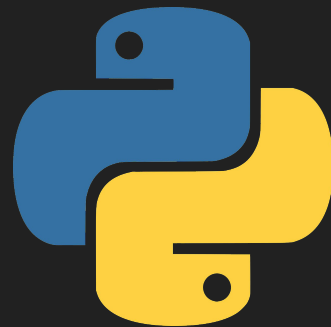


Pythonで  
Webスクレイピングをして  
みよう！

- PC Mac Book Air M1 2020
- 勉強中の言語 Python ,C
- やりたい言語 Go
- 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   | 19. perform  |
| 2. refer     | 20. include  |
| 3. attend    | 21. nowhere  |
| 4. desert    | 22. iron     |
| 5. whip      | 23. vehicle  |
| 6. rot       | 24. chief    |
| 7. seconds   | 25. protest  |
| 8. exercise  | 26. occasion |
| 9. respect   | 27. bean     |
| 10. tap      | 28. exhaust  |
| 11. swallow  | 29. seek     |
| 12. fault    | 30. century  |
| 13. lump     | 31. belong   |
| 14. super    | 32. lesson   |
| 15. process  | 33. magic    |
| 16. material | 34. breathe  |
| 17. trunk    | 35. switch   |
| 18. effect   |              |

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

ターボ検索



respect



と一致する



項目

Weblio 辞書 > 英和辞典・和英辞典 > 英和辞典 > respectの意味・解説

意味

例文 (999件)

類語

共起表現

respect とは 意味・読み方・使い方



発音を聞く プレーヤー再生 ピン留め

単語を追加

英会話で使う

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

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;">
  <div class="addLmFdWr" id="addLmFdWrHdId"></div>
```

```
▶<table class="summaryTbl">...</table>
```

```
▼<div class="summaryM descriptionWrp">
```

```
▼<p>
```

```
  <span class="squareCircle description"> 意味・対訳</span>
```

```
  <span class="content-explanation ej"> 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</span> == $0
```

```
</p>
```

```
</div>
```

```
▶<div class="summaryM">...</div>
```

```
▶<table class="intrst">...</table>
```

```
▶<table class="intrst">...</table>
```

```
... on-member.hlt_SUMRY  div.summaryM.descriptionWrp  p  span.content-explanation.ej  ...
```

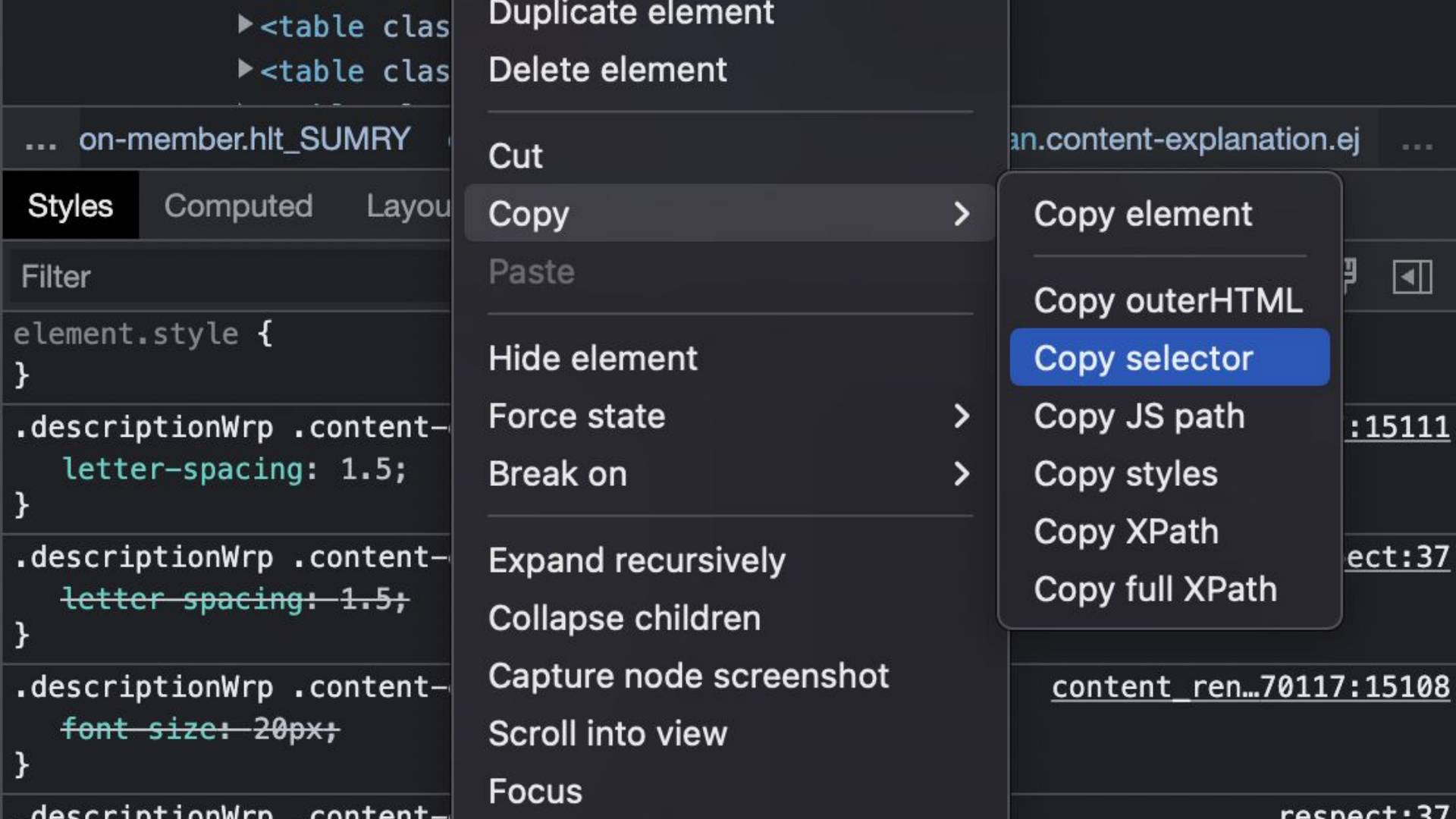
Styles   Computed   Layout   Event Listeners   DOM Breakpoints   Properties   >>

Filter

:hov .cls +



element.style {



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
    #PDFファイルの読み込み
    reader = PyPDF2.PdfFileReader(f)
    page = reader.getPage(0)
    words=page.extractText().split()
```



```
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
```



# 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、軽くぼんとたたく、(...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、特大片

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://gammасoft.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