

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]: 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: charset<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

```
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_="yuRufb")
    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 “yuRufb” 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

```

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 mengulangnya 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\n\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 perlindungan dan log perlindungan untuk informasi selengkapnya. Jika Anda tidak dapat menyunting Artikel ini dan Anda ingin melakukannya, Anda dapat memohon permintaan penyuntingan, diskusi dan perubahan yang ingin dilakukan di halaman pembicaraan, memohon untuk melepaskan perlindungan, masuk, atau buatlah sebuah akun.\nIri. 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\x932023)\n\nPendahulu Susilo Bambang Yudhoyono\nGubernur DKI Jakarta ke-14\nMasa jabatan\n15 Oktober 2012 \Xe2\x80\x9316 Oktober 2014\nWakil Gubernur Basuki Tjahaja Purnama\nPendahulu Fauzi Bowo\nFadjar Panjaitan (Plt.)\n\nPengganti Basuki Tjahaja Purnama\nWali Kota Surakarta ke-16\nMasa jabatan\n28 Juli 2005 \Xe2\x80\x931 Oktober 2012\n\nPresiden Susilo Bambang Yudhoyono\nGubernur Mardiyanto\nAli Mufiz\nBibit Waluyo\nWakil F.X. Hadi Rudyatmo\nPendahulu Slamet Suryanto\nAnwar Cholil (Plh.)\n\nPengganti F.X. Hadi Rudyatmo\nKetua Perhimpunan Bangsa-bangsa Asia Tenggara\n\nPetahana\nMulai menjabat\n01 Januari 2023\n\nPendahulu Hun Sen\n\nInformasi pribadi\nLahir Mulyono\n21 Juni 1961 (umur 61)\n\nSurakarta, Jawa Tengah, Indonesia\n\nKebangsaan Indonesia\n\nPartai politik PDI-P\n\nSuami/istri Iriana\nAnak\nGibran Rakabuming Raka\n\nOrang tua\nWidjiatno Notomihardjo (ayah)\nSudjiatmi (ibu)\n\nKerabat\nBobby Nasution (menantu)\nAnwar Usman (ipar)\n\nMata mater Universitas Gadjah Mada\n\nPekerjaan Pengusaha\n\nPolitikus\n\nTanda tangan\n\nSitus web Website Resmi Presiden RI\n\nArtikel ini merupakan bagian dari seri\nJoko Widodo\nSebelum menjadi presiden\nWali Kota Surakarta BST P1Kada Jakarta Gubernur DKI Jakarta LRT MRT\n\nPresiden Indonesia\nIndonesia\n\nPetahana\n\nPilpres 2014 (kampanye) Pilpres 2019 (kampanye) Pelantikan I Pelantikan II Kepresidenan Kabinet Kerja Kabinet Indonesia Maju\n\nKebijakan\n\nBali Nine Tol Laut Kereta cepat Trans-Sumatra Ibu kota baru\n\nKIT APEC 2014 KIT ASEAN 2014 KIT ASEAN 2015 G20 (2014, 2015, 2016, 2017, 2019, 2020, 2021)\n\nKeluarga Iriana Joko Widodo (istri dan Ibu Negara) Anak Gibran Rakabuming Raka Kahiyang Ayu Kaesang Pangarep\n\nSitus Web\n\nSitus Kepresidenan\n\nMedia sosial\n\nFacebook Twitter Instagram YouTube\n\nIri. H. Joko Widodo (pengucapan bahasa Indonesia: [dʰɔcaˈxɔ2ˈxc9ˈx94kˈxc9ˈx94 widʰɔcaˈx94dʰɔcaˈx94]; 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]\n\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]\n\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]\n\nMasa kecil dan keluarga\n\nLihat pula: Keluarga Joko Widodo\n\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 Tirtoyo yang dikenal sebagai sekolah untuk kalangan menengah ke bawah.[22]\n\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 pengusuran rumah sebanyak tiga kali. Pengusuran 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]\n\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]\n\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"([a-zA-Z0-9]+) ([a-zA-Z0-9]+) ([a-zA-Z0-9]+)", " ", text) # menghapus link yang tidak relevan
    # Hapus karakter non-ASCII
    text = re.sub(r"[^\x00-\x7F]+", "", text, flags=re.MULTILINE)
    # Hapus karakter non-ASCII
    text = text.encode("ascii", "ignore").decode("ascii")
    # Hapus karakter ganda
    text = re.sub(r"\\n\\n", "\\n", text)
    # Hapus spasi ekstra
    text = re.sub(r"\\s+", " ", text)
    # Hapus karakter tab di awal
    text = re.sub(r"\\t", "", text)
    # Hapus karakter /
    text = text.replace("/", "")
    # Hapus karakter \n
    text = re.sub(r"\\n", "", 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 script antara header dan body
    separator_index = len(filtered_data) // 2
    filtered_data = [filtered_data[separator_index], filtered_data[separator_index:]]

