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

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

```
[1]: !pip install beautifulsoup4
     !pip install selenium
     !pip install webdriver-manager
```

```
Requirement already satisfied: beautifulsoup4 in c:\users\lenovo\anaconda3\lib\site-packages (4.9.3)
Requirement already satisfied: soupsieve>1.2 in c:\users\lenovo\anaconda3\lib\site-packages (from beautifulsoup4) (2.2.1)
Requirement already satisfied: selenium in c:\users\lenovo\anaconda3\lib\site-packages (4.8.2)
Requirement already satisfied: urllib3[socks]~=1.26 in c:\users\lenovo\anaconda3\lib\site-packages (from selenium) (1.26.4)
Requirement already satisfied: certifi>=2021.10.8 in c:\users\lenovo\anaconda3\lib\site-packages (from selenium) (2022.12.7)
Requirement already satisfied: trio-websocket~=0.9 in c:\users\lenovo\anaconda3\lib\site-packages (from selenium) (0.9.2)
Requirement already satisfied: trio~=0.17 in c:\users\lenovo\anaconda3\lib\site-packages (from selenium) (0.22.0)
Requirement already satisfied: attrs>=19.2.0 in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (20.3.0)
Requirement already satisfied: sortedcontainers in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (2.3.0)
Requirement already satisfied: sniffio in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.2.0)
Requirement already satisfied: exceptiongroup>=1.0.0rc9 in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.1.0)
Requirement already satisfied: outcome in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.2.0)
Requirement already satisfied: cffi>=1.14 in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.14.5)
Requirement already satisfied: async-generator>=1.9 in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (1.10)
Requirement already satisfied: idna in c:\users\lenovo\anaconda3\lib\site-packages (from trio~=0.17->selenium) (2.10)
Requirement already satisfied: pycparser in c:\users\lenovo\anaconda3\lib\site-packages (from cffi>=1.14->trio~=0.17->selenium) (2.20)
Requirement already satisfied: wsproto>=0.14 in c:\users\lenovo\anaconda3\lib\site-packages (from trio-websocket~=0.9->selenium) (1.2.0)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in c:\users\lenovo\anaconda3\lib\site-packages (from urllib3[socks]~=1.26->selenium) (1.7.1)
Requirement already satisfied: h11<1,>=0.9.0 in c:\users\lenovo\anaconda3\lib\site-packages (from wsproto>=0.14->trio-websocket~=0.9->selenium) (0.9.0)
Requirement already satisfied: webdriver-manager in c:\users\lenovo\anaconda3\lib\site-packages (3.8.5)
Requirement already satisfied: packaging in c:\users\lenovo\anaconda3\lib\site-packages (from webdriver-manager) (20.9)
Requirement already satisfied: tqdm in c:\users\lenovo\anaconda3\lib\site-packages (from webdriver-manager) (4.59.0)
Requirement already satisfied: requests in c:\users\lenovo\anaconda3\lib\site-packages (from webdriver-manager) (2.25.1)
Requirement already satisfied: python-dotenv in c:\users\lenovo\anaconda3\lib\site-packages (from webdriver-manager) (0.21.1)
Requirement already satisfied: pyparsing>=2.0.2 in c:\users\lenovo\anaconda3\lib\site-packages (from packaging->webdriver-manager) (2.4.7)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\lenovo\anaconda3\lib\site-packages (from requests->webdriver-manager) (2022.12.7)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\lenovo\anaconda3\lib\site-packages (from requests->webdriver-manager) (1.26.4)
Requirement already satisfied: idna<3,>=2.5 in c:\users\lenovo\anaconda3\lib\site-packages (from requests->webdriver-manager) (2.10)
Requirement already satisfied: chardet<5,>=3.0.2 in c:\users\lenovo\anaconda3\lib\site-packages (from requests->webdriver-manager) (3.0.4)
```

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

```python
from bs4 import BeautifulSoup
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(ChromeDriverManager().install(), chrome_options=chrome_options)

#Query to obtain links
query = 'Joko Widodo' #term yang akan dicari

#links = [] #Initate empty list to capture final results
#Specify number of pages on google search, each page contains 10 #links
n_pages = 10 #jumlah page yang akan discrap

for page in range(1, n_pages):
    url = "http://www.google.com/search?q=" + query + "&start=" + str((page - 1) * 10)
    driver.get(url)
    soup =  BeautifulSoup(driver.page_source, 'html.parser')
    #soup = BeautifulSoup(r.text, 'html.parser')

    search = soup.find_all('div', class_="yuRUbf")
    for h in search:
        #links.append(h.a.get('herf'))
        print(h.a.text) #judul teks yang akan ditampilkan
        print(h.a.get('herf')) #URL yang akan ditampilkan
```

