M06_S2_Teknikal_WebScraping

In [142…

```
%%html
<style>
.col1{
    font-weight:bold;
    width:20%;
    display: inline-block;
}
.img{
    margin:20px;
    width:64px;
    display: inline-block;
}
.p{
    line-height: 2;
    font-size:18px
}
.hem{
    display:inline;
}
.bind{
    display:inline-block;
}
.bind1{
    font-size:18px;
}
</style>
```

Nama : Muhammad Alfan Irsyadi Hutagalung

Kelas : Persevere

ID : 29

Asal Universitas : Universitas Sumatera Utara

Jurusan/Fakultas : Matematika/FMIPA

Let's Connect :

You can open notebook in this repo with binder launch binder

Latihan

1. Lakukan langkah persiapan untuk scrap web https://books.toscrape.com/
2. Scrap https://books.toscrape.com/ untuk mencari info harga buku berdasarkan input user, dengan judul buku yang anda suka!
3. Tuliskan judul buku untuk rating >= 3 star

In [108…

```
## Gunakan Jupyter Notebook untuk menampilkan hasil yang memuaskan

!jupyter nbextension enable --py widgetsnbextension
import ipywidgets as widgets

from ipywidgets import interact, interact_manual
```

```
Enabling notebook extension jupyter-js-widgets/extension...
      - Validating: ok
```

