# beautifulSoup4でスクレイピングする

| | |
|---|---|
| タグ | Python スクレイピング |
| 概要 | |




## ▼ BeautifulSoupとは

めっちゃ簡潔に、Pythonでスクレイピングをすることができる外部ライブラリ

## ▼ 事前準備

Python3がインストールされている環境であることを前提としている。

ターミナルで以下を入力

```
pip install beautifulsoup4
```

開発環境としてjupyter-labを使ったので、同じ環境を使ってみたい方を以下も（高専の時に書いた資料を移植したのでﾁｮｯﾄｷｮｳﾐｱﾙ人はこちらも）

```
pip install jupyter-lab
```

## ▼ まずはrequestを使ってみよう

beautifulsoupを使う前に、標準ライブラリのrequestを使用して、読売新聞のページHTMLを受け取ってみる

以下のコードを打ち込もう

```
# 先に必要なものをすべてインポートしておく
import request
import re

from bs4 import BeautifuSoup

# urlは自分が見てみたいサイトに適宜変更してみよう
url = 'https://www.yomiuri.co.jp/economy/20220511-OYT1T50121/'
res = requests.get(url)
```

▼ 結果

```
In [1]: res.text

Out[1]: '<!DOCTYPE html>\n<html lang="ja" >\n\n  <head>\n    <meta charset
="UTF-8">\n    <meta http-equiv="X-UA-Compatible" content="IE=Edg
e">\n    <meta name="viewport" content="width=device-width">\n    <
script>\n      (function(){\n        // screen adjustment\n
// ユーザーエージェントの取得\n        var ua = navigator.userAgen
t;\n        if(ua.indexOf(\'iPad\') > 0) document.querySelector("me
ta[name=\'viewport\']").setAttribute("content", "width=1076");\n
})();\n    </script>\n        <script src="/assets/js/pwa-202205091
73130.js"></script>\n    <title>アップルがｉＰｏｄ販売終了へ…２００
１年に初代発売、音楽業界に革命 : 読売新聞オンライン</title>\n<link r
el=\'dns-prefetch\' href=\'//widgets.outbrain.com\' />\n<link rel=
\'dns-prefetch\' href=\'//moviead.cdnext.stream.ne.jp\' />\n<link r
el=\'dns-prefetch\' href=\'//news.google.com\' />\n<link rel=\'dns-
prefetch\' href=\'//apis.google.com\' />\n<link rel=\'stylesheet\'
id=\'style-css\'  href=\'https://www.yomiuri.co.jp/assets/css/style
-20220509173130.css\'  media=\'all\' />\n<link rel=\'stylesheet\' i
d=\'sys-ajax-contents-css\'  href=\'https://www.yomiuri.co.jp/asset
s/css/sys-ajax-contents-20220509173130.css\'  media=\'all\' />\n<li
nk rel=\'stylesheet\' id=\'slick.css-css\'  href=\'https://www.yomi
uri.co.jp/assets/libs/slick/css/slick.css\'  media=\'all\' />\n<lin
k rel=\'stylesheet\' id=\'slick-theme.css-css\'  href=\'https://ww
w.yomiuri.co.jp/assets/libs/slick/css/slick-theme.css\'  media=\'al
l\' />\n<link rel=\'stylesheet\' id=\'slick-base.css-css\'  href=
\'https://www.yomiuri.co.jp/assets/libs/slick/css/slick-base.css\'
media=\'all\' />\n<link rel=\'stylesheet\' id=\'uni-scrap-css\'  hr
ef=\'https://www.yomiuri.co.jp/assets/css/uni-scrap-20220509173130.
css\'  media=\'all\' />\n<link rel=\'stylesheet\' id=\'responsive-l
ightbox-nivo-css\'  href=\'https://www.yomiuri.co.jp/wp-content/plu
gins/responsive-lightbox/assets/nivo/nivo-lightbox.min.css?ver=2.0.
4\'  media=\'all\' />\n<link rel=\'stylesheet\' id=\'responsive-lig
htbox-nivo-default-css\'  href=\'https://www.yomiuri.co.jp/wp-conte
nt/plugins/responsive-lightbox/assets/nivo/themes/default/default.c
ss?ver=2.0.4\'  media=\'all\' />\n<link rel=\'stylesheet\' id=\'yol
-gns-css\'  href=\'https://www.yomiuri.co.jp/assets/css/plugin-yol-
gns-20220509173130.css?ver=4.9.18\'  media=\'all\' />\n<link rel=
\'preload\' href=\'https://www.yomiuri.co.jp/assets/js/ajax-content
s-20220509173130.js\' as=\'script\' />\n<link rel=\'prefetch\' href
=\'https://www.yomiuri.co.jp/assets/js/ajax-contents-20220509173130
```

