PythonでWebスクレイピングをしてみよう！

s1300004 伊藤優汰

自己紹介

- 名前　伊藤優汰
- 学部1年
- PC　Mac Book Air M1 2020
- 勉強中の言語　Python ,C
- テキストエディター　VSCode
- 趣味　アニメ、ゲーム、料理

ゴール

英語のvocabularyの
単語の意味をテキストファイルに
書き込もう！

作るためにやること

① PDFファイルから文字を読み込む

② 読み込んだ単語の意味をWeblioから抽出します

③ テキストファイルに書き込む

使用したライブラリ

- PyPDF2

  PDFファイルの英数字を読み込みます（日本語未対応）
- requests

  HTMLからデータを取得します
- BeautifulSoup

  requestsからの必要なデータを抽出します
- os

  PC内のファイルの存在確認に使用します

Introductory English 2 vocabulary

Week 3

1. nervous
2. refer
3. attend
4. desert
5. whip
6. rot
7. seconds
8. exercise
9. respect
10. tap
11. swallow
12. fault
13. lump
14. super
15. process
16. material
17. trunk
18. effect
19. perform
20. include
21. nowhere
22. iron
23. vehicle
24. chief
25. protest
26. occasion
27. bean
28. exhaust
29. seek
30. century
31. belong
32. lesson
33. magic
34. breathe
35. switch