Jawaban

Nomor 1 : Lakukan langkah persiapan untuk scrap web https://books.toscrape.com/

Langkah 1: siapkan library yang dibutuhkan

In [109…

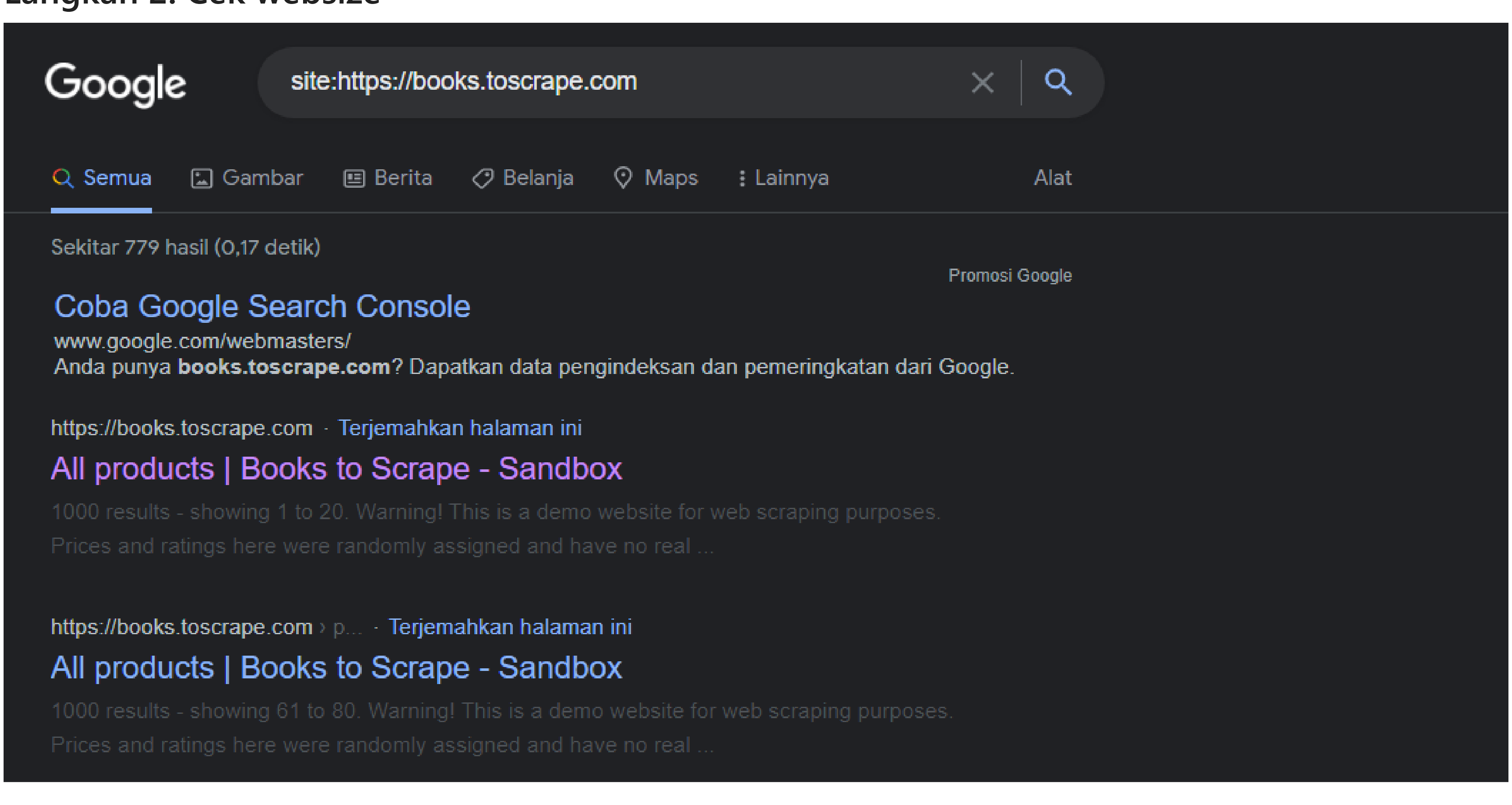
```
# download dan instal modul yang dibutuhkan

!pip install builtwith
!pip install python-whois
!pip install pandas
!pip install numpy
!pip install beautifulsoup4
```

```
Requirement already satisfied: builtwith in d:\users\lenovo\anaconda3\lib\site-packages (1.3.4)
Requirement already satisfied: six in d:\users\lenovo\anaconda3\lib\site-packages (from builtwith) (1.15.0)
Requirement already satisfied: python-whois in d:\users\lenovo\anaconda3\lib\site-packages (0.7.3)
Requirement already satisfied: future in d:\users\lenovo\anaconda3\lib\site-packages (from python-whois) (0.18.
2)
Requirement already satisfied: pandas in d:\users\lenovo\anaconda3\lib\site-packages (1.2.4)
Requirement already satisfied: pytz>=2017.3 in d:\users\lenovo\anaconda3\lib\site-packages (from pandas) (2021.
1)
Requirement already satisfied: python-dateutil>=2.7.3 in d:\users\lenovo\anaconda3\lib\site-packages (from pand
as) (2.8.1)
Requirement already satisfied: numpy>=1.16.5 in c:\users\lenovo\appdata\roaming\python\python38\site-packages
(from pandas) (1.22.2)
Requirement already satisfied: six>=1.5 in d:\users\lenovo\anaconda3\lib\site-packages (from python-dateutil>=
2.7.3->pandas) (1.15.0)
Requirement already satisfied: numpy in c:\users\lenovo\appdata\roaming\python\python38\site-packages (1.22.2)
Requirement already satisfied: beautifulsoup4 in d:\users\lenovo\anaconda3\lib\site-packages (4.9.3)
Requirement already satisfied: soupsieve>1.2 in d:\users\lenovo\anaconda3\lib\site-packages (from beautifulsoup
4) (2.2.1)
```

In [110…

```
import re
import urllib3
import urllib.request
import pandas as pd
import numpy as np
from bs4 import BeautifulSoup
import builtwith
import whois
from IPython.display import clear_output
from time import sleep
```

Langkah 2: Cek websize



Langkah 3: Cek teknologi website dibuat menggunakan library builtwith

In [111…

```
builtwith.parse('https://books.toscrape.com/')
```

Out[111…

```
{'web-servers': ['Nginx'],
 'web-frameworks': ['Twitter Bootstrap'],
 'javascript-frameworks': ['jQuery']}
```

