# LAPORAN PRAKTIKUM
# PYTHON

## WEB SCRAPPING

**Disusun oleh :**

Yunita Kartika Putri (V3923019)

**Dosen :** Yusuf Fadila Rachman. S.Kom., M.Kom

# PS D-III TEKNIK INFORMATIKA
# SEKOLAH VOKASI UNIVERSITAS SEBELAS MARET 2024

```python
import requests
from bs4 import BeautifulSoup
import csv

def fetch_content(url):
    response = requests.get(url)
    if response.status_code == 200:
        return response.content
    else:
        print(f'Failed to retrieve {url}. Status code: {response.status_code}')
        return None

def save_to_csv(data, filename, headers=None):
    with open(filename, 'w', newline='', encoding='utf-8') as csvfile:
        writer = csv.writer(csvfile)
        if headers:
            writer.writerow(headers)
        for row in data:
            writer.writerow(row)

```

Fungsi `fetch_content` bertujuan untuk mengambil konten dari URL yang diberikan menggunakan pustaka `requests`. Jika respons dari URL memiliki status kode 200, yang menunjukkan bahwa permintaan berhasil, fungsi ini mengembalikan konten dari respons tersebut. Jika respons memiliki status kode yang berbeda, fungsi ini mencetak pesan kesalahan yang mencantumkan URL dan status kode respons.

Fungsi `save_to_csv` digunakan untuk menyimpan data ke dalam file CSV. Fungsi ini menerima tiga parameter: data yang akan disimpan, nama file untuk disimpan, dan opsional, headers untuk kolom-kolom dalam file CSV. Fungsi ini membuka file CSV dengan mode 'w' untuk menulis, kemudian menulis header jika diberikan, dan selanjutnya menulis setiap baris data ke dalam file CSV tersebut menggunakan objek penulis dari pustaka `csv`.

## 1. Gambar

### a. Sourcode

```python

def scrape_proxyway_images():
    url = 'https://proxyway.com/news'
    content = fetch_content(url)
    if content:
        soup = BeautifulSoup(content, 'html.parser')
        images_list = []
        images = soup.select('img')
        for image in images:
            src = image.get('src')
            alt = image.get('alt')
            images_list.append([src, alt])
        save_to_csv(images_list, 'data_gambar_yunita.csv', headers=['src', 'alt'])
        for image in images_list:
            print(image)
```

Fungsi `scrape_proxyway_images` adalah sebuah fungsi yang dirancang untuk mengumpulkan gambar dari halaman berita situs "Proxyway". Fungsi ini dimulai dengan mendefinisikan URL tujuan sebagai 'https://proxyway.com/news' dan kemudian menggunakan fungsi `fetch_content` untuk mengambil konten dari halaman tersebut. Jika pengambilan konten berhasil, konten tersebut diuraikan menggunakan `BeautifulSoup`, sebuah perpustakaan Python yang digunakan untuk mengurai dokumen HTML dan XML. Dengan `BeautifulSoup`, fungsi ini mencari semua elemen gambar (`<img>`) dalam halaman tersebut.

Setelah menemukan semua elemen gambar, fungsi ini membuat sebuah daftar yang menyimpan atribut `src` (yang menunjukkan sumber gambar) dan `alt` (yang berisi teks alternatif gambar) dari setiap gambar. Daftar ini kemudian disimpan ke dalam sebuah file CSV bernama 'data_gambar_yunita.csv' dengan kolom `src` dan `alt` untuk setiap gambar. Sebagai langkah terakhir, fungsi ini mencetak setiap entri dalam daftar gambar ke layar, yang memudahkan pengguna untuk melihat gambar-gambar yang telah diambil beserta deskripsi teks alternatifnya.