## ▼ bs4でHTML解析

### HTML解析とは

HTML文書を解析して、プログラム内で扱いやすい形にすること。

今回の場合はHTMLパーサーをつかって条件に合う属性のテキストを抽出する。

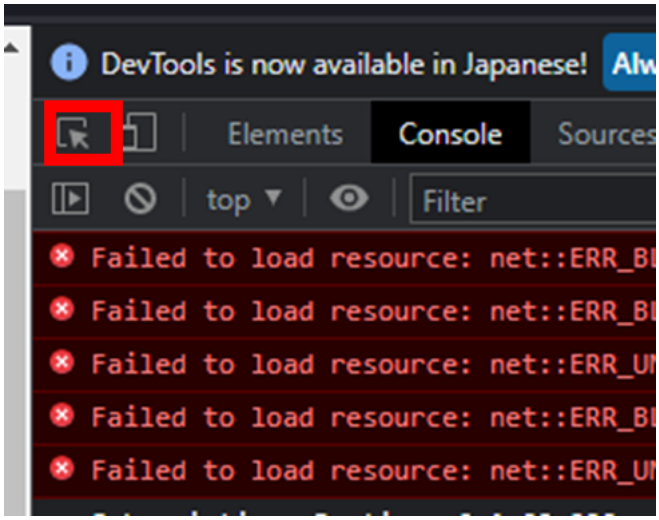
条件を見つけるためにCSSセレクタを使う。

### ▼ CSSセレクタ

CSS セレクターは、一連の CSS のルールが適用される要素を定義するもの。

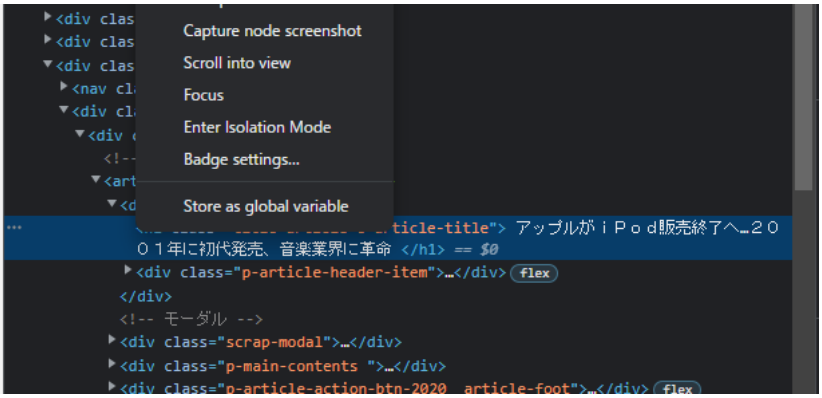
ブラウザのデベロッパーツールからニュースタイトルのセレクタをコピーしてみる

1. F12キーをおしてDevツールを押して、左上の矢印を押す

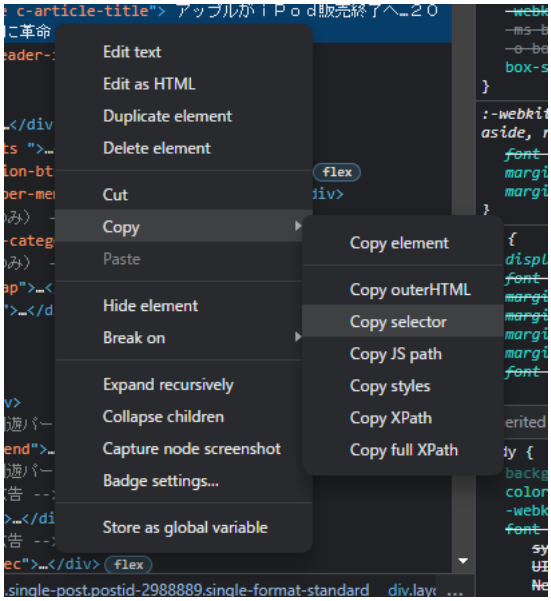



2. 記事のタイトルをクリックする


3. 該当箇所のHTMLが表示されるので、右クリック




4. `copy->copy selector`を選択




5. 試しに貼り付けてみるとこのようなものが出てくる

```
body > div.layout-contents > div.layout-contents__main > div.uni-scrap > article > div.article-header > h1
```