    filtered_data = [clean_text(s) for s in str(data).split("\n") if len(s) >= 5] for data in filtered_data]
    filtered_data = [s.replace("\n", " ") for s in data] for data in filtered_data]
    filtered_data = [s.strip() for s in data] for data in filtered_data]
    filtered_data = [s for s in data if s.strip() != "" and s.strip() != " "] for data in filtered_data]
    filtered_data = [s for s in data if s.strip() != "" and s.strip() != " "] for data in filtered_data]

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

    driver.close()
    f.close()

main()

usage: ipynb_launcher.py [-h] [-i INFILE] [-o OUTFILE]
ipynb_launcher.py: error: unrecognized arguments: -f C:\Users\LENDVO\AppData\Roaming\jupyter/runtime\kernel-6c5384-5db-428f-9944-04f9b4d4519.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

[[('Lihat syarat dan ketentuannya', 'Lihat kebijakan perlindungan dan log perlindungan untuk informasi selengkapnnya', '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 perlindungan, masuk, atau buatlah sebuah akun', '[1]Pengganti Basuki Tjahaja Purnamawall Kota Surakarta ke-16Masa Jabatan28 Juli 2005 1 Oktober 2012Presiden Susilo Bambang YudhoyonoGubernur MardiyantoAll MufizBibit Waluyowakil F', 'Hadri RudyatmoPendahuulu Slamet SuryantonoAnwar Cholli (Plh', '[2]Pengganti F', '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', 'Hadri Rudyatmo sebagai wakil wali kota', '[8]Joko 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 memenangkan Pilkada Jakarta 2012', 'Kemenangannya dianggap mencerminkan dukungan populer untuk seorang pemimpin yang "muda" dan "bersih", meskipun umurnya sudah lebih dari 50 tahun', '[12]Semerjak 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]Masa kecil dan keluarganyaItih pula: Keluarga Joko WidodoJoko Widodo lahir dari pasangan Widjaitno Notomihardjo dan Sudjaitni', 'Ia merupakan anak sulung dan putra satu-satunya dari empat bersaudara', 'Ia memiliki tiga orang adik perempuan bernama Iit Sriyanti, 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]Dengan kesulitan hidup yang dialami, ia terpaksa berdagang, mengojek payung, dan menjadi kuli pangkut 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 pengurusan rumah sebanyak tiga kali', 'Pengurusan 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]Setelah 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]Jokowi menikah dengan Irlana di Surakarta pada 24 Desember 1906, dan memiliki 3 orang anak, yaitu Gibran Rakabuming Raka (1987), Kahiyang Ayu (1991), dan Kaesang Pangarep (1994)', 'Masa kuliah dan beryisruahDengan kemampuan akademis yang dimiliki, ia diterima di Jurusan Kehutanan, Fakultas Kehutanan Universitas Gajah Mada', 'Kesempatan ini dimanfaatkan untuk belajar struktur kayu, pemanfaatan, dan teknologinya', 'Ia berhasil menyelesaikan pendidikannya dengan judul skripsi "Studi tentang Pola Konsumsi Kayu Lapis pada Pemakaian Akhir di Kodya Surakarta" dan dengan gelar Insinyur (Ir', 'Selain kuliah, ia juga tercatat aktif sebagai anggota Mapala Silvagama, unit kegiatan mahasiswa pecinta alam di fakultasnya', 'Setelah lulus pada 1985, ia bekerja di BUN PT Kertas Kraft Aceh, dan ditempatkan di area Tumbuhan Pinus Merkusii di Dataran Tinggi Gayo, Aceh Tengah', 'Namun ia merasa tidak betah dan pulang menyusul istrinya yang sedang hamil tujuh bulan', 'Ia bertekad berbisnis di bidang kayu dan bekerja di usaha miliknya, Miyono, di bawah bendera CV Roda Jati', 'Pada tahun 1988, ia memberanikan diri membuka usaha sendiri dengan nama CV Rakabu, yang diambil dari nama anak pertamanya', 'Usahnya sempat berjaya dan juga naik turun karena tertipu pesanan yang akhirnya tidak dibayar', 'Namun pada tahun 1990 ia bangkit kembali dengan pinjaman modal Rp30 juta yang ia peroleh dari ibunya', '[27]Usaha ini membawanya bertemu Bernard