

はじめに。

今回は ネットに繋いで情報取得するプログラム作りますので、
以下のWi-Fi設定をしてください。

Python_Lab

20191114

1. 基本構文など

他プログラム言語を少しは触ったことが有る前提としているので
言語予約語、変数、など説明しません。

な感じでお願いします。



python 基本構文



侍エンジニア塾ブログ

<https://www.sejuku.net/blog/49951>

学生のためのPython講座

<http://python4study.9isnine.com/abc>

★解らないときは周りに聞いてください。仲良くなれます。

2. WEBから情報取得について

WEBページから情報を取得する技術のことをWEBスクレイピングと言います。

そもそも、WEBページはHTMLなどの文字列を受信してブラウザで表示していますので。

受信した文字列の欲しいところだけ抜き出せば良いだけです。

注意:

WEBスクレイピングを禁止しているサービスもありますのでPG組む前に目的サイトの注意事項を要確認です。

Twitter 、 AMAZON などは明確に禁止されている。

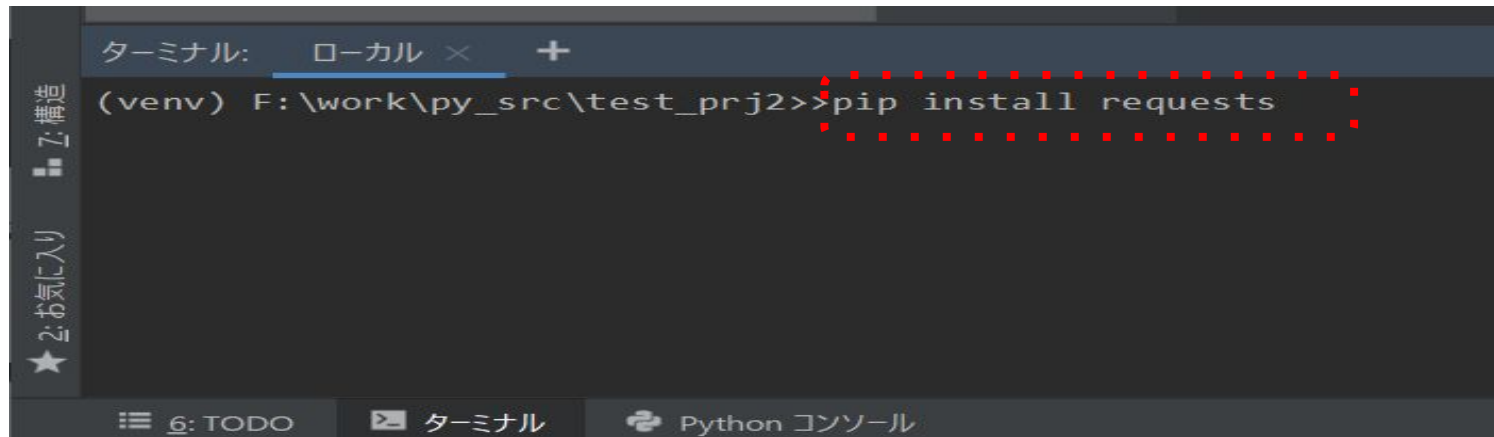
3. ライブラリの使用

WEBからの情報取得は色々な方法がありますが、今回は、BeautifulSoupを使用します。

requestsと

前回と同様にPyCharmを使用してプロジェクトを作成して進めます。

PyCharm下部ウィンドのターミナルからpipコマンドではrequestsをインストールします。



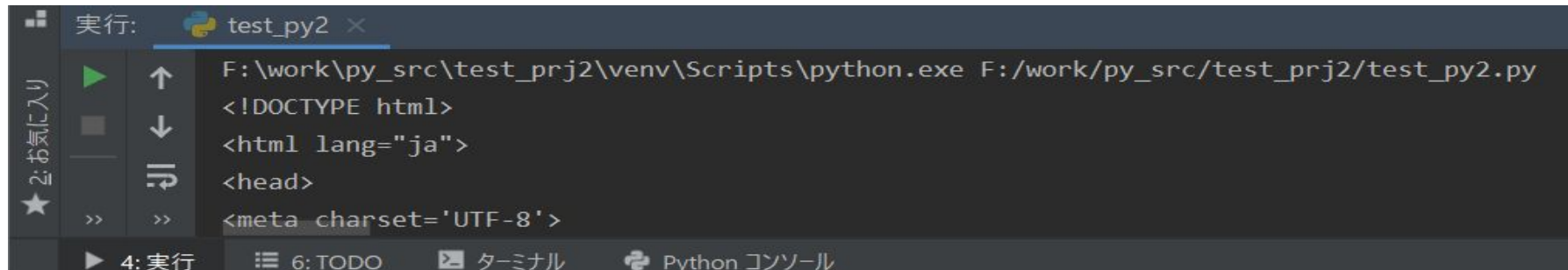
4. WEBページ情報を取得する

目的のWEBサイトのURLを指定してHTMLテキストを取得します。

以下のソースをpyファイルに打ち込み実行してください。

```
import requests  
get_data = requests.get('https://www.chancelab.jp/')  
print(get_data.text)
```

実行結果は下部ウィンドの実行タブに表示されます。



5. 特定の情報を取得するー1

BeautifulSoupライブラリを使用して特定の情報だけを取得します。

BeautifulSoupはいわゆるXMLパーサを使い易くした物です。

以下の様にpipコマンドでインストールしてください。

```
>pip install beautifulsoup4
```

以下のソースをpyファイルに打ち込み実行してください。

```
import requests
from bs4 import BeautifulSoup
get_data = requests.get('https://www.chancelab.jp/')
soup_data = BeautifulSoup(get_data.text, 'html.parser')
print(soup_data.p)  # pタグの文字列を取得する ※最初の行しか取れない
```

5. 特定の情報を取得するー2

タグ指定での取得は最初に見つけたタグしか取得できません。

セレクトタを使用して取得してみましょう。

```
import requests
from bs4 import BeautifulSoup
get_data = requests.get('https://www.chancelab.jp/')
soup_data = BeautifulSoup(get_data.text, 'html.parser')
p_teams = soup_data.select('p')    # pタグの全行を取得する
for p_man in p_teams:
    print(p_man)
```

便利なBS4ドキュメント <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

今回使用した資料について

今回使用した資料等は、以下のgithubに置いていきますので
ダウンロードして使用可能です。

https://github.com/okamotomasatosi/py_doc