## b. Output

```
['https://proxyway.com/wp-content/uploads/2023/04/proxyway.svg?ver=1681290142', 'Proxyway']
['https://proxyway.com/wp-content/uploads/2023/04/Newsfeed-image-min.png', 'Adam sitting in a chair reading a newspaper']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2020/07/oxylabs-logo-png.png?ver=1704718753', 'Oxylabs logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964', 'Bright Data logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/07/rayobyte-logo.png?ver=1704718347', 'rayobyte logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964', 'Bright Data logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2020/08/infatica-logo.png?ver=1704718655', 'infatica logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/07/rayobyte-logo.png?ver=1704718347', 'rayobyte logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964', 'Bright Data logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2020/07/oxylabs-logo-png.png?ver=1704718753', 'Oxylabs logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964', 'Bright Data logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/03/zyte-logo.png?ver=1704718975', 'Zyte logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2019/09/Smartproxy-Logo.png?ver=1704719397', 'smartproxy-logo']
['https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g', 'Adam Dubois']
['https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964', 'Bright Data logo']
['https://proxyway.com/wp-content/uploads/2023/04/proxyway.svg?ver=1681290142', 'Proxyway']
```

## c. CSV

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | src,alt | | | | | | | | | | | | | |
| 2 | https://proxyway.com/wp-content/uploads/2023/04/proxyway.svg?ver=1681290142,Proxyway | | | | | | | | | | | | | |
| 3 | https://proxyway.com/wp-content/uploads/2023/04/Newsfeed-image-min.png,Adam sitting in a chair reading a newspaper | | | | | | | | | | | | | |
| 4 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 5 | https://proxyway.com/wp-content/uploads/2020/07/oxylabs-logo-png.png?ver=1704718753,Oxylabs logo | | | | | | | | | | | | | |
| 6 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 7 | https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964,Bright Data logo | | | | | | | | | | | | | |
| 8 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 9 | https://proxyway.com/wp-content/uploads/2022/07/rayobyte-logo.png?ver=1704718347,rayobyte logo | | | | | | | | | | | | | |
| 10 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 11 | https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964,Bright Data logo | | | | | | | | | | | | | |
| 12 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 13 | https://proxyway.com/wp-content/uploads/2020/08/infatica-logo.png?ver=1704718655,infatica logo | | | | | | | | | | | | | |
| 14 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 15 | https://proxyway.com/wp-content/uploads/2022/07/rayobyte-logo.png?ver=1704718347,rayobyte logo | | | | | | | | | | | | | |
| 16 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 17 | https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964,Bright Data logo | | | | | | | | | | | | | |
| 18 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 19 | https://proxyway.com/wp-content/uploads/2020/07/oxylabs-logo-png.png?ver=1704718753,Oxylabs logo | | | | | | | | | | | | | |
| 20 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 21 | https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964,Bright Data logo | | | | | | | | | | | | | |
| 22 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 23 | https://proxyway.com/wp-content/uploads/2022/03/zyte-logo.png?ver=1704718975,Zyte logo | | | | | | | | | | | | | |
| 24 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 25 | https://proxyway.com/wp-content/uploads/2019/09/Smartproxy-Logo.png?ver=1704719397,smartproxy-logo | | | | | | | | | | | | | |
| 26 | https://secure.gravatar.com/avatar/40de7ce27e119cfc9a04eef5e77cc6c2?s=96&r=g,Adam Dubois | | | | | | | | | | | | | |
| 27 | https://proxyway.com/wp-content/uploads/2022/05/bright-data-logo.png?ver=1704718964,Bright Data logo | | | | | | | | | | | | | |
| 28 | https://proxyway.com/wp-content/uploads/2023/04/proxyway.svg?ver=1681290142,Proxyway | | | | | | | | | | | | | |
| 29 | | | | | | | | | | | | | | |

## 2. Subjudul

### a. Sourcode

```
40  def scrape_proxyway_subtitles():
41      url = 'https://proxyway.com/news'
42      content = fetch_content(url)
43      if content:
44          soup = BeautifulSoup(content, 'html.parser')
45          paragraphs = soup.find_all('h2')
46          subtitles = [[paragraph.text] for paragraph in paragraphs]
47          save_to_csv(subtitles, 'subjudul_yunita.csv')
48          for subtitle in subtitles:
49              print(subtitle[0])
```

Fungsi `scrape_proxyway_subtitles` bertujuan untuk mengekstrak subjudul dari halaman berita situs "Proxyway". Mirip dengan fungsi sebelumnya, fungsi ini dimulai dengan menetapkan URL tujuan sebagai 'https://proxyway.com/news' dan mengambil kontennya menggunakan fungsi `fetch_content`. Setelah berhasil mendapatkan konten, konten tersebut diurai menggunakan `BeautifulSoup` untuk mempermudah pencarian elemen HTML.