## ▼ タイトルの抽出

Devツールで抽出したCSSセレクタを使ってHTML解析をしていく。

今回は例としてタイトルを抽出してみる

```
# HTMLパーサーの定義
soup = BeautifulSoup(res.text, "html.parser")
element = soup.select('body > div.layout-contents > div.layout-contents__main > div.uni-scrap > article > div.article-header >

print(element[0].contents[0])
"""
出力：
'\n                アップル、ｉＰｏｄ販売終了へ…２００１年に初代発売、音楽業界に革命
"""
```

## ▼ find, find_allメソッド

`find()`

引数に一致する最初の一つの要素を取得する

`find_all()`

引数に一致するすべての要素を取得する

Yahooのニュースの記事のURLを習得してみる

```
url = 'https://www.yahoo.co.jp/'
res = requests.get(url)
# htmlパーサーの定義
soup = BeautifulSoup(res.text, "html.parser")
print(soup.find("a"))
# 正規表現で記事URLが入っている要素だけ取得
elems = soup.find_all(href=re.compile("news.yahoo.co.jp/pickup"))
for i in elems:
    print(i.attrs['href']) #リンクだけ出力
```

- 結果

```
3]: <a class="yMWCYupQNdgppL-NV6sMi _3sAlKGsIBCxTUbNi86oSjt" data-cl-pa
    rams="_cl_vmodule:header;_cl_link:logo;_cl_position:0" data-ylk="rs
    ec:header;slk:logo;pos:0" href="https://www.yahoo.co.jp">Yahoo! JAP
    AN</a>
```

```
1 for i in elems:
2     print(i.attrs['href'])

https://news.yahoo.co.jp/pickup/6426388
https://news.yahoo.co.jp/pickup/6426401
https://news.yahoo.co.jp/pickup/6426386
https://news.yahoo.co.jp/pickup/6426396
https://news.yahoo.co.jp/pickup/6426390
https://news.yahoo.co.jp/pickup/6426393
https://news.yahoo.co.jp/pickup/6426402
https://news.yahoo.co.jp/pickup/6426395
```

## ▼ クローリング

先ほど取得したニュース記事から更にニュース記事の内容を取得してみる

```
pickup_links = [elem.attrs['href'] for elem in elems]
for pickup_link in pickup_links:
    pickup_res = requests.get(pickup_link)
    # 標準より早いlxmlパーサーを使ってみている
    pickup_soup = BeautifulSoup(pickup_res.text, ["lxml", "xml"])
    # 開発者ツールを使うと、特定の名前のクラス名に記事の内容が入っている
    pickup_elem = pickup_soup.find("p", class_="sc-fsGQkc gnQXTK")
    news_link = pickup_elem.contents[0].attrs['href']

    news_res = requests.get(news_link)
    news_soup = BeautifulSoup(news_res.text, ["lxml", "xml"])

    print(news_soup.title.text)
    print(news_link)

    detail_text = news_soup.find(class_=re.compile("sc-ipXKqB LKGIH yjSlinkDirectlink highLightSearchTarget"))
    print(detail_text.text if hasattr(detail_text, "text") else '', end='\n\n\n')
```

- 結果

```
いじめ加害者に学校への立ち入り制限…自民作業部会が提言案（読売新聞オンライン） - Yahoo!ニュース
https://news.yahoo.co.jp/articles/58b0e5978fc036b192f6dc8baa2b063f0cd2b3a1
  いじめの被害者を守るため、加害者側の児童生徒について、学校の敷地に入らないことを命じる新たな懲戒制度の創設が検討される見通しであることがわかった。いじめ対策を検討する自民党の作業部会（座長・三谷英弘衆院議員）が提言案をまとめた。文部科学省は提言を受け、具体的な検討を始める考えだ。

【写真】公園に手向けられた花、一方で学校側はいじめ否定文書配布


EU大統領が広島訪問　ロシアの核「世界に脅威」と批判（共同通信） - Yahoo!ニュース
https://news.yahoo.co.jp/articles/a303670c8391ea8b776e438ae227827c7291ae4d
  欧州連合（EU）のミシェル大統領は13日、被爆地の広島市を訪れ、声明を発表。ウクライナに侵攻したロシアが「許し難いことに核兵器の使用に言及している」と非難。北朝鮮も「違法で挑発的なミサイル実験を繰り返している」として「世界の安全保障の脅威となっている」と批判した。
```