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

```
[1]: !python selenium-url.py > hasil-search-url2.txt

selenium-url.py:6: DeprecationWarning: executable_path has been deprecated, please pass in a Service object
  driver = webdriver.Chrome(ChromeDriverManager().install(), chrome_options=chrome_options)
selenium-url.py:6: DeprecationWarning: use options instead of chrome_options
  driver = webdriver.Chrome(ChromeDriverManager().install(), chrome_options=chrome_options)
Traceback (most recent call last):
  File "selenium-url.py", line 24, in <module>
    print(h.a.text) #judul teks yang akan ditampilkan
  File "C:\Users\LENOVO\anaconda3\lib\encodings\cp1252.py", line 19, in encode
    return codecs.charmap_encode(input,self.errors,encoding_table)[0]
UnicodeEncodeError: 'charmap' codec can't encode character '\u202a' in position 0: character maps to <undefined>
```

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

```
Joko Widodo - Wikipedia bahasa Indonesia, ensiklopedia ...https://id.wikipedia.org › wiki › Joko...
None
Presiden Joko Widodohttps://www.presidenri.go.id › presid...
None
Laman Resmi Presiden Republik Indonesia • Presiden RIhttps://www.presidenri.go.id
None
Joko Widodo (@jokowi) • Instagram photos and videoshttps://www.instagram.com › jokowi
None
Joko Widodo (@jokowi) / Twitterhttps://twitter.com › jokowi
None
Presiden Joko Widodo - Facebookhttps://www.facebook.com › ... › Presiden Joko Widodo
None
Presiden Joko Widodo - YouTubehttps://www.youtube.com › channel
None
Berita dan Informasi Joko widodo Terkini dan Terbaru Hari inihttps://www.detik.com › tag › joko-...
None
Sekretariat Kabinet Republik Indonesia | Kabinet Indonesia ...https://setkab.go.id
None
Herbertus Handoko Joko Widodo - Kominfohttps://www.kominfo.go.id › detail
None
Berita Terbaru Joko Widodo - Tempo.cohttps://www.tempo.co › tag › joko-w...
None
Berita Harian Jokowi - CNN Indonesiahttps://www.cnnindonesia.com › tag
None
Berita Presiden dan Pemerintahan - Sekretariat Negarahttps://www.setneg.go.id › listberita
None
#Joko Widodo | Kemenaghttps://kemenag.go.id › tags › tags=J...
None
Prof. Dr. Joko Widodo, M. Pd. - Universitas Negeri Semaranghttps://staff.unnes.ac.id › dosen › jok...
None
Presiden Joko Widodo Berikan Perhatian Tinggi pada ... - PPIDhttp://ppid.menlhk.go.id › browse
```

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

```
1 https://id.wikipedia.org/wiki/Joko_Widodo
2 https://www.detik.com/tag/joko-widodo
```

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

```python
from selenium import webdriver
import sys, getopt
import argparse
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By

options = Options()
options.add_argument("--headless")
driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()),options=options)
full_text=[]

def parse_args():
    parser = argparse.ArgumentParser()
    parser.add_argument('-i', '--infile', default='', help='input filename')
    parser.add_argument('-o', '--outfile', default='', help='output filename')
    return parser.parse_args()

def main():
    args = parse_args()
    outfile = args.outfile
    infile = args.infile

    with open(infile) as f:
        content = f.read().splitlines()
    f.close()

    f=open(outfile,"w")
    for u in content:
        driver.get(u)
        elems = driver.find_element(By.TAG_NAME, 'body').text.encode("utf-8")
        full_text.append(elems)
    print(full_text)

    f.write(str(full_text))

    #print(full_text)
    driver.close()
    f.close()
main()
```

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

```
[2]: !python selenium-browse.py -i hasil-search-url.txt -o hasil-browse-url2.txt
```