Dalam hal ini, fungsi mencari semua elemen `<h2>`, yang sering digunakan untuk menandai subjudul dalam dokumen HTML. Subjudul-subjudul ini kemudian disimpan dalam sebuah daftar menggunakan pemrosesan daftar, dan daftar tersebut disimpan ke dalam file CSV bernama 'subjudul_yunita.csv'. Selanjutnya, setiap subjudul dalam daftar diprint ke layar untuk ditampilkan kepada pengguna. Dengan demikian, fungsi ini memungkinkan pengguna untuk melihat daftar subjudul dari halaman berita "Proxyway".

### b. Output

```
Oxylabs Lowers Mobile Proxy Prices by Up to 60%
Bright Data Equalizes Residential and Mobile Proxy Rates
Rayobyte Publishes a Whitepaper on Preventing Proxy Abuse
ScrapeCon 2024: A Recap
Infatica Makes Its Residential Proxies Up to 43% Cheaper
Rayobyte Decreases Residential Proxy Prices, Moves to Subscription
Bright Data’s ScrapeCon to Take Place on April 2
Oxylabs Cuts Residential Proxy Rates by Up to 20%
Bright Data Reduces Residential Proxy Prices by 20%
Zyte Adds AI Scraping Functionality to Its API
Smartproxy Slashes Residential Proxy Prices by Up to 25%
Meta Drops the Case Against Bright Data
The provider also reorganized its pricing scheme.
```

### c. CSV

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Oxylabs Lowers Mobile Proxy Prices by Up to 60% | | | | | | | | | |
| 2 | Bright Data Equalizes Residential and Mobile Proxy Rates | | | | | | | | | |
| 3 | Rayobyte Publishes a Whitepaper on Preventing Proxy Abuse | | | | | | | | | |
| 4 | ScrapeCon 2024: A Recap | | | | | | | | | |
| 5 | Infatica Makes Its Residential Proxies Up to 43% Cheaper | | | | | | | | | |
| 6 | Rayobyte Decreases Residential Proxy Prices, Moves to Subscription | | | | | | | | | |
| 7 | Bright Dataâ€™s ScrapeCon to Take Place on April 2 | | | | | | | | | |
| 8 | Oxylabs Cuts Residential Proxy Rates by Up to 20% | | | | | | | | | |
| 9 | Bright Data Reduces Residential Proxy Prices by 20% | | | | | | | | | |
| 10 | Zyte Adds AI Scraping Functionality to Its API | | | | | | | | | |
| 11 | Smartproxy Slashes Residential Proxy Prices by Up to 25% | | | | | | | | | |
| 12 | Meta Drops the Case Against Bright Data | | | | | | | | | |
| 13 | | | | | | | | | | |

3. **Deskripsi**

a. **Sourcode**

```python
def scrape_proxyway_descriptions():
    url = 'https://proxyway.com/news'
    content = fetch_content(url)
    if content:
        soup = BeautifulSoup(content, 'html.parser')
        div_elements = soup.find_all('div', attrs={'data-widget_type': 'theme-post-excerpt.default'})
        descriptions = []
        for div_element in div_elements:
            inner_div = div_element.find('div', class_='elementor-widget-container')
            if inner_div:
                text_content = inner_div.get_text(strip=True)
                descriptions.append([text_content])
                print(text_content)
        save_to_csv(descriptions, 'keterangan_yunita.csv')

```

Fungsi `scrape_proxyway_descriptions` bertujuan untuk mengekstrak deskripsi atau keterangan dari halaman berita situs "Proxyway". Prosesnya dimulai dengan menetapkan URL target sebagai 'https://proxyway.com/news' dan mengambil kontennya menggunakan fungsi `fetch_content`. Setelah mendapatkan konten dengan sukses, konten tersebut diurai menggunakan `BeautifulSoup` untuk memudahkan pencarian elemen HTML.