これが実際の
vocabularyです
このPDFのテキストを
読み込んで
単語の意味を持ってき
ます

単語の意味はweblioの青で選択した部分から取得します

単語の意味の取得方法

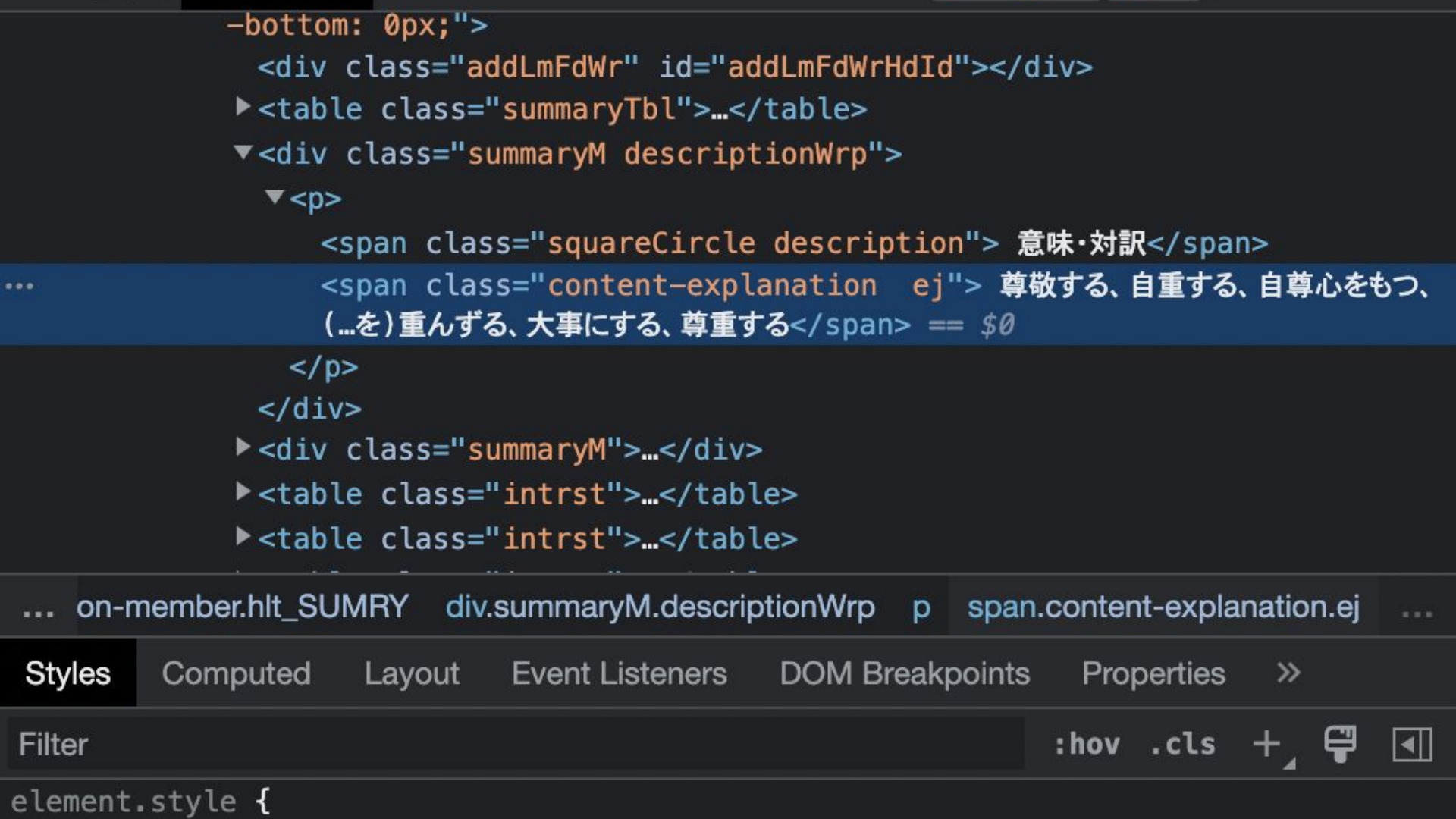
① Google Chrome で欲しい情報の部分を選択します

② 右クリックをして「検証」を選択すると、

デベロッパーツールが表示されます。

③ 選択されている部分で右クリックでCopyを選択して

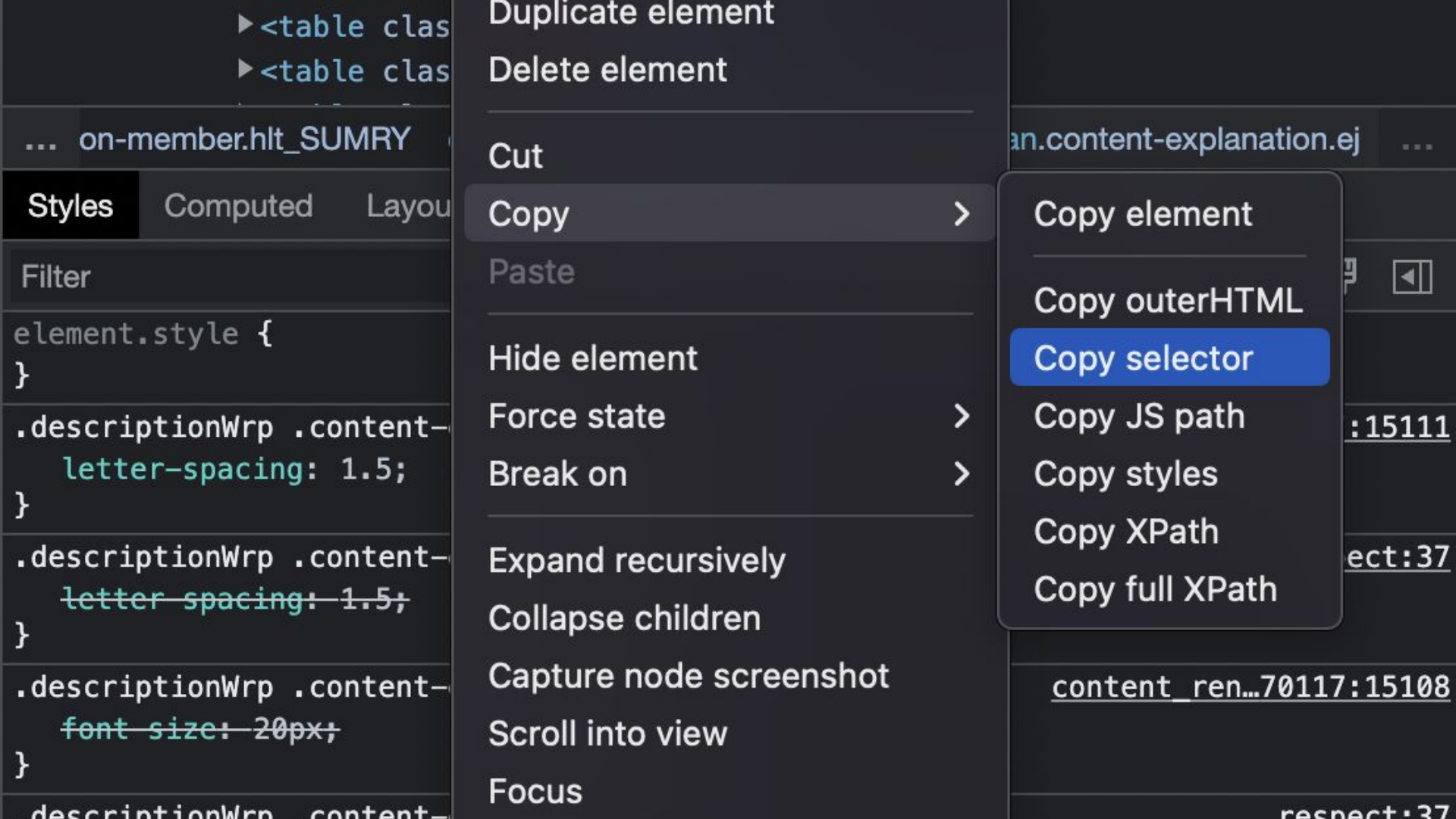
Copy selectorを選択します。この情報を使います。


-bottom: 0px;">

<table class="summaryTbl">…</table>

 意味・対訳
 尊敬する、自重する、自尊心をもつ、(…を)重んずる、大事にする、尊重する == $0

…
<table class="intrst">…</table>
<table class="intrst">…</table>
… on-member.hlt_SUMRY div.summaryM.descriptionWrp p span.content-explanation.ej …
Styles
Computed
Layout
Event Listeners
DOM Breakpoints
Properties
Filter
:hov .cls
element.style {


