Laporan Praktikum Web Scraping

Denis Muhammad Jethro/ 162112133028

Github : https://github.com/Juethro/pythonWeb-scraper

Import Library

```
import requests as rq
from bs4 import BeautifulSoup as bs
import pandas as pd
import re
import scrapy as sc
```

Scraping Beautifulsoup

Step:

1. Ambil html dengan requests
2. Masukkan dalam file .html
3. Buka file dan masukkan dalam Beautifulsoup
4. Find_all tag yang dicari, dalam hal ini *h3*
5. Bersihkan hasilnya dan buat dataframe dari kumpulan data tersebut

```
url = rq.get("https://unair.ac.id/news")
```

Save in html format and Open it

```
with open("index.html", 'w', encoding='utf_8') as nulis:
    nulis.write(url.text)
```

```
with open("index.html", 'r', encoding='utf_8') as htmll:
    soup = bs(htmll, 'html.parser')

listt_title = soup.find_all('h3' , class_='elementor-post__title')
```

Cleaning Data *listt_title*

```
data_judul = []
filter = []

for i in listt_title:
    splitted = re.split(r'[\n\t\f\v\r]+', i.find('a').get_text())

    # Hilangkan enter depan
    for x in splitted[1:]:
        filter.append(x)

# Hilangkan enter belakang
for su in filter[0::2]:
    data_judul.append(su)
```

Make DataFrame and Export

```
df = pd.DataFrame(data_judul)
df.columns = ['Judul_Headline']
df.to_csv('../scraping-headlines.csv')

df.head()
```

	Judul_Headline
0	Kontribusi Terhadap Keberlanjutan Lingkungan, ...
1	Jalani Hidup Sehat, Ini Yang Perlu Disiapkan
2	5 Mahasiswa UNAIR Ikuti Student Exchange dan K...
3	Waspadai Resistensi Antibiotik Saat Sakit
4	Kontribusi Terhadap Keberlanjutan Lingkungan, ...

Terlihat file csv sudah terekspor dan isinya sesuai dengan apa yang ada di website *unair news* tersebut

Crawling Beautifulsoup

```
urls = rq.get('https://unair.ac.id/news')
```

Melakukan http requests ke server dan disimpan dalam variabel `urls`

Save html File and Open it

```
with open('index2.html', 'w', encoding= 'utf_8') as f:
    f.write(urls.text)

with open('index2.html', 'r') as duar:
    soup = bs(urls.text, 'html.parser')
```

Untuk menghemat tempat di ipynb, saya ekspor file html tersebut dengan nama `index2.html`. Jika ingin melihat isinya bisa dilihat di sana.

Make crawler route

```
#Dapatkan semua link artikel
all_links = soup.find_all('a', class_='elementor-post__read-more')
news_links = []

for i in all_links:
    ab = i['href']
    news_links.append(ab)
```

Dari file html tersebut, saya cari semua tag a dengan class `elementor-post__read-more` kemudian melakukan iterasi untuk memfilter `href` dan dimasukkan dalam variabel `news_links`

Crawl from news_links

```
list_judul = []
list_date = []
#crawler unit (masih bisa dikembangkan)
for i in news_links:
    buang = rq.get(i.lstrip("\'"))

    bakar = bs(buang.text, 'html.parser')
    judul = bakar.find_all('h3', class_='elementor-heading-title')
    date = bakar.find_all('span', class_='elementor-post-info__item--type-date')

    list_judul.append(judul[0].string)
```

```
    list_date.append(date[0].string)
```

Melakukan crawl dengan data link dari `news_links` kemudian di tiap link mengambil tag `h3` dengan class `elementor-heading-title` dan tag `span` dengan class `elementor-post-info__item--type-date` yakni judul artikel dan tanggal uploadnya. Kedua data tersebut dimasukkan ke dalam dua list yakni `list-judul` dan `list-date`

Cleaning Data *Date*

```
raw_date = []
filtered_date = []

for so in list_date:
    splitted = re.split(r'[\n\t\f\v\r]+', so)

    # Hilangkan enter depan
    for x in splitted[1:]:
        raw_date.append(x)

# Hilangkan enter belakang
for su in raw_date[0::2]:
    filtered_date.append(su)
```

Kode diatas digunakan untuk membersihkan data *date* yang tidak bersih dikarenakan banyak tulisan \n \t dsb. Gangguan tersebut berada pada depan judul dan diakhir masing-masing judul, sehingga mulai melakukan penghapusan *noise* tersebut

Make and Export DataFrame

```
df2 = pd.DataFrame({'Judul' : list_judul, 'Date' : filtered_date})
df2.to_csv('../crawling-headlines.csv')

df2.head()
```

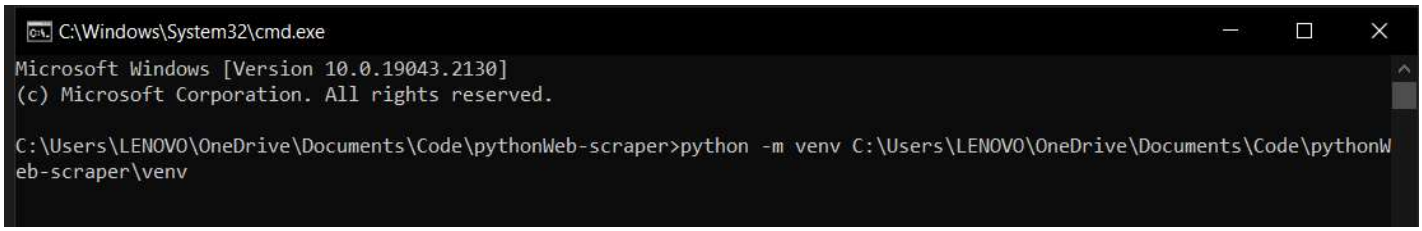
	Judul	Date
0	Kontribusi Terhadap Keberlanjutan Lingkungan, ...	Oktober 25, 2022