Fungsi ini kemudian mencari semua elemen `<div>` yang memiliki atribut `data-widget_type` yang sama dengan 'theme-post-excerpt.default', yang umumnya digunakan untuk menampung deskripsi atau keterangan sebuah posting. Setiap deskripsi yang ditemukan diambil dan dimasukkan ke dalam sebuah daftar. Kemudian, deskripsi-deskripsi tersebut diprint ke layar untuk ditampilkan kepada pengguna. Selanjutnya, daftar deskripsi disimpan ke dalam sebuah file CSV bernama 'keterangan_yunita.csv' menggunakan fungsi `save_to_csv`. Dengan demikian, pengguna dapat dengan mudah melihat deskripsi atau keterangan yang terkandung dalam halaman berita "Proxyway".

b. **Output**

```
The provider also reorganized its pricing scheme.
In essence, the provider’s mobile proxies just got 65% cheaper.
The document describes Rayobyte’s three-pronged approach against bad actors.
Our impressions from Bright Data’s first virtual conference on web scraping.
The provider’s plans now give significantly more traffic for the same price.
The new rates are up to 50% cheaper in lower ranges.
The rescheduled web scraping event will proceed in a week.
The reduction affects three entry plans.
The second round of price cuts continues.
The tool can now crawl, unblock, and parse websites using AI and an optional no-code interface.
It’s the second price cut in less than a year.
The dispute results in a complete win for the data collection infrastructure provider.
```

**c. CSV**

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | The provider also reorganized its pricing scheme. | | | | | | | | |
| 2 | In essence, the providerâ€™s mobile proxies just got 65% cheaper. | | | | | | | | |
| 3 | The document describes Rayobyteâ€™s three-pronged approach against bad actors. | | | | | | | | |
| 4 | Our impressions from Bright Dataâ€™s first virtual conference on web scraping. | | | | | | | | |
| 5 | The providerâ€™s plans now give significantly more traffic for the same price. | | | | | | | | |
| 6 | The new rates are up to 50% cheaper in lower ranges. | | | | | | | | |
| 7 | The rescheduled web scraping event will proceed in a week. | | | | | | | | |
| 8 | The reduction affects three entry plans. | | | | | | | | |
| 9 | The second round of price cuts continues. | | | | | | | | |
| 10 | The tool can now crawl, unblock, and parse websites using AI and an optional no-code interface. | | | | | | | | |
| 11 | Itâ€™s the second price cut in less than a year. | | | | | | | | |
| 12 | The dispute results in a complete win for the data collection infrastructure provider. | | | | | | | | |
| 13 | | | | | | | | | |

```
66 if __name__ == "__main__":
67
68
69
70     # Scrape proxyway.com/news for images and save to CSV
71     scrape_proxyway_images()
72
73     # Scrape proxyway.com/news for subtitles (h2 tags) and save to CSV
74     scrape_proxyway_subtitles()
75
76     # Scrape proxyway.com/news for specific div descriptions and save to CSV
77     scrape_proxyway_descriptions()
78
```

Pada blok kode yang disediakan, `if __name__ == "__main__":` digunakan untuk mengeksekusi tiga fungsi secara berurutan ketika file ini dijalankan sebagai program utama. Pertama, fungsi `scrape_proxyway_images()` dipanggil untuk mengambil gambar dari halaman berita situs "Proxyway" dan menyimpannya ke dalam file CSV. Kemudian, fungsi `scrape_proxyway_subtitles()` dipanggil untuk mengekstrak subjudul (yang ditandai dengan tag `<h2>`) dari halaman berita yang sama dan menyimpannya ke dalam file CSV. Terakhir, fungsi `scrape_proxyway_descriptions()` dipanggil untuk mengambil deskripsi atau keterangan khusus dari halaman berita tersebut dan juga menyimpannya dalam format CSV. Dengan cara ini, ketiga proses ini dapat dijalankan secara otomatis saat file dieksekusi, memudahkan penggunaan dan pemeliharaan.