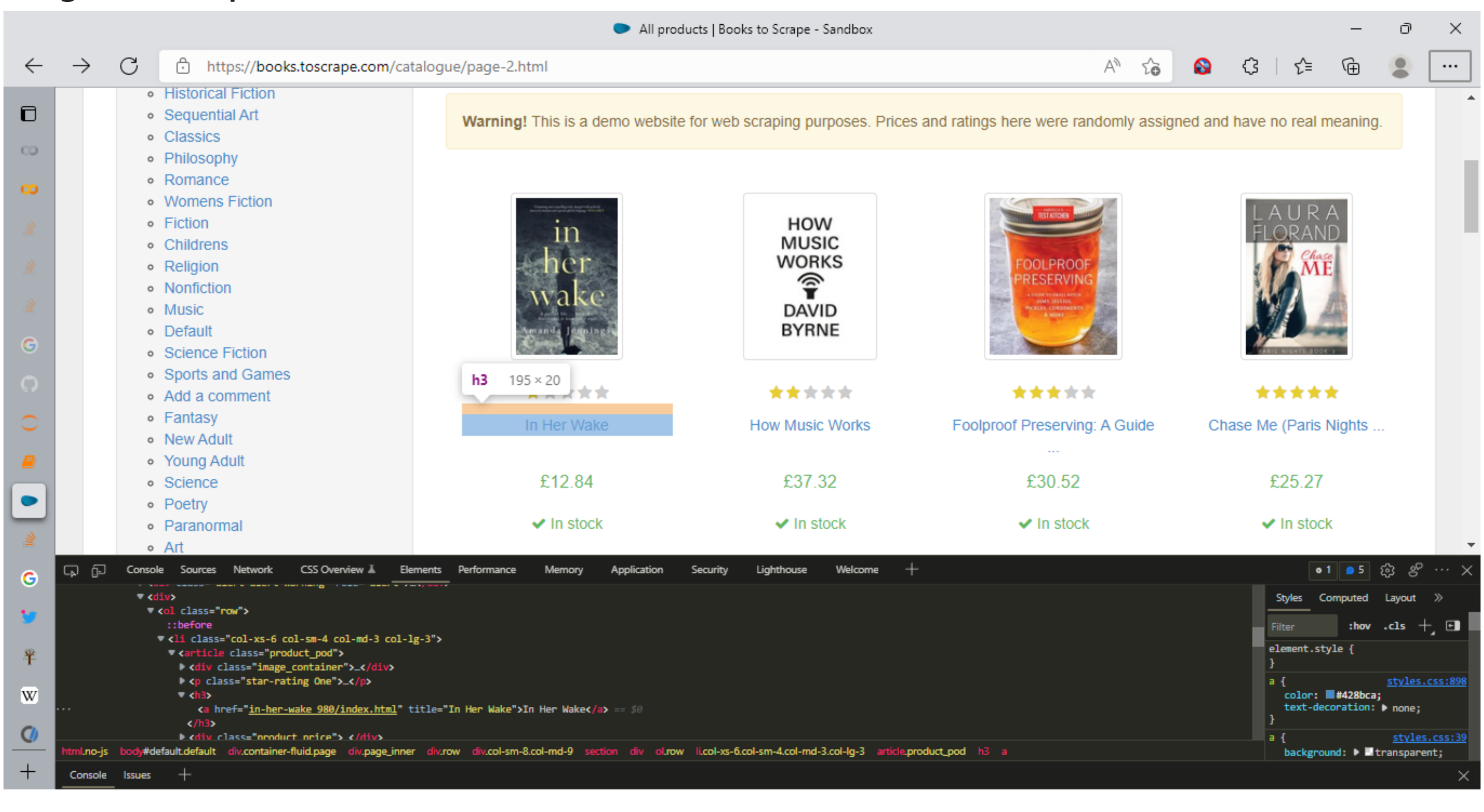
Langkah 4: Cek pemilik website menggunakan protokol whois dengan library python-whois

