スクレイビング(1)

現代の html 記述には一定のルールがある。そのルールが守られているサイトであれば、その構造を解析することで、html 内の任意の情報を機械的に抽出することができる。

◆Web ページの送受信に使われる HTTP 通信

スクレイビングを実行するためには、Python からサーバに HTTP のリクエストを送り、そのレスポンスから HTML を取得する。

■HTTP 通信

HTTP 通信を行うライブラリが「Requests」[1]。
Requests は Anaconda でインストールされている。

◆書籍ページのレスポンスを取得

```
#1
import requests
res = requests.get('https://gihyo.jp/book')
res.status_code
```

「200 番台」の数字が返されれば、リクエストが成功している。

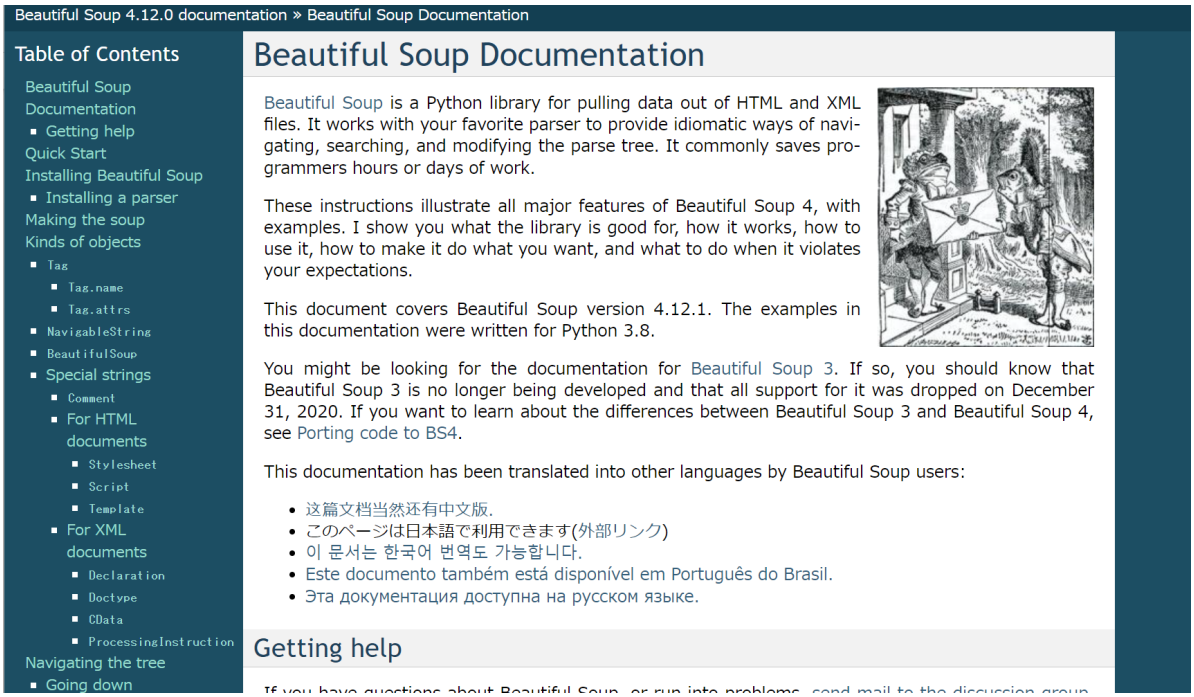
◆レスポンスから HTML を取得

```
#2
html_doc = res.text
print(html_doc[:300])
```

※実行には#1 が必要

[1] Python の標準ライブラリには「urllib.requests」がある

■HTML からデータを抽出 `Beautiful Soup`



Beautiful Soup 4.12.0 documentation » Beautiful Soup Documentation

Table of Contents

Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

This document covers Beautiful Soup version 4.12.1. The examples in this documentation were written for Python 3.8.

You might be looking for the documentation for Beautiful Soup 3. If so, you should know that Beautiful Soup 3 is no longer being developed and that all support for it was dropped on December 31, 2020. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see Porting code to BS4.

This documentation has been translated into other languages by Beautiful Soup users:

- 这篇文档当然还有中文版.
- このページは日本語で利用できます(外部リンク)
- 이 문서는 한국어 번역도 가능합니다.
- Este documento também está disponível em Português do Brasil.
- Эта документация доступна на русском языке.

Getting help

If you have questions about Beautiful Soup, or run into problems, send mail to the discussion group