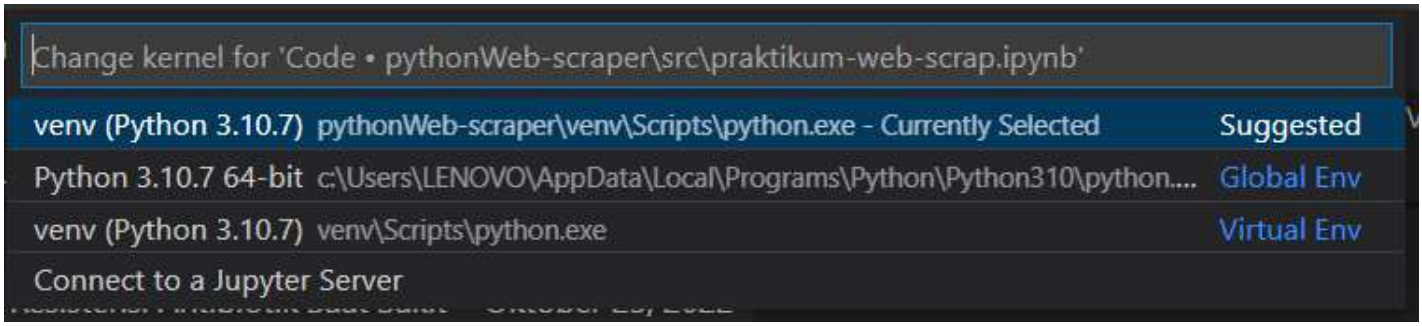
Melakukan ekspor dataframe dalam format csv dan terlihat, dataset sudah sesuai dengan yang diinginkan sehingga bisa dilakukan analisis lebih lanjut

3 Waspadai Resistensi Antibiotik Saat Sakit Oktober 25, 2022

Scrapy Crawler

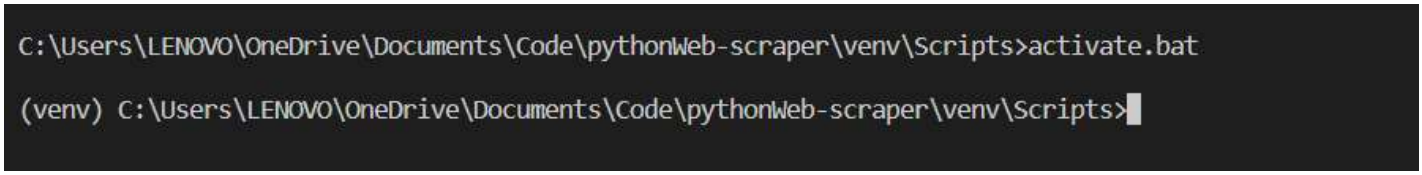
Setup VirtualENV





Membuat venv atau virtual environment, dikarenakan rumor yang beredar tentang scrapy suka merusak evironment python. Untuk mengatasinya saya membuat venv yang nanti akan menjalankan crawling scrapy selanjutnya.

Setup Scrapy Environment



```
(venv) C:\Users\LENOVO\OneDrive\Documents\Code\pythonWeb-scraper>scrapy startproject ps_crawl
```

Step:

1. Aktifkan venv `(gambar-1)`
2. Start project `(gambar-2)`

Membuat projek scrapy, ini dilakukan agar scrapy membuat folder berisi komponen spider. Komponen ini berfungsi agar crawling dapat dilakukan

Spider code

```
#Spider code
import scrapy

class Ayomain_Spider(scrapy.Spider):
    name = 'ayomain' #nama spider
    start_urls = ["https://store.playstation.com/en-id/category/29696e1b-a942-4832-935d-ebd11


    def parse(self, response):
        url = response.url

        for i in range(1,47): #iterasi sampe page ke-46
            yield scrapy.Request(url=url+str(i), callback=self.parse_details) #url awal ditam

    def parse_details(self, response):
        for text in response.css(".psw-product-tile__details"):
            yield{
                "title":text.css(".psw-t-body::text").get(), #filter yang memiliki class .psw
                "price":text.css(".psw-m-r-3::text").get()}  #filter yang memiliki class .psw
        pass
```

Membuat sebuah file python baru dengan nama file `play_spider.py` di dalam direktori spider. Di dalam direktori tersebut berisi spider yang akan melakukan crawl.

Start Spider Crawl

```
(venv) C:\Users\LENOVO\OneDrive\Documents\Code\pythonWeb-scraper\src\ps_crawl>scrapy crawl ayomain -t json -o titleNprice.json
```

Memulai crawling dengan perintah diatas `scrapy crawl ayomain -t json -o titleNprice.json`.

`scrapy` : memanggil modul scrapy

`crawl` : command untuk memulai crawl

`ayomain` : nama spider yang sudah dibuat tadi

`json` : mengeluarkan output file json berisi data yang diambil berdasarkan codingan pada spider. Dalam kasus ini ambil title game dan pricenya