Python: how to check the URLs of a sitemap

The Google sitemap is an important tool to ensure that search engines crawl all pages of the website correctly. However, a sitemap does not guarantee that all URLs are up and running properly. In this article, we will explain how to check that all URLs are live using Python.

The first step is to import the necessary libraries. In particular, we will use the `requests` and `lxml` libraries to make HTTP requests and parse the XML sitemap.

```
import requests
from lxml import etree
```

After importing the libraries, we download the sitemap and parse the XML file using etree. We can do this using the following function:

```
def get_sitemap_urls(url):
    try:
        resp = requests.get(url)
        tree = etree.fromstring(resp.content)
        return [loc.text for loc in
tree.findall('{*}url/{*}loc')]
    except requests.exceptions.RequestException:
        return []
```

This function takes the URL of the sitemap as an argument and returns a list of all the URLs in the sitemap. The function uses the requests library to make the HTTP request to the sitemap and the `lxml` library to parse the XML content of the sitemap.

At this point, we can use the requests library to make an HTTP request to each sitemap URL and verify that the response status code is 200, indicating that the page is active. We can do this using the following code:

```
def check_urls(urls):
    for url in urls:
        resp = requests.get(url)
        if resp.status_code != 200:
            print(f"URL {url} is down (status code {resp.status_code})")
        else:
            print(f"OK {url}")
```

This function takes as argument the list of URLs returned by the `get_sitemap_urls()` function and uses the `requests` module to make an HTTP request to each URL. If the response status code is not 200, the function prints a message indicating that the URL is down, otherwise it informs that the request was successful.

```
sitemap_url = "https://nodejstutorial.it/sitemap.xml"
sitemap_urls = get_sitemap_urls(sitemap_url)

if len(sitemap_urls) > 0:
    check_urls(sitemap_urls)
else:
    print("Request error.")
```

In this example it is important to note that this solution only works for XML sitemaps. If you are using a sitemap in a different format, such as an HTML sitemap, you will need to adapt your code accordingly.

In conclusion, verifying that all URLs in the sitemap are active is an important step to ensure that search engines crawl all pages of the website correctly. Using Python and the `requests` and `lxml` libraries, you can automate this process and quickly verify that all URLs are up and running correctly.