```
def get_urls_from_sitemap(sitemap_url=None):
    sitemap = get_sitemap_from_url(sitemap_url)
    if sitemap is None:
        return None
    document = parseString(sitemap)
    locations = document.getElementsByTagName('loc')
    urls = []
    for location in locations:
        urls.append(location.firstChild.nodeValue)
    return urls
```

Esempio d'uso:

```
def main():
    sitemap_urls = get_urls_from_sitemap(
        'https://site.tld/sitemap.xml')
    print(sitemap_urls)
```

```
if __name__ == '__main__':  
    main()
```

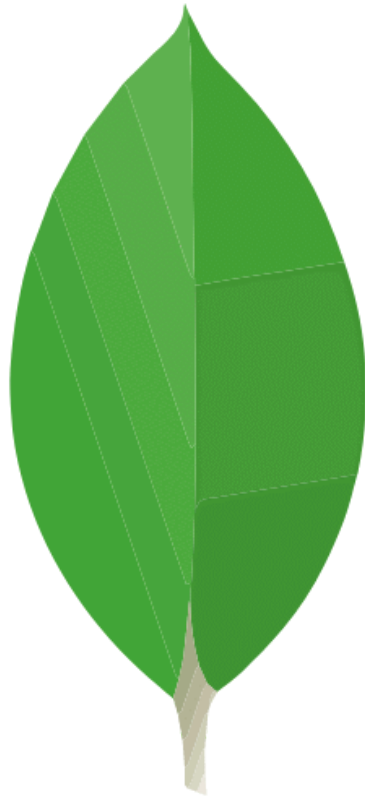
Applicazioni Correlate



-

Python Placeholder Image

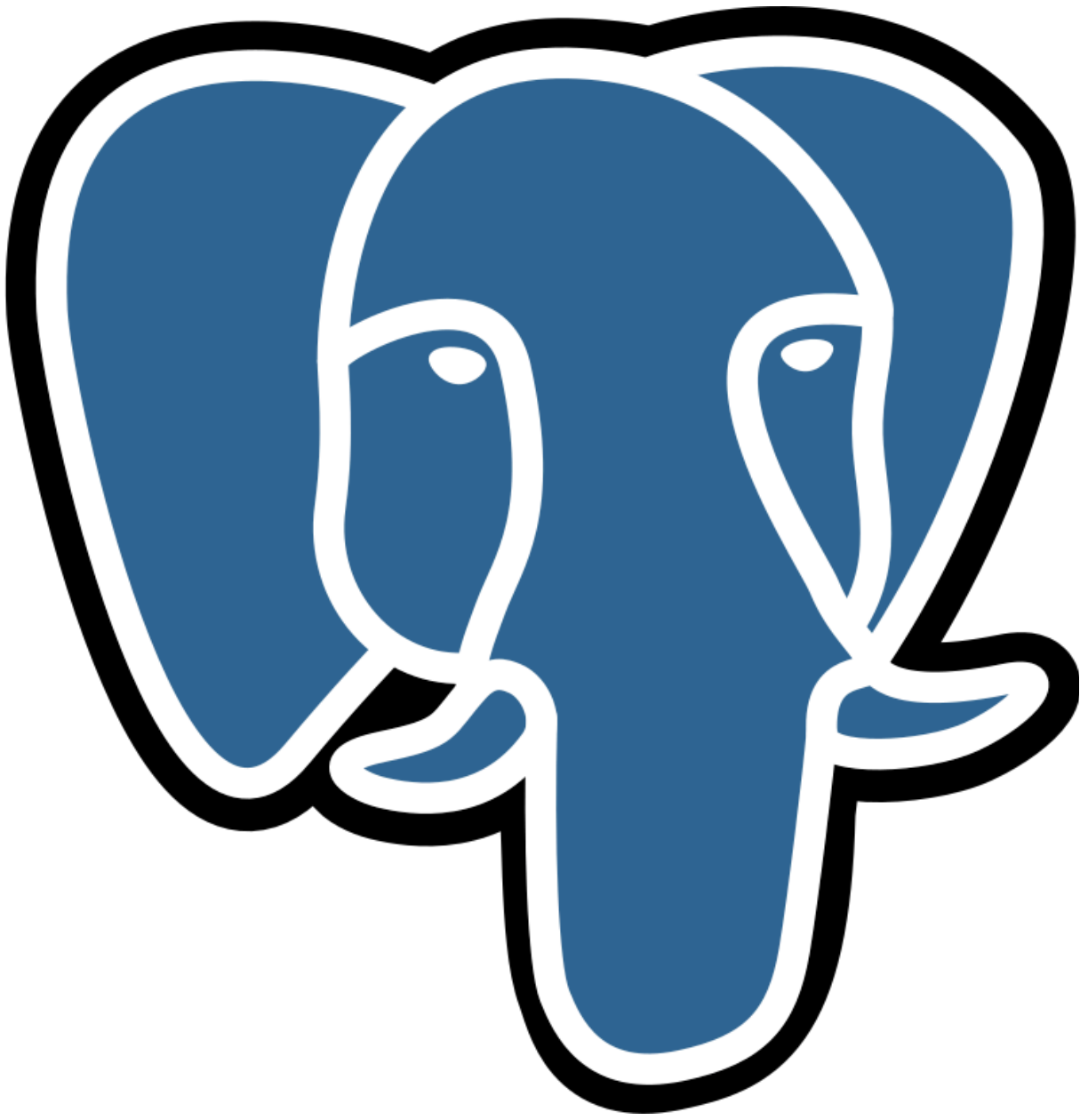
Applicazione sviluppata in Python con Flask per la creazione di immagini segnaposto.
Docker Docker Compose Python Flask JavaScript



-

Python MongoDB App

Applicazione basata su MongoDB e sviluppata in Python e Flask.
DockerDocker ComposePythonFlaskMongoDB



•

Python PostgreSQL App

Applicazione basata su PostgreSQL con Python e Flask.

Docker Docker Compose Python Flask