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

```
[b'Lompat ke isi\nBuka/tutup bilah samping\nPencarian\nBuat akun baru\nMasuk log\nPerkakas pribadi\nAyo ikut kompetisi Proyek Yuwana di Wikibuku bahasa
Indonesia!\nPeriode pendaftaran 1 Januari\xe2\x80\x9330 April 2023. Lihat syarat dan ketentuannya.\nIkuti Wikipedia bahasa Indonesia di Facebook, Twitter, Instagram,
dan Telegram\nToggle the table of contents\nJoko Widodo\n98 bahasa\nHalaman\nPembicaraan\nBaca\nLihat sumber\nLihat riwayat\nDari Wikipedia bahasa Indonesia,
ensiklopedia bebas\nPenyuntingan Artikel oleh pengguna baru atau anonim untuk saat ini tidak diizinkan.\nLihat kebijakan pelindungan dan log pelindungan untuk informasi
selengkapnya. Jika Anda tidak dapat menyunting Artikel ini dan Anda ingin melakukannya, Anda dapat memohon permintaan penyuntingan, diskusikan perubahan yang ingin
dilakukan di halaman pembicaraan, memohon untuk melepaskan pelindungan, masuk, atau buatlah sebuah akun.\nIr. H.\nJoko Widodo\nPresiden Indonesia ke-7\nPetahana\nMulai
menjabat\n20 Oktober 2014\nWakil Presiden Jusuf Kalla (2014\xe2\x80\x932019)\nMa\'ruf Amin (2019\xe2\x80\x93)\nPendahulu Susilo Bambang Yudhoyono\nGubernur DKI Jakarta
ke-14\nMasa jabatan\n15 Oktober 2012 \xe2\x80\x93 16 Oktober 2014\nWakil Gubernur Basuki Tjahaja Purnama\nPendahulu Fauzi Bowo\nFadjar Panjaitan (Plt.)[1]\nPengganti
Basuki Tjahaja Purnama\nWali Kota Surakarta ke-16\nMasa jabatan\n28 Juli 2005 \xe2\x80\x93 1 Oktober 2012\nPresiden Susilo Bambang Yudhoyono\nGubernur Mardiyanto\nAli
Mufiz\nBibit Waluyo\nWakil F.X. Hadi Rudyatmo\nPendahulu Slamet Suryanto\nAnwar Cholil (Plh.)[2]\nPengganti F.X. Hadi Rudyatmo\nKetua Perhimpunan Bangsa-bangsa Asia
Tenggara\nPetahana\nMulai menjabat\n01 Januari 2023\nPendahulu Hun Sen\nInformasi pribadi\nLahir Mulyono\n21 Juni 1961 (umur 61)\nSurakarta, Jawa Tengah,
Indonesia[3]\nKebangsaan Indonesia\nPartai politik  PDI-P\nSuami/istri Iriana\nAnak\nGibran Rakabuming Raka[4]\nKahiyang Ayu[5]\nKaesang Pangarep[5]\nOrang
tua\nWidjiatno Notomihardjo (ayah)\nSudjiatmi[6] (ibu)\nKerabat\nBobby Nasution (menantu)\nAnwar Usman (ipar)\nAlma mater Universitas Gadjah Mada\nPekerjaan
Pengusaha\nPolitikus\nTanda tangan\nSitus web Website Resmi Presiden RI\nArtikel ini merupakan bagian dari seri\nJoko Widodo\nSebelum menjadi presiden\nWali Kota
Surakarta BST Pilkada Jakarta Gubernur DKI Jakarta LRT MRT\nPresiden Indonesia\nPetahana\nPilpres 2014 (kampanye) Pilpres 2019 (kampanye) Pelantikan I Pelantikan II
Kepresidenan Kabinet Kerja Kabinet Indonesia Maju\nKebijakan\nBali Nine Tol Laut Kereta cepat Trans-Sumatra Ibu kota baru\nKTT yang Dihadiri\nKTT APEC 2014 KTT ASEAN
2014 KTT ASEAN 2015 G20 (2014, 2015, 2016, 2017, 2019, 2020, 2021)\nKeluarga\nSilsilah keluarga Iriana Joko Widodo (istri dan Ibu Negara) Anak Gibran Rakabuming Raka
Kahiyang Ayu Kaesang Pangarep\nSitus Web\nSitus Kepresidenan\nMedia sosial\nFacebook Twitter Instagram YouTube\nlbs\nIr. H. Joko Widodo (pengucapan bahasa Indonesia:
[d\xca\x92\xc9\x94k\xc9\x94 wid\xc9\x94d\xc9\x94]; lahir 21 Juni 1961) adalah presiden Indonesia yang mulai menjabat sejak tanggal 20 Oktober 2014. Terpilih dalam
Pemilu Presiden 2014, Jokowi menjadi presiden Indonesia pertama yang bukan berasal dari elite politik atau militer Indonesia. Dia terpilih bersama Wakil Presiden Jusuf
Kalla dan kembali terpilih bersama Wakil Presiden Ma\'ruf Amin dalam Pemilu Presiden 2019. Sebelumnya, Jokowi pernah menjabat sebagai gubernur Jakarta sejak 15 Oktober