Chene,[28][29] seorang pria berkebangsaan Prancis, yang akhirnya memberinya panggilan yang populer hingga kini, "Jokowi", 'Dengan kejujuran dan kerja kerasnya, ia mendapat kepercayaan dan bisa berkeilling Eropa yang membuka mata', 'Pengaturan kota yang baik di Eropa menjadi inspirasinya untuk diterapkan di Solo dan menginspirasi untuk memasuki dunia politik', 'Ia ingin menerapkan kepemimpinan manusiawi dan mewujudkan kota yang bersahabat untuk penghuninya yaitu daerah Surakarta', 'Ia berhasil memenangkan pemilihan tersebut dengan persentase suara sebesar 36,62%', '[10] Setelah terpilih, dengan berbagai pengalaman pada masa muda, ia mengembangkan Solo yang sebelumnya buruk penataannya dan menghadapi berbagai penolakan masyarakat untuk ditingkatkan', 'Di bawah kepemimpinannya, Solo mengalami perubahan dan menjadi kajian di universitas dalam dan luar negeri', '[30]Di bawah kepemimpinannya, bus Batik Solo Trans diperkenalkan,[31] berbagai kawasan seperti Jalan Slamet Riyadi dan Ngarsopuro diremajakan,[32] dan Solo menjadi tuan rumah berbagai acara internasional', '[32] Selain itu, Jokowi juga dikenal akan pendekatannya dalam merelokasi pedagang kaki lima yang "memanusiakan manusia", '[33] Berkat pencapaiannya ini, pada tahun 2010, ia terpilih lagi sebagai Wali Kota Surakarta dengan suara melebihi 90%', '[34] Kemudian, pada tahun 2012, ia dicalonkan oleh PDI-P sebagai calon Gubernur DKI Jakarta', '[8]Gubernur DKI JakartaArtikel utama: Karier Joko Widodo sebagai Gubernur DKI JakartaFoto resmi Jokowi sebagai Gubernur DKI Jakarta, 2012', 'Pilkada 2012Lihat pula: Pemilihan umum Gubernur DKI Jakarta 2012Susana di posko pemenang Jokowi di Jalan Borobudur 22', 'Jokowi diminta secara pribadi oleh Jusuf Kalla untuk mencalonkan diri sebagai Gubernur DKI Jakarta pada Pilgub DKI tahun 2012', '[35] Karena merupakan kader PDI Perjuangan, maka Jusuf Kalla meminta dukungan dari Megawati Soekarnoputri, yang awalnya terlihat masih ragu', 'Sementara itu, Prabowo Subianto dari Partai Gerindra juga melobi PDI Perjuangan agar bersedia mendukung Jokowi sebagai calon gubernur karena membutuhkan 9 kursi lagi untuk bisa mengkalikan calon gubernur', '[36] Sebagai wakilnya, Basuki Tjahaja Purnama yang saat itu menjadi anggota DPR dicalonkan mendampingi Jokowi', '[37] Pasangan ini awalnya tidak diunggulkan', '[38] Namun hasil pilgub putaran pertama dari KPU memperlihatkan Jokowi memimpin dengan 42,6% suara, sementara Fauzi Bowo di posisi kedua dengan 34,05% suara', '[39]Pasangan ini berbalik diunggulkan memenangkan pilnikuda DKI 2012 karena kedekatan Jokowi dengan Hidayat Nur Mahid saat pilkada Wali Kota Surakarta 2010[40] serta pendukung Faizal Basri dan Alex Noerdin dari hasil survei cenderung beralih kepadanya', '[41] Namun keadaan berbalik setelah partai-partai pendukung calon lainnya di putaran pertama malah menyalakan dukungan kepada Fauzi Bowo', '[42][43] Jokowi akhirnya mendapat dukungan dari tokoh-tokoh penting seperti Mibakhun dari PKS,[44] Jusuf Kalla dari Partai Golkar,[45] Indra J', 'Piliani dari Partai Golkar, [46] serta Roro Heri yang merupakan adik ipar Fauzi Bowo', '[47]Pertarungan politik juga merambah ke media sosial dengan peluncuran Jasew,[48] pembentukan media center,[49] serta pemanfaatan media baru seperti Youtube', '[50] Putaran kedua juga diwarnai tuduhan kampanye hitam yang antara lain berkisar dalam isu SARA,[51] isu kebakaran yang disengaja,[52] korupsi,[53] dan politik transaksional', '[54] Pada 29 September 2012, KPU DKI Jakarta menetapkan pasangan Jokowi - Ahok sebagai gubernur dan wakil gubernur DKI yang baru untuk masa bakti 2012[2017 menggantikan Fauzi Bowo - Prijanto', '[55][56]KebijakanKebijakan Joko Widodo selama menjabat Gubernur DKI Jakarta banyak yang disebut populis, seperti Kampung Deret, Kartu Jakarta Sehat dan Kartu Jakarta Pintar', 'Namun

