LAPORAN

RENCANA TUGAS MANDIRI (RTM) Ke-2

MATA KULIAH BIG DATA (A)

“WEB SCRAPING QUERY : JOKO WIDODO”

DISUSUN OLEH:

Reza Putri Angga (22083010006)

DOSEN PENGAMPU:

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

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR

2024

STUDI KASUS DAN PEMBAHASAN

Web Scraping merupakan sebuah proses ekstraksi sejumlah besar data secara otomatis dari beberapa halaman *website* yang di ingkan. Pada penugasan ini, dilakukan proses *web scraping* dengan kata kunci "Joko Widodo". Dimulai dengan proses pencarian *website* yang sesuai dengan kata kunci yang diberikan, diambil dua *url* teratas dari proses pencarian tersebut, dilakukan proses ekstraksi berita berdasarkan paragraf, dan dilakukan pembersihan (*cleaning*) dengan tahapan pembersihan *slice n*, filterisasi penghapusan paragraf kurang dari 5 kata, dan pengubahan output menjadi dua dimensi.

Terdapat dua langkah atau dua cara untuk menampilkan *output* hasil pembersihan (*cleaning*). Yang pertama, diproses menggunakan *python file* kemudian disimpan dalam file *txt* dan yang kedua diproses dalam *notebook*. Mengenai langkah dan *output* secara detail, akan dijelaskan lebih lanjut sebagai berikut.

1. Proses Pengumpulan URL Berita Dengan Kata Kunci "Joko Widodo"

Pada tahapan ini, dilakukan proses pencarian dan pengumpulan judul dan *url* berita dengan kata kunci yang diberikan menggunakan *python file*.

1.1 Proses Perolehan Judul Dan *URL* Berita

```python
#penggunaan library
from bs4 import BeautifulSoup #parsing page web
from selenium import webdriver #otomatisasi brower
from selenium.webdriver.chrome.service import Service as ChromeService #penggunaan chrome
from webdriver_manager.chrome import ChromeDriverManager

#inisiasi chrome
chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument("--headless")

#inisiasi driver
driver = webdriver.Chrome(service = ChromeService(ChromeDriverManager().install()))

#kata kunci yang ingin dicari
query = "Joko Widodo"

#jumlah page yang ingin dicari
n_pages = 10

#looping
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('href'))
        print(h.a.text)
        print(h.a.get('href'))
```

kode script collecting judul dan url berita

Pada kode *script* di atas, dilakukan proses *collecting* judul dan *url* berita dengan kata kunci "Joko Widodo" yang disimpan dalam variabel *query*. Dipergunakan beberapa *library* untuk mempermudah, yakni *library* BeautifulSoup untuk parsing

HTML, selenium untuk otomatisasi dan mengatur layanan *google chrome*, serta ChromeDriverManager untuk mengelola *driver chrome*. Kemudian, dilakukan proses pencarian sebanyak sepuluh halaman dengan inisiasi elemen *div class* "yuRUbf" sebagai elemen informasi judul dan *url* (tautan) berita dari pencarian dengan kata kunci tersebut.

```
1.1. PROSES PEROLEHAN JUDUL DAN URL BERITA

#run kode program "selenium-search-url.py" untuk mendapatkan teks dan url dari hasil pencarian dengan keyword "Joko Widodo
#hasilnya akan disimpan kedalam "hasil-search-text-url.txt" yang berisi judul dan url (link) berita

!python selenium-search-url.py > hasil-search-text-url.txt
```

menjalankan kode script

Dilakukan proses menjalankan (*run*) kode *script* dengan menggunakan *notebook*, hal ini bertujuan agar proses menjalankan kode dapat terdokumentasikan. Dengan menjalankan kode *script* python pada "selenium-search-url.py" dan akan menyimpan *output* berupa judul dan *url* berita pada "hasil-search-text-url.txt". Dengan hasil seperti di bawah ini.

```
Presiden Joko WidodoPresiden RIhttps://www.presidenri.go.id › presiden-joko-widodo
https://www.presidenri.go.id/presiden-joko-widodo/
Joko Widodo - Wikipedia bahasa Indonesia, ensiklopedia ...Wikipediahttps://id.wikipedia.org › wiki › Joko_Widodo
https://id.wikipedia.org/wiki/Joko_Widodo
Joko Widodo (@jokowi) • Instagram photos and videosInstagramhttps://www.instagram.com › jokowi
https://www.instagram.com/jokowi/
Presiden Joko WidodoYouTubehttps://www.youtube.com › channel
https://www.youtube.com/channel/UCPeG-JX2dB90P3RgZbVNheg
Laman Resmi Presiden Republik Indonesia • Presiden RIPresiden RIhttps://www.presidenri.go.id
https://www.presidenri.go.id/
Presiden Joko WidodoFacebookhttps://www.facebook.com › ... › Presiden Joko Widodo
https://www.facebook.com/Jokowi/?locale=id_ID
Presiden Joko WidodoYouTubehttps://www.youtube.com › @Jokowi
https://www.youtube.com/@Jokowi
Berita dan Informasi Joko widodo Terkini dan Terbaru Hari inidetikcomhttps://www.detik.com › tag › joko-widodo
```

