

LAPORAN TUGAS HARIAN 2

DATA SCRAPING

Disusun untuk Memenuhi Tugas Harian 2 Mata Kuliah Data Scraping
Dibimbing oleh Pelsri Ramadar N.S. M.Kom



Oleh:

Sherly Puput Rahmawati (1123102147)

PROGRAM STUDI S1 TEKNIK INFORMATIKA
SEKOLAH TINGGI ILMU KOMPUTER PGRI BANYUWANGI
2022

1. Kode Program (MainFungsi.py)

```
import os

def CreateDirectory(path):
    if not os.path.exists(path):
        os.makedirs(path)

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

Penjelasan Kode Program:

Kode program MainFungsi.py berisi dua fungsi utama untuk menangani pembuatan direktori dan penyimpanan file. Fungsi pertama adalah `CreateDirectory`, yang bertujuan untuk memeriksa apakah sebuah direktori sudah ada di lokasi tertentu, dan jika belum, direktori tersebut akan dibuat menggunakan `os.makedirs`. Ini memastikan bahwa setiap file atau data yang akan disimpan memiliki folder tujuan yang valid. Fungsi kedua adalah `WriteToFile2`, yang digunakan untuk menyimpan file dalam mode biner. Fungsi ini menerima tiga parameter: path untuk lokasi folder tempat file disimpan, data sebagai nama file, dan response yang berisi konten file dari permintaan HTTP. Path lengkap file dibentuk menggunakan `os.path.join`, kemudian file dibuka dalam mode `wb` (write binary), sehingga konten file dari response dapat disimpan.

2. Kode Program (img.py)

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

url = 'https://www.lazada.co.id'
headers = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4664.110 Safari/537.36"
}
page = requests.get(url, headers=headers)
soup = BeautifulSoup(page.content, 'html.parser')

datax = soup.find_all('img')

images = []
```

```

for img in datax:
    img_url = img.get('data-lazy-src') or img.get('src')
    if img_url and img_url.startswith('http') and img_url.endswith((''.jpg',
'.png', '.gif')):
        images.append(img_url)

print("Gambar yang ditemukan:", images)

folder_tugas = "Tugas 2"
direktori = os.path.join(folder_tugas, "Hasil Gambar")

if not os.path.exists(folder_tugas):
    os.makedirs(folder_tugas)

if not os.path.exists(direktori):
    os.makedirs(direktori)

for gmb in images:
    try:
        response = requests.get(gmb, headers=headers, stream=True)
        if response.status_code == 200:
            filename = os.path.basename(gmb.split('?')[0])
            alt_text = img.get('alt') or 'default_name'
            filename = f"{alt_text}_{filename}"
            file_path = os.path.join(direktori, filename)
            with open(file_path, 'wb') as f:
                f.write(response.content)
            print(f"Sukses menyimpan: {filename}")
        else:
            print(f"Gagal mengunduh: {gmb}")
    except Exception as e:
        print(f"Error saat mengunduh {gmb}: {e}")

```

Penjelasan Kode Program:

Kode program `img.py` digunakan untuk scraping gambar dari website Lazada. Pertama, modul seperti `os` untuk manipulasi direktori, `MainFungsi` untuk menggunakan fungsi tambahan, `requests` untuk melakukan HTTP request, dan `BeautifulSoup` untuk parsing HTML diimpor. Program ini dimulai dengan mendefinisikan URL Lazada dan sebuah header HTTP `User-Agent`, yang membantu menyamarkan request sebagai permintaan dari browser biasa untuk menghindari blokir server. Selanjutnya, halaman web dimuat dengan `requests.get`, dan kontennya diparsing menjadi objek BeautifulSoup agar elemen HTML dapat diakses.

Program kemudian mencari semua elemen `` pada halaman menggunakan `soup.find_all('img')`. Semua URL gambar diekstrak dari atribut `data-lazy-src` atau `src` dengan validasi tambahan agar hanya URL yang dimulai dengan "http" dan memiliki ekstensi gambar seperti `.jpg`, `.png`, atau `.gif` yang dimasukkan ke dalam daftar `images`. Daftar ini dicetak untuk memastikan URL gambar ditemukan.

Program membuat direktori bernama "Tugas 2" dan subdirektornya "Hasil Gambar" untuk menyimpan file gambar. Hal ini dilakukan dengan mengecek apakah direktori tersebut sudah ada menggunakan `os.path.exists`. Jika tidak ada, maka direktori dibuat menggunakan `os.makedirs`.

Setelah itu, program mengunduh setiap gambar dalam daftar `images` menggunakan loop. Untuk setiap URL, request HTTP dikirim menggunakan `requests.get`. Jika respons HTTP berhasil (status kode 200), nama file diambil dari URL dengan `os.path.basename`, sementara nama alternatif (`alt_text`) digunakan sebagai fallback jika tidak ada nama gambar dalam URL. File kemudian disimpan ke dalam direktori yang telah dibuat sebelumnya. Jika terjadi kesalahan selama proses pengunduhan, seperti URL yang tidak dapat diakses atau masalah jaringan, pesan error akan dicetak untuk gambar tersebut.