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

```
[1]: import requests
from bs4 import BeautifulSoup

# melakukan request ke URL Wikipedia
url = "https://id.wikipedia.org/wiki/Joko_Widodo"
response = requests.get(url)

# parsing teks HTML menggunakan BeautifulSoup
soup = BeautifulSoup(response.content, "html.parser")

# mengambil teks dari tag <p> dalam HTML
p_tags = soup.find_all("p")

# membuat list kosong untuk menyimpan hasil rapikan
clean_text = []

# menghapus karakter yang tidak diperlukan dan memech teks menjadi kalimat-kalimat
for tag in p_tags:
    text = tag.get_text().replace("\n", "")
    sentences = text.split(".")
    for sentence in sentences:
        sentence = sentence.strip()
        if len(sentence.split()) >= 5: # hanya menyimpan kalimat dengan item kata >= 5
            clean_text.append(sentence)

# menghapus kalimat yang tidak relevan
clean_text = [text for text in clean_text if "Artikel utama" not in text]

# menggabungkan kembali kalimat-kalimat yang relevan menjadi satu teks
result = " ".join(clean_text)

# mencetak hasil rapikan
print(result)

Joko Widodo (pengucapan bahasa Indonesia: [dʒɔkɔ wɪdɔdɔ]; lahir 21 Juni 1961) adalah presiden Indonesia yang mulai menjabat sejak tanggal 20 Oktober 2014 terpilih dalam Pemilu Presiden 2014. Jokowi menjadi presiden pertama 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 s ebagai wakil gubernur Sebelumnya, ia adalah wali kota Surakarta, sejak tanggal 28 Juli 2005 hingga 1 Oktober 2012, didampingi F Hadri Rudyatmo sebagai wakil wali kota [7] Dua tahun menjelang 1 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 Joko 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 [1 0] 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 be rhasil memenangi Pilkada Jakarta 2012 Kemenangannya dianggap mencerminkan dukungan populer untuk seorang pemimpin yang "muda" dan "bersih", meskipun umurnya sudah lebih dari 50 tahun Semenjak terpilih sebagai gubernur, popularitasnya terus naik dan menjadi sorotan media [13][14] Akibatnya, muncul wacana untuk menjadikannya calon presiden untuk Pemilu Presiden 2014 [15] ita mba 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 d
```