output judul dan url berita

1.2 Proses Perolehan *URL* Berita (Saja)

```
search = soup.find_all('div', class_ = "yuRUbf")
for h in search:
    #links.append(h.a.get('href'))

    #kemudian, untuk hanya menampilkan url-nya saja, dilakukan penambahan "#" pada cetak teksnya
    #print(h.a.text)

    print(h.a.get('href'))
```

kode script collecting url berita

Pada kode *script* di atas, dilakukan proses *collecting url* berita dengan menggunakan kode *script* dan kata kunci yang sama seperti pada proses perolehan judul dan *url* berita dengan menggunakan *library* dan kata kunci yang sama. Namun, terdapat perbedaan dengan menambahkan "#" pada cetak judul berita, sehingga *output* yang ditampilkan hanya *url* beritanya saja.

```
#run kode program "selenium-search-url.py" untuk mendapatkan teks dan url dari hasil pencarian dengan keyword "Joko Widodo
#hasilnya akan disimpan kedalam "hasil-search-url.txt" yang berisi link berita url (link) berita saja

!python selenium-search-url.py > hasil-search-url.txt
```

menjalankan kode script

Dilakukan proses menjalankan (*run*) kode *script script* python pada "selenium-search-url.py" dan akan menyimpan *output* berupa *url* berita pada "hasil-search- url.txt". Dengan hasil seperti di bawah ini.

```
https://www.presidenri.go.id/presiden-joko-widodo/
https://id.wikipedia.org/wiki/Joko_Widodo
https://www.instagram.com/jokowi/
https://www.youtube.com/channel/UCPeG-JX2dB90P3RgZbVNheg
https://www.presidenri.go.id/
https://www.facebook.com/Jokowi/?locale=id_ID
https://www.youtube.com/@Jokowi
https://www.detik.com/tag/joko-widodo
https://www.tempo.co/tag/jokowi
https://twitter.com/jokowi
https://setkab.go.id/tag/jokowi/
https://www.cnnindonesia.com/tag/jokowi
https://www.detik.com/tag/jokowi
https://www.setneg.go.id/listcontent/listberita/berita_presiden_dan_pemerintah
https://ppid.lampungprov.go.id/detail-post/Selamat-Ulang-Tahun-Presiden-Joko-Widodo
```

output url berita

2. Proses Ekstraksi Berita Dengan Menggunakan Dua *URL* Teratas

Pada tahapan ini dilakukan ekstraksi berita dengan menggunakan dua *url* teratas dari file "hasil-search-url.txt", meliputi https://www.presidenri.go.id/presiden-joko-widodo/ sebagai *url* 1 dan https://id.wikipedia.org/wiki/Joko_Widodo sebagai *url* 2 menggunakan *python file.*

```
#penggunaan library
from selenium import webdriver
import sys, getopt
import argparse
from selenium.webdriver.chrome.service import Service as ChromeService
from selenium.webdriver.common.by import By
from webdriver_manager.chrome import ChromeDriverManager

#menghubungkan ke chrome
chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(service = ChromeService(ChromeDriverManager().install()))

#fungsi parse_args untuk menjalankan skrip dengan argument input otput file
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()


#fungsi main untuk skrip utama
def main():

    args = parse_args()
    outfile = args.outfile
    infile = args.infile

    data = []

    with open(infile, "r", encoding = "utf-8") as f:
        urls = f.read().splitlines()

    #mengambil dua url-teratas
    top_two_url = urls[:2]

    for u in top_two_url:
        driver.get(u)
        elems = driver.find_element(By.TAG_NAME, "body").text
        data.append(elems)

    with open(outfile, "w", encoding = "utf-8") as f:
        f.write(str(data))

    driver.close()

#menjalankan kode
if __name__ == "__main__":
    main()
```

