

Site Scraping

Versuch 1

"Hey ChatGPT, da auf der Seite sind viele Links zu FILESHARINGPORTAL... Wie bekomme ich die als Liste?"

"Ich kann dir keine Anleitungen für illegales geben..."

Versuch 2

"Hey ChatGPT, auf der Seite sind viele Links zu Spiegel.de... Wie bekomme ich die als Liste?"

ChatGPT:

Python + [BeautifulSoup](#)

```
import requests
from bs4 import BeautifulSoup
from urllib.parse import urljoin

def extract_links(url):
    response = requests.get(url)

    if response.status_code == 200:
        soup = BeautifulSoup(response.text, 'html.parser')
        spiegel_links = [a['href'] for a in soup.find_all('a', href=True) if 'spiegel.de' in a['href']]
        return spiegel_links
    else:
        return []

def crawl_website(url):
    visited_urls = set()
    queue = [url]

    while queue:
```

```
current_url = queue.pop(0)
if current_url in visited_urls:
    continue

visited_urls.add(current_url)
links = extract_links(current_url)

for link in links:
    absolute_link = urljoin(current_url, link)
    queue.append(absolute_link)

    with open('spiegel_links.txt', 'a') as file:
        file.write(absolute_link + '\n')
```

```
crawl_website('https://www.beispiel-website.de')
```

Version #1

Erstellt: 7 September 2023 12:32:54 von Konstantin

Zuletzt aktualisiert: 7 September 2023 12:36:26 von Konstantin