▶<table clas
▶<table clas
... on-member.hlt_SUMRY
an.content-explanation.ej ...
Styles
Computed
Layou
Filter
element.style {
}
.descriptionWrp .content-
letter-spacing: 1.5;
}
.descriptionWrp .content-
letter-spacing: 1.5;
}
.descriptionWrp .content-
font-size: 20px;
}
:15111
ect:37
content_ren…70117:15108
respect:37
Duplicate element
Delete element
Cut
Copy
Paste
Hide element
Force state
Break on
Expand recursively
Collapse children
Capture node screenshot
Scroll into view
Focus
Copy element
Copy outerHTML
Copy selector
Copy JS path
Copy styles
Copy XPath
Copy full XPath

単語の検索の仕方

https://ejje.weblio.jp/content/apple

これが「apple」のURLです。

URLの末尾に検索したい単語を入力すると

その単語をweblio内で検索してくれます。

今回のコードです

```
import PyPDF2
import os
import requests
from bs4 import BeautifulSoup

print("第何回ですか？",end="")
num = input()
print("実行中")

with open("IE2 vocabulary Week "+num+".pdf", "rb") as f:#ここでPDFファイルの読み込み
    reader = PyPDF2.PdfFileReader(f)
    page = reader.getPage(0)
    words=page.extractText().split()
```

```
for l in words:#一個ずつ入れる
    url1 = 'https://ejje.weblio.jp/content/'
    url = url1+str(l)
    res = requests.get(url)
    soup = BeautifulSoup(res.text, "html.parser")
    elems1 = soup.select('#summary >
div.summaryM.descriptionWrp > p >
span.content-explanation.ej')
    #単語の意味を取り出す↑
    filepath = 'vocabulary'+num+'.txt'
    exists = os.path.exists(filepath)#ファイルがあるかの確認
```

```
if str(exists) =="True":
        try:
            with open("vocabulary"+num+".txt",
mode='a') as f:
                f.write(l)#英単語を書き込む
                f.write(elems1[0].contents[0])#意味を
書き込む
                f.write("\n")
        except IndexError:
            continue
```

```
else:#存在しなければファイルを作成
        path = 'vocabulary.txt'
        f = open(path, 'w')
        f.write('')#何も入れないテキストファイルの作成
        f.close()
        try:
            elems1[0].contents[0]
            with open("vocabulary"+num+".txt", mode='a') as f:
                f.write(l)
                f.write(elems1[0].contents[0])
                f.write("\n")
        except IndexError:
            continue


print("実行完了")
```

実行結果の一部

```
8.exercise
    |   |   |   (体の)運動、練習、けいこ、実習、習作、試作、(軍隊・艦隊などの)演習、軍事演習、練習問題、課題
9.respect
    |   |   |   尊敬する、自重する、自尊心をもつ、(…を)重んずる、大事にする、尊重する
10.tap
    |   |   |   (…を)軽くたたく、軽くぽんとたたく、(…で)(…を)コツコツたたく、コツコツと立てる、コツコツたたいて送る、灰をたたいて落とす、
11.swallow
    |   |   |   ぐっと飲む、飲み込む、(…を)飲み込む、見えなくする、使い尽くす、なくす、うのみにする、軽信する、忍ぶ、抑える
12.fault
    |   |   |   (性格上の)欠点、短所、欠陥、きず、誤り、過失、失策、落ち度、(過失の)責任、罪
13.lump
    |   |   |   (不定形の)かたまり、角砂糖 1 個、こぶ、腫(は)れ物、ずんぐりした人、まぬけ、のろま、たくさん、どっさり、批判
14.super
    |   |   |   (アパートなどの)管理人、監督(など)、警視、警察本部長、(せりふなしの)端役、特製品、特大品
15.process
    |   |   |   過程、経過、成り行き、進行、(ものを造る)方法、手順、工程、処置、訴訟手続き、令状
16.material
    |   |   |   原料、材料、(服などの)生地、資料、データ、用具、道具、人材
```

参考にしたサイト

- 図解！PythonでWEB スクレイピングを極めよう！(サンプルコード付きチュートリアル)

  https://ai-inter1.com/python-webscraping/
- PythonでPDFからテキストを読み取る方法について

  https://gammasoft.jp/blog/python-parse-pdf-contents/
- pythonでファイルの存在を確認する - Qiita

  https://qiita.com/tortuepin/items/4a0669d8f275e966229e

ご清聴ありがとうございました