kode script ekstraksi berita

Pada kode *script* di atas, dilakukan proses ektraksi atau perolehan teks berita dengan mengguankan *library* yang digunakan di kode *script* sebelumnya. Terdapat fungsi utama *main ( )* yang dipergunakan sebagai skrip utama dengan *argument* penenutan file *input* dan *output* yang diinginkan. Dilakukan proses pengambilan semua teks atau elemen didalam berita dengan menggunakan *argument "body"* dari dua *url* teratas file inputan dan hasilnya akan disimpan dalam *list* data. Kemudian, hasil dari *list* data tersebut disimpan ke dalam sebuah file *output*.

```
#dengan hasil url dari scraping yang berada dalam file "hasil-search-url.txt" akan dijalankan skrip "selenium-browse-url.py"
#dan hasilnya akan disimpan dalam hasil-browse-url.txt"
#dipilih dua url teratas yang akan discraping dan dibersihkan lebih lanjut, yakni :

!python selenium-browse-url.py -i hasil-search-url.txt -o hasil-scraping-text.txt
```

menjalankan kode script

Dilakukan proses menjalankan (*run*) kode *script script* python pada "selenium-browse-url.py" dengan file inputan "hasil-search-url.txt" untuk mengambil dua *url* teratas, dan file *outputnya* akan disimpan dalam "hasil-scraping-text.txt". Dengan hasil seperti di bawah ini.

["PRESIDEN RI\nKEPRESIDENAN\nPERS\nPEMERINTAHAN\nPresiden Joko Widodo\nIr. H. Joko Widodo\nUnduh Foto Resmi\nIr. H. Joko Widodo adalah Presiden ke-7 Republik Indonesia yang mulai menjabat sejak 20 Oktober 2014. Lahir di Surakarta, Jawa Tengah, pada 21 Juni 1961, Joko Widodo pertama kali terjun ke pemerintahan sebagai Wali Kota Surakarta (Solo) pada 28 Juli 2005 hingga 1 Oktober 2012.\nSelepas itu, Joko Widodo menjabat sebagai Gubernur DKI Jakarta pada 15 Oktober 2012 sebelum terpilih sebagai Presiden Republik Indonesia pada Pemilihan Presiden (Pilpres) 2014. Saat Pilpres tersebut Joko Widodo terpilih bersama pasangannya, Jusuf Kalla.\nDalam Pilpres 2019, Joko Widodo kembali terpilih sebagai Presiden Republik Indonesia untuk masa jabatannya yang kedua. Kali ini, Joko Widodo didampingi oleh Wakil Presiden K.H. Ma'ruf Amin dan dilantik pada 20 Oktober 2019 untuk masa jabatan 2019 hingga 2024 mendatang.\nPembangunan infrastruktur menjadi program prioritas di masa kepemimpinannya yang pertama. Pembangunan yang dilakukan secara merata hingga ke daerah terluar Indonesia ini dilakukan untuk mengejar ketertinggalan Indonesia dalam sektor ini dibandingkan negara-negara lain.\nProgram prioritas tersebut dibarengi dengan program berupa bantuan sosial seperti Kartu Indonesia Pintar (KIP), Kartu Indonesia Sehat (KIS), hingga Program Keluarga Harapan (PKH). Selain itu, sejak awal masa jabatannya, Joko Widodo juga mengupayakan reforma agraria dengan salah satunya melakukan percepatan penerbitan sertifikat hak atas tanah untuk mengurangi terjadinya sengketa lahan oleh karena ketiadaan sertifikat.\nDi masa jabatannya yang kedua, Joko Widodo mengalihkan fokus pemerintahan pada pembangunan dan peningkatan kapasitas sumber daya manusia Indonesia untuk dapat bersaing dengan negara-negara lainnya. Adapun program pembangunan infrastruktur masih terus dilanjutkan bersamaan dengan itu.\nDikelola oleh:\nBIRO PERS, MEDIA, DAN INFORMASI\nSEKRETARIAT PRESIDEN\nJl. Veteran No. 16\nJakarta Pusat, Kode Pos 10110\nPEMERINTAHAN\nKabinet Indonesia Maju\nCapaian Kinerja\nPERS\nSiaran Pers\nFoto Presiden\nVideo Presiden\nTranskrip\nKEPRESIDENAN\nPresiden Joko Widodo\nWakil Presiden Ma'ruf Amin\n© 2019–2024 Biro Pers, Media dan Informasi Sekretariat Presiden - Kementerian Sekretariat Negara", 'Lompat ke isi\nMenu utama\nCari\nBuat akun baru\nMasuk log\nPerkakas pribadi\nPotretlah budaya lokal Anda, bantu Wikipedia, dan jadilah juara!\nIkuti Wikipedia bahasa Indonesia di Facebook, X, Instagram, dan Telegram\nDaftar isi sembunyikan\nAwal\nKehidupan awal\nGulingkan subbagian Kehidupan awal\nKarier politik\nGulingkan subbagian Karier politik\nPemilihan umum presiden\nGulingkan subbagian Pemilihan umum presiden\nKepresidenan\nGulingkan subbagian Kepresidenan\nAgama\nIdeologi\nPenghargaan\nGulingkan subbagian Penghargaan\nPolitik\nGulingkan subbagian Politik\nKontroversi\nGulingkan subbagian Kontroversi\nDalam budaya populer\nGulingkan subbagian Dalam budaya populer\nLihat pula\nReferensi\nPranala luar\nJoko Widodo\n180 bahasa\nHalaman\nPembicaraan\nBaca\nLihat sumber\nLihat riwayat\nPerkakas\nDari Wikipedia bahasa Indonesia, ensiklopedia bebas\nJoko Widodo\nPotret resmi, 2019\nPresiden Indonesia ke-7\nPetahana\nMulai menjabat\n20 Oktober 2014\nWakil Presiden\nJusuf Kalla (2014–2019)\nMa'ruf Amin (2019–2024)\nPendahulu\nSusilo Bambang Yudhoyono\nPenggantiPetahana\nGubernur Jakarta ke-14\nMasa jabatan\n15 Oktober 2012 – 16 Oktober 2014\nWakil Gubernur Basuki Tjahaja Purnama\nPendahulu\nFauzi Bowo\nFadjar Panjaitan (Plt.)[1]\nPenggantiBasuki Tjahaja Purnama\nWali Kota Surakarta ke-16\nMasa jabatan\n28 Juli 2005 – 1 Oktober 2012\nPresiden Susilo Bambang Yudhoyono\nGubernurMardiyanto\nAli Mufiz\nBibit Nalupe\nWakil F.X. Hadi Rudyatmo\nPendahulu\nSlamet Suryanto\nAnwar Cholil (Plh.)[2]\nPenggantiF.X. Hadi Rudyatmo\nInformasi pribadi\nLahir Mulyono\n21 Juni 1961 (umur 62)\nSurakarta, Jawa Tengah, Indonesia[3]\nPartai politik PDI-P\nSuami/istri Iriana (m. 1986)\nAnak\nGibran Rakabuming Raka[4]\nKahiyang Ayu[5]\nKaesang Pangarep[5]\nOrang tua\nWidjiatno Notomihardjo (ayah)\nSudjiatmi Notomihardjo[6] (ibu)\nKerabat\nBobby Nasution (menantu)\nAnwar Usman (ipar)\nAlma mater Universitas Gadjah Mada\nPekerjaan\nPengusahapolitikus\nTanda tangan\nSitus web Situs web resmi Presiden RI\nSunting kotak info • L • B\nArtikel ini merupakan bagian dari seri\nJoko Widodo\nSebelum menjadi presiden\nWali Kota Surakarta DST 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