2012 hingga 16 Oktober 2014 didampingi Basuki Tjahaja Purnama sebagai wakil gubernur. Sebelumnya, ia adalah wali kota Surakarta, sejak tanggal 28 Juli 2005 hingga 1
Oktober 2012, didampingi F.X. Hadi Rudyatmo sebagai wakil wali kota.[7] Dua tahun menjalani periode keduanya menjadi wali kota Surakarta, Jokowi ditunjuk oleh
partainya, Partai Demokrasi Indonesia Perjuangan (PDI-P), untuk bersaing dalam Pilkada DKI Jakarta 2012 berpasangan dengan Basuki Tjahaja Purnama.[8]\nJoko Widodo
berasal dari keluarga sederhana, rumahnya pernah digusur sebanyak tiga kali ketika dia masih kecil,[9] tetapi dia mampu menyelesaikan sekolahnya di Fakultas Kehutanan
Universitas Gadjah Mada. Setelah lulus, dia menekuni profesinya sebagai pengusaha mebel.[9] Karier politiknya dimulai dengan menjadi wali kota Surakarta pada tahun
2005.[10] Namanya mulai dikenal setelah dianggap berhasil mengubah wajah Surakarta menjadi kota pariwisata, kota budaya, dan kota batik yang populer.[11] Pada tanggal
20 September 2012, Jokowi berhasil memenangi Pilkada Jakarta 2012. Kemenangannya dianggap mencerminkan dukungan populer untuk seorang pemimpin yang "muda" dan "bersih",
meskipun umurnya sudah lebih dari 50 tahun.[12]\nSemenjak terpilih sebagai gubernur, popularitasnya terus naik dan menjadi sorotan media.[13][14] Akibatnya, muncul
wacana untuk menjadikannya calon presiden untuk Pemilu Presiden 2014.[15] Ditambah lagi, hasil survei menunjukkan nama Jokowi selalu unggul.[16] Pada awalnya, Ketua
Umum PDI-P Megawati Soekarnoputri menyatakan bahwa dia tidak akan mengumumkan calon presiden dari PDI-P sampai setelah Pemilu Legislatif 2014.[17] Pada tanggal 14 Maret
2014, Jokowi menerima mandat dari Megawati untuk maju sebagai calon presiden, tiga pekan sebelum pemilu legislatif dan dua hari sebelum kampanye.[18]\nMasa kecil dan
keluarga\nLihat pula: Keluarga Joko Widodo\nJoko Widodo lahir dari pasangan Widjiatno Notomihardjo dan Sudjiatmi. Ia merupakan anak sulung dan putra satu-satunya dari
empat bersaudara. Ia memiliki tiga orang adik perempuan bernama Iit Sriyantini, Ida Yati, dan Titik Relawati.[19] Ia sebenarnya memiliki seorang adik laki-laki bernama
Joko Lukito, tetapi meninggal saat persalinan. Sebelum berganti nama, Joko Widodo memiliki nama kecil Mulyono.[20] Ayahnya berasal dari Karanganyar, sementara kakek dan
neneknya berasal dari sebuah desa di Boyolali.[21] Pendidikannya diawali dengan masuk SD Negeri 112 Tirtoyoso yang dikenal sebagai sekolah untuk kalangan menengah ke
bawah.[22]\nDengan kesulitan hidup yang dialami, ia terpaksa berdagang, mengojek payung, dan menjadi kuli panggul untuk membiayai sendiri keperluan sekolah dan uang
jajan sehari-hari. Saat anak-anak lain ke sekolah dengan sepeda, ia memilih untuk tetap berjalan kaki. Mewarisi keahlian bertukang kayu dari ayahnya, ia mulai bekerja
sebagai penggergaji di umur 12 tahun.[9][23] Jokowi kecil telah mengalami penggusuran rumah sebanyak tiga kali. Penggusuran yang dialaminya sebanyak tiga kali pada masa
kecil mempengaruhi cara berpikirnya dan kepemimpinannya kelak setelah menjadi Wali Kota Surakarta saat harus menertibkan permukiman warga.[24]\nSetelah lulus SD, ia
kemudian melanjutkan pendidikan di SMP Negeri 1 Surakarta.[25] Ketika ia lulus SMP, ia sempat ingin masuk ke SMA Negeri 1 Surakarta, namun gagal sehingga pada akhirnya
ia masuk ke SMA Negeri 6 Surakarta.[26]\nJokowi menikah dengan Iriana di Surakarta pada 24 Desember 1986, dan memiliki 3 orang anak, yaitu Gibran Rakabuming Raka
```

