Nama: Rhegysa Alvyanthi Juniartha
NIM: 1123102098
Mata Kuliah: Data Scraping
Tugas Individu 2

DATA SCRAPING

TUGAS INDIVIDU 2

1.1 Capaian Praktikum

- Mahasiswa mampu memahami etika dan praktik terbaik dalam pengambilan data gambar (scraping)
- Mahasiswa mampu menggunakan alat dan metode pengambilan gambar yang tepat dalam scraping
- Mahasiswa mampu menyimpan hasil scraping gambar dengan cara yang terstruktur dan efisien

1.2 Indikator Capaian

- Mahasiswa dapat menjelaskan potensi dampak scraping terhadap beban server dan penyalahgunaan informasi
- Mahasiswa dapat mengidentifikasi teknik lazy loading yang diterapkan pada elemen gambar di halaman web
- Mahasiswa dapat menyimpan data gambar ke dalam array list yang terstruktur
- Mahasiswa dapat menentukan dan mengatur direktori untuk menyimpan gambar yang diambil

1.3 Landasan Teori

Web scraping merupakan teknik yang digunakan untuk mengambil data secara otomatis dari sebuah situs web yang tidak jarang melibatkan pengambilan elemen gambar. Scraping yang dilakukan secara agresif dapat membebani server bahkan kerusakan pada server yang ditargetkan. Oleh karena itu, penting untuk mempertimbangkan etika dalam pengambilan data dengan memastikan bahwa scraping yang dilakukan menggunakan cara yang tidak merugikan pihak manapun.

Selain etika dalam pengambilan data, teknik lazy loading adalah konsep penting yang perlu dipahami dalam scraping gambar. Teknik ini mengurangi waktu

muat halaman dan penggunaan bandwidth, sehingga lebih efisien dibandingkan dengan memuat semua elemen ketika halaman pertama kali dimuat.

Dalam hal pengambilan gambar, library seperti BeautifulSoup di Python sering digunakan untuk mengekstrak tag img pada halaman web. Penyimpanan hasil scraping juga memerlukan perhatian khusus, terutama terkait pengaturan direktori. Penggunaan fungsi seperti path.join() di Python memungkinkan gambar disimpan di direktori yang sesuai, sehingga memudahkan manajemen file dalam proyek yang lebih besar dan menghindari duplikasi file.

Dengan memadukan pemahaman etika, teknik lazy loading, dan pengelolaan penyimpanan yang efisien, mahasiswa dapat melakukan web scraping secara terstruktur dan dapat dipertanggungjawabkan.

1.4 Pelaksanaan Praktikum

1.4.1 Percobaan

Pada percobaan ini, mahasiswa diperintahkan untuk melakukan analisa pada website lain untuk mengambil data gambar dan melakukan scraping. Dalam percobaan ini, telah dilakukan percobaan pada website liputan6 dengan kategori pemilu. Berikut kode scraping yang telah dilakukan.

a. Script / Setting Program

MainFungsi.py

```
import os
def CreateDirectory(namaFolder):
    if not os.path.exists(namaFolder):
        os.makedirs(namaFolder);

def CreateNewFile(path):
    f = open(path, "w");
    f.write("");
    f.close();

def WriteToFile(path, data):
    with open(path, "a") as file:
        file.write(data + "\n");
```

```
def DoesFileExist(path):
    return os.path.isfile(path)

def WriteToFile2(path, data, response):
    fullPath = os.path.join(path, data)
    with open(fullPath, 'wb') as f:
        f.write(response.content)
```

coba2.py

```
import requests
import MainFungsi
import os
from bs4 import BeautifulSoup

url = 'https://www.liputan6.com/pemilu'
page = requests.get(url)
soup = BeautifulSoup(page.content, 'html.parser')
datax = soup.find_all('img')

images = []

for img in datax:
    img_url = img.get('src')
    if img_url is not None and img_url.endswith(('.jpg', '.png', '.gif', '.webp', '.jpeg')):
        images.append(img_url)

print(images)

direktori = "Pertemuan 9/Hasil Gambar"
MainFungsi.CreateDirectory(direktori)

for gmb in images:
    response = requests.get(gmb)
    fileName = os.path.basename(gmb)
    MainFungsi.WriteToFile2(direktori, fileName, response)
    print(response)
    print(fileName)
```

Penjelasan Kode MainFungsi.py

1. Mengimpor modul os yang digunakan untuk berinteraksi dengan sistem operasi
2. def CreateDirectory(namaFolder) berfungsi untuk membuat folder jika folder dengan namaFolder belum ada

3. def CreateNewFile(path) berfungsi untuk membuat file baru dengan path yang diberikan (path)
4. def WriteToFile(path, data) berfungsi untuk menambahkan data pada file yang sudah ada
5. def DoesFileExist(path) berfungsi untuk memeriksa apakah file ada di dalam path yang sudah diberikan
6. def WriteToFile2(path. Data, response) berfungsi untuk menulis konten dari response ke file dengan nama file yang disesuaikan

Penjelasan Kode coba2.py

1. import requests berfungsi untuk melakukan HTTP requests seperti mengunduh halaman web atau gambar
2. import MainFungsi memanggil file MainFungsi.py yang berisi fungsi-fungsi untuk mmbuat folder, file, dan menambahkan data ke file
3. import os berfungsi untuk berinteraksi dengan sistem file seperti manipulasi path
4. from bs4 import BeautifulSoup berfungsi untuk memparsing HTML dan mengekstrak data seperti gambar
5. url = 'https://www.liputan6.com' adalah web yang akan diambil isinya. Semisal sasaran web nya berubah, tinggal mengubah isi value url tersebut
6. page = requests.get(url) berfungsi untuk mengambil konten halaman web dari url yang telah ditentukan sebelumnya
7. soup = BeautifulSoup(page.content, 'html.parser') berfungsi untuk memparsing kontenh html dari halaman yang diunduh
8. datax = soup.find_all('img') berfungsi untuk menemukan semua elemen gambar (<img>) dalam halaman web
9. images = [ ] menyiapkan list kosong untuk menampung URL gambar yang ditemukan
10. for img in datax adalah looping melalui semua elemen gambar yang ditemukan