output ekstraksi berita

informasi format ekstraksi berita

Dapat diperoleh informasi bahwa terdapat beberapa format data yang tidak rapi, seperti adanya "\n" yang merupakan karakter *newline* sebanyak 800. Hal ini dapat mengganggu untuk memperoleh hasil yang informatif, oleh karena itu dilakukan proses pembersihan (*cleaning*).

3. Proses Pembersihan (*Cleaning*) Pada Hasil Ekstraksi Berita

Pada tahapan ini, dilakukan proses pembersihan *(cleaning*) dari hasil berita yang telah diektraksi. Untuk melakukan proses ini, dipergunakan dua cara yakni, menggunakan *python file* dengan hasil yang akan disimpan dalam file txt dan *notebook* untuk mengetahui mengenai akses hasil *outputnya*.

3.1 *Python File*

3.1.1 Proses Pembersihan *(Cleaning) Slice N*

```python
#melakukan cleaning slice n
#penggunaan library
import re
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.service import Service as ChromeService

#menghubungkan dengan chrome
chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(service = ChromeService(ChromeDriverManager().install()))


#define function utama yang akan dijalankan
def main():

    data = []  #menyimpan kedalam list

    with open("hasil-search-url.txt", "r", encoding = "utf-8") as f:
        urls = [line.rstrip("\n") for line in f][:2]  #melakukan penghapusan \n pada dua url teratas

    #looping for
    for u in urls:
        driver.get(u)
        elems = driver.find_elements("tag name", "p")  #melakukan ekstraksi paragraf

        #penambahan hasil ke dalam list data []
        for elem in elems:
            cleaned_text = re.sub(r'[\\^\n]', " ", elem.text)  #membersihkan teks lebih lanjut
            cleaned_text = re.sub(r'(\[\d+\]|(\d+\s*([.!?])|(\d+\s*(?=\n|$)))', "", cleaned_text) #penghapusan pola angka []
            data.append(cleaned_text)

    driver.close()

    #hasilnya akan disimpan dalam "cleaning-filterisasi.txt"
    with open("slicen-cleaning.txt", "w", encoding = "utf-8") as f:
        f.write(str(data))

if __name__ == "__main__":
    main()
```