In [112…

```
print(whois.whois('https://books.toscrape.com'))
```


```
{
  "domain_name": [
    "TOSCRAPE.COM",
    "toscrape.com"
  ],
  "registrar": "Amazon Registrar, Inc.",
  "whois_server": "whois.registrar.amazon.com",
  "referral_url": null,
  "updated_date": [
    "2021-05-25 22:54:15",
    "2021-05-25 22:54:16.257000"
  ],
  "creation_date": "2016-06-28 20:26:52",
  "expiration_date": "2022-06-28 20:26:52",
  "name_servers": [
    "NS-1192.AWSDNS-21.ORG",
    "NS-1782.AWSDNS-30.CO.UK",
    "NS-437.AWSDNS-54.COM",
    "NS-740.AWSDNS-28.NET",
    "ns-1192.awsdns-21.org",
    "ns-1782.awsdns-30.co.uk",
    "ns-437.awsdns-54.com",
    "ns-740.awsdns-28.net"
  ],
  "status": [
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "renewPeriod https://icann.org/epp#renewPeriod"
  ],
  "emails": [
    "abuse@amazonaws.com",
    "owner-1961310@toscrape.com.whoisprivacyservice.org",
    "admin-1961310@toscrape.com.whoisprivacyservice.org",
    "tech-1961310@toscrape.com.whoisprivacyservice.org"
  ],
  "dnssec": "unsigned",
  "name": "On behalf of toscrape.com owner",
  "org": "Whois Privacy Service",
  "address": "P.O. Box 81226",
  "city": "Seattle",
  "state": "WA",
  "zipcode": "98108-1226",
  "country": "US"
}
```


Langkah 5: Inspect element



Nomor 2: Scrap https://books.toscrape.com/ untuk mencari info harga buku berdasarkan input user, dengan judul buku yang anda suka!

In [113…

```
%%time
urllib3.disable_warnings() #disable annoying warnings


judul = []

harga = []

rating = []
# <p class="star-rating Two">
def to_number(kata):

    if kata == 'One' : return 1

    elif kata =='Two': return 2

    elif kata =='Three': return 3

    elif kata =='Four': return 4

    else: return 5


print('Sedang memproses....')

for page in range(1,51):

    response = urllib.request.urlopen(f'https://books.toscrape.com/catalogue/page-%d.html'%page)

    html = response.read()

    soup = BeautifulSoup(html, 'html.parser')

    for i in soup.find_all("article", class_="product_pod"):

        for j in i.find("h3"):

            judul.append(j.get('title'))


    text = html.decode()

    rate = re.findall('<p class="star-rating (.*?)">',text)

    price = re.findall('<p class="price_color">(.*?)</p>',text)

    rating.append(rate)

    harga.append(price)

    clear_output(wait=True)

    if(page != 50):

        string = 'Sedang memproses'+'.'*(page%4)

    else:
        string = 'Selesai :D'
    print(string)

    print('Page %2d/50: '%page, '🔴'*int(np.floor(page/5))+\

        '⭕'*int(np.floor(10- page/5))+' '+str(page*2)+'%',end='')

    sleep(0.01)


print(' ')

harga = np.array(harga).flatten()

rating = np.array(rating).flatten()


df = pd.DataFrame({'Judul':judul, 'Harga':harga, 'Rating':rating})

df.Rating = df.Rating.apply(lambda x: int(to_number(x)))

df.Judul = df.Judul.apply(lambda x: x.lower())
```

```
Selesai :D
Page 50/50:  ⚫⚫⚫⚫⚫⚫⚫⚫⚫⚫ 100%
Wall time: 1min 21s
```

In [114…

```
@interact
def show_title_contain(Judul='frank'):

    return df[df.Judul.str.contains(Judul.lower())]
```

Nomor 3: Tuliskan judul buku untuk rating >= 3 star

In [115…

```
@interact
def show_rating_more_than(rate=(1,5,1)):

    return df[df.Rating >= rate]
```

Thank You!!