Ln 1, Col 1 Spaces: 4 hasil-browse-url2

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

```
from selenium import webdriver
import sys
import argparse
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
import re

options = Options()
options.add_argument("--headless")
driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()), options=options)
filtered_data = []

def parse_args():
    parser = argparse.ArgumentParser()
    parser.add_argument('-i', '--infile', default='', help='input filename')
    parser.add_argument('-o', '--outfile', default='', help='output filename')
    return parser.parse_args()

def clean_text(text):
    text = re.sub(r'http\S+', '', text) # menghapus tautan
    text = re.sub(r'Baca\nLihat sumber\nLihat riwayat\n.+?bahasa Indonesia, ensiklopedia bebas', '', text) # menghapus kalimat yang tidak relevan
    # Hapus kalimat pencarian
    text = re.sub(r'^\s*(?:From|In the news):.*\n?', '', text, flags=re.MULTILINE)
    # Hapus karakter non-ASCII
    text = text.encode('ascii', 'ignore').decode('ascii')
    # Hapus karakter ganda
    text = re.sub(r'\n\s*\n', '\n\n', text)
    # Hapus spasi ekstra
    text = re.sub(r'\s{2,}', ' ', text)
    # Hapus karakter tak dikenal
    text = "".join(c for c in text if ord(c) < 128)
    # Hapus karakter /
    text = text.replace('/', ' ')
    # Pemisahan kalimat
    text = re.sub(r'(?<=[.,])\s+(?=[A-Z])', '  ', text)
    # Hapus karakter \n
    text = text.replace('\n', '')
    return text.strip()

def main():
    args = parse_args()
    outfile = args.outfile
    infile = args.infile

    with open(infile) as f:
        items = f.read().splitlines()
    f.close()

    filtered_data = []

    for i in items:
        driver.get(i)
        text = driver.find_element(By.TAG_NAME, 'body').text
        text = clean_text(text)
        # Filter berdasarkan panjang kata
        words = text.split()
        if len(words) >= 5:
            filtered_data.append(text)

    # Pisahkan hasil scrap antara kedua website
    separator_index = len(filtered_data) // 2
    filtered_data = [filtered_data[:separator_index], filtered_data[separator_index:]]

    filtered_data = [[clean_text(sent) for sent in str(data).split('.') if len(sent) >= 5] for data in filtered_data]
    filtered_data = [[sent.replace('\n', ' ').replace('/', ' ') for sent in data] for data in filtered_data]
    filtered_data = [[sent.strip() for sent in data] for data in filtered_data]
    filtered_data = [list(filter(lambda sent: 'Wikipedia' not in sent and 'Indonesia' not in sent, data)) for data in filtered_data]
    filtered_data = [text for text in filtered_data if "Artikel utama" not in text]

    f = open(outfile, "w")
    f.write(str(filtered_data))
    f.close()

    driver.close()
    f.close()

main()
```

```
usage: ipykernel_launcher.py [-h] [-i INFILE] [-o OUTFILE]
ipykernel_launcher.py: error: unrecognized arguments: -f C:\Users\LENOVO\AppData\Roaming\jupyter\runtime\kernel-6c53ac84-5ebe-4d8f-9944-d4f90cd4e519.json
An exception has occurred, use %tb to see the full traceback.

SystemExit: 2
```

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

```
[20]: !python selen-coba.py -i hasil-search-url.txt -o hasil-coba.txt
```

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