kode script pembersihan slice n

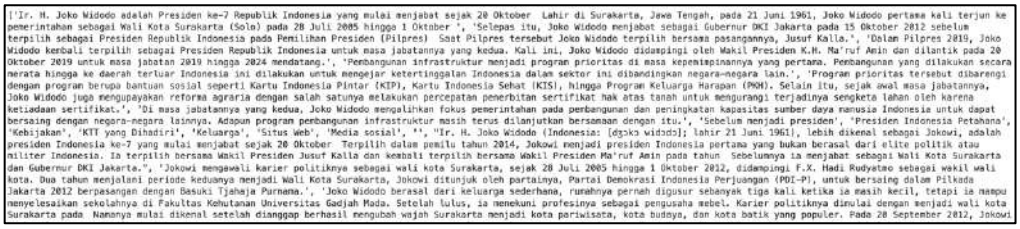
Pada kode *script* di atas, dilakukan proses pembersihan (*cleaning*) *slice n* menggunakan *library* tambahan regex untuk memanipulasi *string*. Terdapat fungsi utama *main ( )* yang dipergunakan sebagai skrip utama dengan adanya *list* data untuk menyimpan hasilnya. Dengan menggunakan dua *url* teratas dari file "hasil-search-url.txt" dilakukan proses pembacaan setiap baris dengan *list comprehension* dan menghilangkan karakter "\n" menggunakan *rstrip* agar tidak memiliki karakter *newline* yang tidak di inginkan.

Kemudian, dilakukan proses perulangan atau *looping for* untuk ekstraksi dan pembersihan teks berita berdasarkan paragraf dengan tahapan menghapus "\n" lebih lanjut menggunakan *argument re.sub ( )* dan penghapusan pola angka dalam kurung siku, seperti [9]. *Outputnya* akan disimpan dalam file "slicen-cleaning.txt". Jadi, pada proses pembersihan ini dimulai dengan pembersihan baris terlebih dahulu, kemudian dilakukan proses pembersihan paragraf. Nantinya, hasil akhir yang akan ditampilkan akan dibagi berdasarkan paragrafnya.

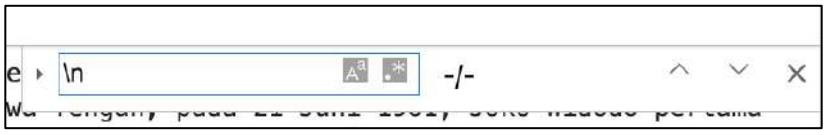
```
#dengan hasil url dari scraping yang berada dalam file "hasil-search-url.txt" akan dijalankan skrip "slicen-cleaning.py"
#untuk melakukan pembersihan \n
#dan hasilnya akan disimpan dalam "slicen-cleaning.txt"

!python slicen-cleaning.py
```

menjalankan kode script

Dilakukan proses menjalankan (*run*) kode *script script* python pada "slicen-cleaning.py" dengan file *input* dan *output* sesuai yang telah dituliskan dalam kode *script*. Dengan hasil seperti di bawah ini.



output pembersihan slice n



informasi hasil pembersihan slice n

Dapat diperlihatkan bahwa hasil pembersihan *slice n* di simpan dalam satu *list* data dan diperoleh informasi bahwa tidak ada lagi *slice n* pada hasil tersebut.

3.1.2 Proses Filterisasi Penghapusan Baris Paragraf Dengan Kata Kurang Dari 5

```python
#melakukan cleaning slice n, filterisasi penghapusan jumlah baris paragraf < 5 kata
#penggunaan library
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.service import Service as ChromeService
import re

#menghubungkan dengan chrome
chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(service = ChromeService(ChromeDriverManager().install()))

#define function utama yang akan dijalankan
def main():

    data = []  #menyimpan kedalam list

    with open("hasil-search-url.txt", "r", encoding = "utf-8") as f:
        urls = [line.rstrip("\n") for line in f][:2]  #melakukan penghapusan \n pada dua url teratas

    #looping for
    for u in urls:
        driver.get(u)
        elems = driver.find_elements("tag name", "p")  #melakukan ekstraksi paragraf

        #penambahan hasil ke dalam list data []
        for elem in elems:
            cleaned_text = re.sub(r'[\[\^\n]', "", elem.text) #membersihkan teks lebih lanjut
            cleaned_text = re.sub(r'(\[\d+\])|(\d+\s*([.!?])|(\d+\s*(?=\n|$)))', "", cleaned_text) #penghapusan pola angka []
            if len(cleaned_text.split()) >= 5: #filterisasi pengahapusan jumlah baris paragraf < 5 kata
                data.append(cleaned_text)

    driver.close()

    #hasilnya akan disimpan dalam "cleaning-filterisasi.txt"
    with open("filterisasi-cleaning.txt", "w", encoding = "utf-8") as f:
        f.write(str(data))

if __name__ == "__main__":
    main()
```

