**LAPORAN**

**RENCANA TUGAS MANDIRI (RTM) Ke-2**

**MATA KULIAH BIG DATA**

**“SCRAPING DATA DARI GOOGLE”**



**DISUSUN OLEH:**

Citra Amelia Intan Permadani (21083010004)

**DOSEN PENGAMPU:**

Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

**PROGRAM STUDI SAINS DATA**

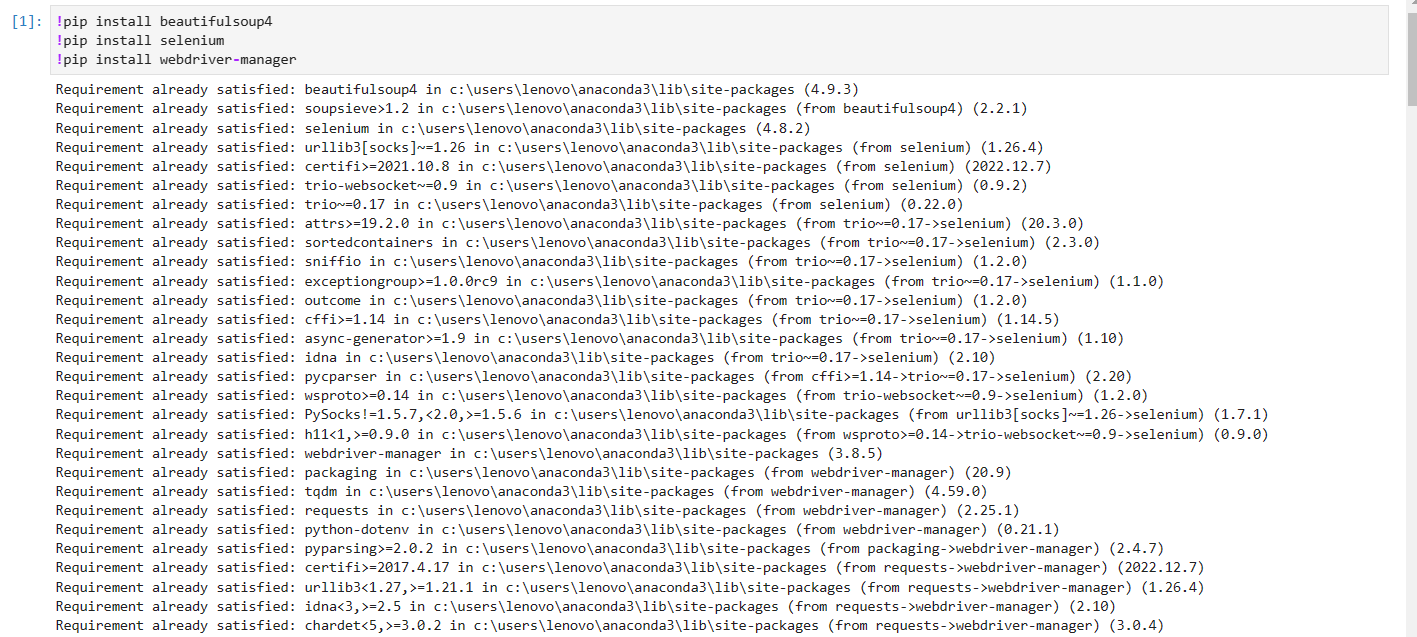
**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR**