11. img_url = img.get(‘src’) mengambil atribut src dari setiap elemen gambar yang menunjukkan url gambar
12. if img_url is not None and img_url.endswith((‘.jpg’, ‘.png’, ‘.gif’, ‘.webp’, ‘.jpeg’)) mengecek apakah url gambar valid dan apakah ekstensi file gambar sudah sesuai
13. images.append(img_url) menambahkan url gambar yang valid ke dalam list images
14. print(images) menampilkan daftar URL gambar yang ditemukan di terminal
15. direktori = “Pertemuan 9/Hasil Gambar” menetapkan path folder untuk menyimpan gambar yang diunduh
16. MainFungsi.CreateDirectory(direktori) membuat folder untuk menyimpan gambar jika folder tersebut belum ada
17. for gmb in images adalah looping melalui setiap url gambar dalam list images
18. response = requests.get(gmb) mengunduh gambar dari url yang diberikan
19. fileName = os.path.basename(gmb) mengambil nama file gambar dari url (misalnya image.jpg)
20. MainFungsi.WriteToFile2(direktori, fileName, response) menyimpan gambar yang diunduh ke dalam folder yang ditentukan dengan nama file yang sesuai
21. print(response) menampilkan informasi mengenai response HTTP seperti status code di terminal
22. print(fileName) menampilkan nama file yang telah diunduh di terminal

b. Langkah Uji Coba

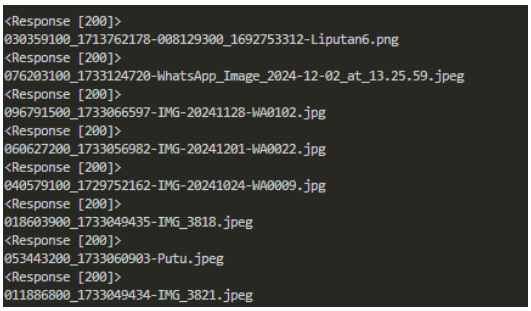
1. Pastikan Python dan library yang diperlukan terinstal dan dapat digunakan.
2. Siapkan file MainFungsi.py dan coba2.py
3. Pastikan folder Pertemuan 9/Hasil Gambar belum ada sebelumnya. Jika belum ada, folder akan dibuat secara otomatis
4. Jalankan program dengan menjalankan file coba2.py

5. Program akan mengakses halaman web https://www.liputan6.com/pemilu dan mengirimkan permintaan HTTP menggunakan library requests
6. Gambar yang valid akan disimpan di dalam folder Pertemuan 9/Hasil Gambar dengan ekstensi yang sudah ditentukan di awal.
7. Verifikasi output dengan memastikan respons HTTP adalah 200
8. Buka folder Pertemuan 9/Hasil Gambar dan pastikan gambar-gambar yang valid telah terunduh
9. Setelah semua gambar terunduh, program akan berhenti dan tidak akan mengunduh gambar lebih lanjut

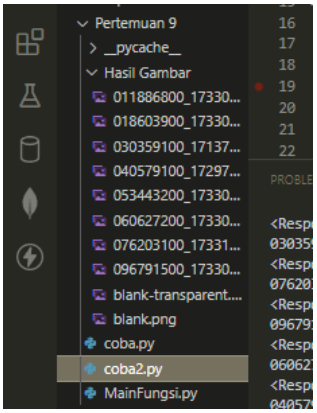
c. Hasil Uji Coba

1. Daftar gambar telah ditemukan

2. Gambar dapat diunduh



3. Folder Pertemuan 9/Hasil Gambar berhasil dibuat dan gambar dapat ditambahkan



d. Analisa Hasil

1. Fungsi pencarian gambar (soup.find_all('img')) berhasil mengekstrak elemen gambar dari halaman HTML. Namun tidak ada jaminan kalau semua gambar yang ditemukan ekstensi file nya valis. Oleh karena itu, filter ekstensi file berfungsi dengan baik untuk memastikan hanya gambar dengan ekstensi yang tepat dan sesuai kebutuhan yang akan diunduh.
2. Pembuatan folder berjalan dengan baik. Kode memeriksa apakah folder sudah ada atau belum yang dapat mencegah kesalahan jika folder sudah ada
3. Kode berhasil mengunduh dan menyimpan gambar ke dalam folder yang ditentukan.
4. Tidak ada kesalahan atau pengecualian yang muncul dalam proses pengunduhan gambar. Semua gambar yang valid berhasil diunduh dengan status HTTP 200

1.5 Kesimpulan

1.5.1 Kesimpulan Percobaan

Secara keseluruhan, kode ini berhasil melaksanakan tugas utamanya, yaitu mengambil gambar dari halaman web https://www.liputan6.com/pemilu dan menyimpannya ke folder yang sudah dibuat. Hasil uji coba menunjukkan bahwa skrip bekerja dengan baik dalam mengekstrak gambar dan menyimpannya dengan benar. Meskipun demikian, ada beberapa aspek yang bisa diperbaiki, seperti penanganan kesalahan, pengelolaan URL relatif, dan peningkatan efisiensi.