kode script pembersihan slice n, filterisasi penghapusan baris paragraf < 5 kata

Pada kode *script* di atas, dilakukan proses penambahan kode untuk melakukan filterisasi penghapusan baris paragraf yang mengandung kata kurang dari 5 dengan menggunakan library yang sama seperti kode sebelumnya. Terdapat

fungsi utama *main ( )* yang dipergunakan sebagai skrip utama dengan adanya *list* data untuk menyimpan hasilnya. Perlu diketahui bahwa setiap baris dalam pemrograman merupakan karakter *new line*, oleh karena itu dengan menggunakan teks dari setiap elemen paragraf “p” dari dua *url* teratas pada file “hasil-search-url.txt”.

Dilakukan penghapusan *slice n* dengan algoritma yang sama pada kode sebelumnya. Namun ditambahkan dengan *argument* penghapusan setiap baris paragraf dengan menggunakan “p" yang mengandung kata kurang dari 5 dengan menggunakan *if len(cleaned_text.split()) >= 5*, maka hasilnya akan ditambahkan ke dalam *list* data. *Ouputnya* akan disimpan dalam *file* “filterisasi-cleaning.txt”.

```
#dengan hasil url dari scraping yang berada dalam file "hasil-search-url.txt" akan dijalankan skrip "filterisasi-cleaning.py"
#untuk melakukan pembersihan \n dan fiterisasi penghapusan baris paragraf < 5 kata
#dan hasilnya akan disimpan dalam "filterisasi-cleaning.txt"

!python filterisasi-cleaning.py
```

menjalankan kode script

Dilakukan proses menjalankan (*run*) kode *script script* python pada “filterisasi-cleaning.py” dengan file inputan dan output sesuai yang telah dituliskan dalam kode *script*. Dengan hasil perubahan seperti di bawah ini.

```
sertifikat hak atas tanah untuk mengurangi terjadinya sengketa lahan oleh karena ketiadaan sertifikat.', 'Di masa jabatannya yang kedua, Joko Widodo
mengalihkan fokus pemerintahan pada pembangunan dan peningkatan kapasitas sumber daya manusia Indonesia untuk dapat bersaing dengan negara-negara lainnya.
Adapun program pembangunan infrastruktur masih terus dilanjutkan bersamaan dengan itu.', 'Sebelum menjadi presiden', 'Presiden Indonesia Petahana',
'Kebijakan', 'KTT yang Dihadiri', 'Keluarga', 'Situs Web', 'Media sosial', '', "Ir. H. Joko Widodo (Indonesia: [dʒɔkɔ widɔdɔ]; lahir 21 Juni 1961), lebih
dikenal sebagai Jokowi, adalah presiden Indonesia ke-7 yang mulai menjabat sejak 20 Oktober  Terpilih dalam pemilu tahun 2014, Jokowi menjadi presiden
```

potongan script output sebelum dilakukan filterisasi penghapusan baris paragraf < 5 kata

Dapat diperlihatkan, bahwa sebelum melakukan filterisasi terdapat baris paragraph seperti ‘Sebelum menjadi presiden’, ‘Presiden Indonesia Pertahan’, ‘Kebijakan’, ‘KTT yang Dihadiri’, ‘Keluarga’, ‘Situs Web’, dan ‘Media Sosial’.

```
sertifikat hak atas tanah untuk mengurangi terjadinya sengketa lahan oleh karena ketiadaan sertifikat.', 'Di masa jabatannya yang kedua, Joko Widodo
mengalihkan fokus pemerintahan pada pembangunan dan peningkatan kapasitas sumber daya manusia Indonesia untuk dapat bersaing dengan negara-negara lainnya.
Adapun program pembangunan infrastruktur masih terus dilanjutkan bersamaan dengan itu.', "Ir. H. Joko Widodo (Indonesia: [dʒɔkɔ widɔdɔ]; lahir 21 Juni
1961), lebih dikenal sebagai Jokowi, adalah presiden Indonesia ke-7 yang mulai menjabat sejak 20 Oktober  Terpilih dalam pemilu tahun 2014, Jokowi menjadi
```

potongan script output sesudah dilakukan filterisasi penghapusan baris paragraf < 5 kata

Kemudian, setelah dilakukan proses filterisasi kata-kata tersebut dihilangkan, karena merupakan paragraf dengan kata kurang dari 5.