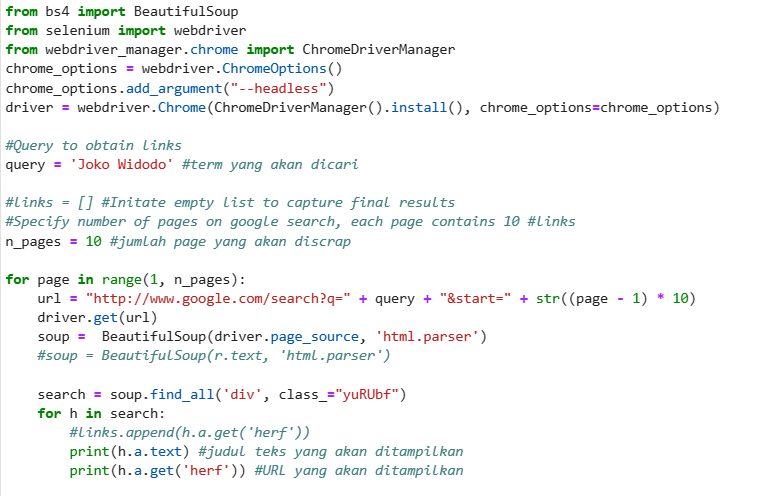
**2022**

Scraping Data dari Google

1. Sebelum melakukan scripting kita install terlebih dahulu library di bawah ini



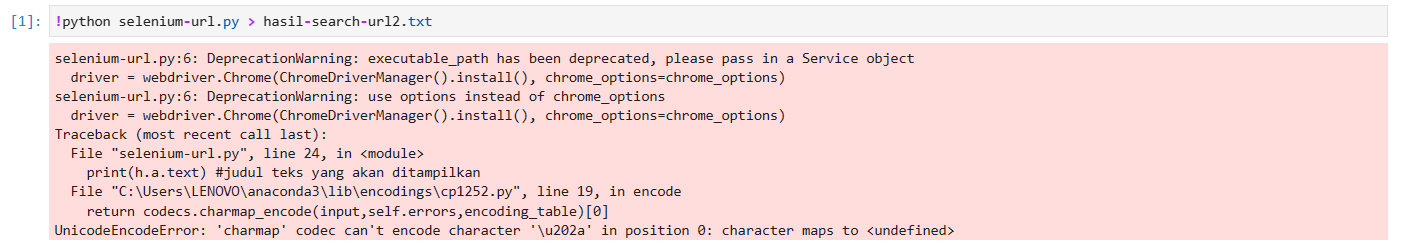
1. Kita buat terlebih dahulu modul dengan nama selenium-url.py diisi dengan script seperti di bawah ini



Ket: dari script di atas kita meng-import BeautifulSoup, webdriver dari selenium dan import Chrome Drive Manager. Lalu kita definisikan query yang akan dicari di google, disini kita menggunakan query Joko Widodo serta menentukan jumlah halaman yang akan dipotong.

Dalam looping for, kode URL yang dibuat digunakan untuk setiap halaman hasil penelusuran dan mengarahkan menggunakan driver Chrome. Kemudian konten HTML tersebut di ekstrak menggunakan BeautifulSoup dan mencari elemen div dengan kelas “yuRUbf” yang berisi judul dan URL dari setiap hasil pencarian.

Lalu kita run terlebih dahulu dan buat lembar baru lalu ketikkan



Maka lembar hasil-search-url2.txt akan terbuat dan hasil url dengan kata kunci Joko Widodo dari google ada didalamnya. Jika kita meng-run selenium-url.py maka hasil URL akan diletakkan di hasil-search-url.txt



1. Selanjutnya, kita buat txt baru untuk memasukkan URL terpilih yang akan dibuka isinya



1. Kemudian, kita buat modul baru untuk membuka isi URL yang dimasukkan di atas dan kemudian kita buat kode script seperti dibawah ini



Ket: script berjalan dengan langkah pertama mengimport pustaka yang diperlukan lalu menyiapkan driver Chrome menggunakan Chrom Drive Manager dan Options dan mendefinisikan fungsi untuk menjelaskan argumen baris perintah menggunakan library argparse. Skrip diatas memiliki fungsi utama untuk membaca daftar URL dari file inputan dan mengulanginya menggunakan looping for. Setiap URL akan ditujukan ke halaman menggunakan driver Chrome dan menemukan isinya menggunakan metode find\_element dari objek driver. Kemudian teks elemen body dikodekan sebagai UTF-8 dan menambahkannya ke daftar yakni full\_text. Setelah semua URL diproses, script menulis isi full\_text ke file output dan menutup driver Chrome dan file output.

Modul diatas dirun lalu buka file python baru kemudian ketikkan script di bawah ini



Dari script di atas kita memasukkan -i sebagai inputan berisi URL di atas dan -o berisi output yang akan dikeluarkan

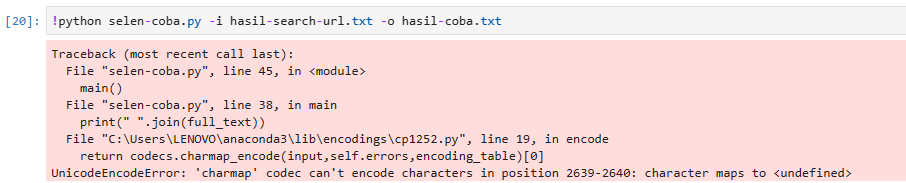


1. Selanjutnya adalan membersihkan hasil browser dengan script di bawah ini



Ket: script di atas untuk membersihkan hasil browser dari kata-kata yang tidak diperlukan serta bagaimana memisahkan kalimat hasil browser dari kedua URL yang dimasukkan

Lalu script di atas di run dan ketikkan script dibawah ini untuk mengeluarkan hasil cleaning text



Maka akan muncul hasilnya di file hasil-coba.txt



Namun, hasilnya belum cukup sesuai dengan yang diinginkan dan masih belum rapi juga masih terdapat kata-kata yang tidak diperlukan.

Namun, jika memakai manual seperti dibawah ini maka akan cukup bersih dan rapi akan tetapi harus memasukkan URL satu per satu

