Nama : Ahmad Rafi Wirana
NPM : 2006595873
Mata Kuliah : Integrasi Aplikasi Perusahaan

Dokumentasi Assignment 4 - Web Scraping

Panduan Menjalankan Assignment 4 - Web Scraping

1. Konfigurasi Environment Virtual dan Instalasi Package
    a. Buka *Terminal* atau *Command Prompt.*
    b. Navigasikan ke direktori folder 2006595873_Assigment4_Source_code pada komputer lokal.
    c. Buat environment virtual Python jika belum dibuat:
        i. python3 -m venv env; atau
        ii. python -m venv env
    d. Aktifkan environment virtual:
        i. Pada Unix atau MacOS:
            1. source env/bin/activate
        ii. Pada Windows:
            1. env\Scripts\activate
    e. Instalasi Library:
        i. pip3 install Flask requests beautifulsoup4; atau
        ii. pip install Flask requests beautifulsoup4
2. Eksekusi dan Pengujian Program
    a. Menjalankan aplikasi Flask:
        i. flask run
    b. Pengujian hasil:
        i. Lakukan pengujian kode program dengan membuka URL: http://127.0.0.1:5000/ di browser lokal komputer.
        ii. Kemudian, lakukan pengujian juga dengan membuka file csv yang dibuat di dalam directory 2006595873_Assigment4_Source_code untuk memastikan data diambil dan disimpan dengan benar sesuai dengan kriteria yang diberikan.

Nama : Ahmad Rafi Wirana
NPM : 2006595873
Mata Kuliah : Integrasi Aplikasi Perusahaan

3. Dokumentasi Struktur Program Web Scraping
    a. Deskripsi Umum
        i. Struktur program untuk proyek web scraping ini dirancang untuk mengumpulkan informasi tentang buku dari situs web Books to Scrape. Program ini mengambil data seperti judul buku, gambar, rating, harga, dan ketersediaan stok.
    b. File Utama
        i. app.py: File ini adalah script utama yang berisi kode untuk melakukan web scraping, pemrosesan data, dan penyimpanan output ke dalam file CSV.
        ii. templates/index.html: File ini adalah kode html yang menampilkan input teks dan juga dropdown tag yang berisi kategori yang ada di dalam website https://books.toscrape.com/.
        iii. templates/results.html: File ini adalah kode html yang menampilkan output buku yang sudah di-*scraping*.
    c. Libraries yang Digunakan
        i. Requests: Library ini digunakan untuk membuat HTTP requests ke situs web target. Hal ini memungkinkan program untuk mengunduh konten halaman yang akan diproses.
        ii. BeautifulSoup: Library ini digunakan untuk parsing HTML. Library ini memfasilitasi ekstraksi data spesifik dari halaman web yang di-scrape, seperti judul buku, rating, dan harga.
        iii. CSV: Module ini digunakan untuk menulis data yang telah di-extract ke dalam file CSV, yang memungkinkan penyimpanan data secara terstruktur dan dapat diakses dengan mudah menggunakan berbagai aplikasi pengolah data.
    d. Struktur Kode:
        i. Fungsi Scraping Utama (scrape_books): Fungsi ini bertanggung jawab untuk menghubungkan ke situs web, mengunduh konten HTML, dan

mengekstrak data menggunakan BeautifulSoup. Fungsi ini juga menyaring buku berdasarkan judul dan kategori.

ii. Konversi Rating (convert_rating): Fungsi ini mengonversi rating dari format teks (misal, 'Three') menjadi format numerik dengan kata 'bintang' untuk kemudahan pembacaan dan konsistensi dengan output yang diinginkan.

iii. Pengurutan dan Penyimpanan Data (save_books_to_csv): Setelah data diekstrak dan diolah, fungsi ini menulis data ke dalam file CSV. Data disusun dalam format yang mudah dibaca, dengan setiap buku muncul sebagai baris dalam CSV, termasuk judul, rating, harga, dan informasi stok.

iv. Pengurutan Buku sesuai dengan rating (sort_books_by_rating): Fungsi ini bertujuan untuk mengurutkan buku yang sudah di-*scraping* dan diurutkan berdasarkan ratingnya, dari yang terbaik (5 bintang) ke yang terburuk (1 bintang).

v. Pengambilan data categories (fetch_categories): Fungsi ini mengambil daftar kategori buku dari situs web "https://books.toscrape.com/" dan menyimpannya secara lokal. Jika kategori sudah diambil sebelumnya, maka fungsi akan mengembalikan kategori yang telah disimpan tersebut dari cache. Jika belum, fungsi akan mengambil kategori dari situs web tersebut, kemudian menyaring kategori 'Erotica', dan menyimpannya.