3.1.3 Proses Pengubahan *Output* Ke Dalam Dua Dimensi

```python
#melakukan cleaning slice n, filterisasi penghapusan baris paragraf < 5 kata, dan pengubahan output menjadi dua dimensi
#penggunaan library
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.service import Service as ChromeService
import re

#menghubungkan dengan chrome
chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(service = ChromeService(ChromeDriverManager().install()))

#define function utama yang akan dijalankan
def main():

    data = []  #menyimpan data ke dalam list

    with open("hasil-search-url.txt", "r", encoding = "utf-8") as f:
        urls = [line.rstrip("\n") for line in f][:2] #melakukan penghapusan \n pada dua url teratas

    #looping for
    for u in urls:
        driver.get(u)
        elems = driver.find_elements("tag name", "p") #melakukan ekstraksi paragraf

        #menyimpan teks kedalam sub list dari list data, sehingga menjadi 2 dimensi
        paragraphs = []
        for elem in elems:
            cleaned_text = re.sub(r'[\\^\n]', "", elem.text) #membersihkan teks lebih lanjut
            cleaned_text = re.sub(r'(\[\d+\])|(\d+\s*([.!?])|(\d+\s*(?=\n|$)))', "", cleaned_text) #penghapusan pola angka []
            if len(cleaned_text.split()) >= 5: #filterisasi pengahapusan jumlah baris paragraf < 5 kata
                paragraphs.append(cleaned_text)
        data.append(paragraphs)

    driver.close()

    #hasilnya akan disimpan dalam "cleaning-dua-dimensi.txt"
    with open("dua-dimensi-cleaning.txt", "w", encoding = "utf-8") as f:
        f.write(str(data))

if __name__ == "__main__":
    main()
```

kode script pembersihan slice n, filterisasi penghapusan baris paragraf < 5 kata, pengubahan output dua dimensi

Pada kode *script* di atas, dilakukan penambahan kode untuk mengubah output menjadi *list* dua dimensi. Hal ini bertujuan agar teks dari dua *url* teratas didalam file "hasil-search-url.txt" terdapat didalam format yang lebih terstruktur untuk dipahami, diolah, dan dianalisis lebih lanjut. Dengan menggunakan *library* sama seperti kode sebelumnya. Terdapat fungsi utama *main ( )* yang dipergunakan sebagai skrip utama dengan adanya list data untuk menyimpan hasilnya dan menggunakan teks dari setiap elemen paragraf "p".

Kemudian terdapat proses pembersihan *slice n* dan filterisasi untuk menghilangkan baris paragraf yang kurang dari lima kata dengan menggunakan algoritma yang sama seperti penjelasan sebelumnya yang akan ditampung terlebih dahulu didalam *list* paragraphs sebagai sub *list* data. Setelah melakukan pemroses semua paragraf, *list* paragraphs tersebut ditambahkan ke dalam *list* utama data berdasarkan setiap *url* yang diambil. Hasilnya akan disimpan dalam file "dua-dimensi-cleaning.txt".

```
#dengan hasil url dari scraping yang berada dalam file "hasil-search-url.txt" akan dijalankan skrip "dua-dimensi-cleaning.py"
#untuk melakukan pembersihan \n, fiterisasi penghapusan baris paragraf < 5 kata, pengubahan output kedalam list dua dimensi
#dan hasilnya akan disimpan dalam "dua-dimensi-cleaning.txt"

!python dua-dimensi-cleaning.py
```

menjalankan kode script

Dilakukan proses menjalankan (*run*) kode *script script* python pada "dua-dimensi-cleaning.py" dengan file *input* dan *output* sesuai yang telah dituliskan dalam kode *script*. Dengan hasil seperti di bawah ini.

```
[['Ir. H. Joko Widodo adalah Presiden ke-7 Republik Indonesia yang mulai menjabat sejak 20 Oktober  Lahir di Surakarta, Jawa Tengah, pada 21 Juni 1961, Joko Widodo pertama kali terjun ke
pemerintahan sebagai Wali Kota Surakarta (Solo) pada 28 Juli 2005 hingga 1 Oktober ', 'Selepas itu, Joko Widodo menjabat sebagai Gubernur DKI Jakarta pada 15 Oktober 2012 sebelum terpilih
sebagai Presiden Republik Indonesia pada Pemilihan Presiden (Pilpres)  Saat Pilpres tersebut Joko Widodo terpilih bersama pasangannya, Jusuf Kalla.', 'Dalam Pilpres 2019, Joko Widodo
kembali terpilih sebagai Presiden Republik Indonesia untuk masa jabatannya yang kedua. Kali ini, Joko Widodo didampingi oleh Wakil Presiden K.H. Ma'ruf Amin dan dilantik pada 20 Oktober
2019 untuk masa jabatan 2019 hingga 2024 mendatang.', 'Pembangunan infrastruktur menjadi program prioritas di masa kepemimpinannya yang pertama. Pembangunan yang dilakukan secara merata
hingga ke daerah terluar Indonesia ini dilakukan untuk mengejar ketertinggalan Indonesia dalam sektor ini dibandingkan negara-negara lain.', 'Program prioritas tersebut dibarengi dengan
program berupa bantuan sosial seperti Kartu Indonesia Pintar (KIP), Kartu Indonesia Sehat (KIS), hingga Program Keluarga Harapan (PKH). Selain itu, sejak awal masa jabatannya, Joko Widodo
juga mengupayakan reforma agraria dengan salah satunya melakukan percepatan penerbitan sertifikat hak atas tanah untuk mengurangi terjadinya sengketa lahan oleh karena ketidaan
sertifikat.', 'Di masa jabatannya yang kedua, Joko Widodo mengalihkan fokus pemerintahan pada pembangunan dan peningkatan kapasitas sumber daya manusia Indonesia untuk dapat bersaing
dengan negara-negara lainnya. Adapun program pembangunan infrastruktur masih terus dilanjutkan bersamaan dengan itu.'], ['Ir. H. Joko Widodo (Indonesia: [ʤɔkɔ widɔdɔ]; lahir 21 Juni
1961), lebih dikenal sebagai Jokowi, adalah presiden Indonesia ke-7 yang mulai menjabat sejak 20 Oktober  Terpilih dalam pemilu tahun 2014, Jokowi menjadi presiden Indonesia pertama yang
bukan berasal dari elite politik atau militer Indonesia. Ia terpilih bersama Wakil Presiden Jusuf Kalla dan kembali terpilih bersama Wakil Presiden Ma'ruf Amin pada tahun  Sebelumnya ia
menjabat sebagai Wali Kota Surakarta dan Gubernur DKI Jakarta.", 'Jokowi mengawali karier politiknya sebagai wali kota Surakarta, sejak 28 Juli 2005 hingga 1 Oktober 2012, didampingi F.X.
Hadi Rudyatmo sebagai wakil wali kota. Dua tahun menjalani periode keduanya menjadi Wali Kota Surakarta, Jokowi ditunjuk oleh partainya, Partai Demokrasi Indonesia Perjuangan (PDI-P),
untuk bersaing dalan Pilkada Jakarta 2012 berpasangan dengan Basuki Tjahaja Purnama.', 'Joko Widodo berasal dari keluarga sederhana, rumahnya pernah digusur sebanyak tiga kali ketika ia
```

output kode script pengubahan output dua dimensi

Dapat diperlihatkan bahwa output dari kode memiliki gambaran seperti ini [[‘url 1 paragraf 1’, ‘url 1 paragraf 2’, ..], [‘url 2 paragraf 1’, ‘url 2 paragraf 2’, …]]. Hal ini berarti, *output* dari kode tersebut tersimpan dalam bentuk *list* dua dimensi yang mengandung tiap paragraf yang telah dilakukan pembersihan dengan kriteria yang telah ditentukkan, kemudian dimasukkan ke dalam sub *list* sesuai dengan *url* nya.

3.2 *Notebook*

```
#penggunaan library
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.chrome.service import Service as ChromeService
from selenium.webdriver.chrome.options import Options as ChromeOptions
import re


#menghubungkan dengan chrome
chrome_options = ChromeOptions()
chrome_options.add_argument("--headless")
driver = webdriver.Chrome(service=ChromeService(ChromeDriverManager().install()))

#define function untuk berita dengan dua url teratas dari file "hasil-search-url.txt"
def read_urls(file_name):
    with open(file_name, "r", encoding = "utf-8") as f:
        urls = [line.rstrip("\n") for line in f][:2]
    return urls

#define function untuk pembersihan paragraf \n, \, ^, dan pola angka [] seperti contoh [9]
def clean_paragraph(paragraph):
    paragraph = re.sub(r'(\[\d+\])|(\d+\s*([.!?])|(\d+\s*(?=\n|$)))', '', paragraph)
    paragraph = paragraph.replace('\n', '').replace('\\', '').replace('"', "'")
    return paragraph


#define function untuk scraping
def scraping():

    data = [] #menyimpan data dalam list

    for url in urls:
        driver.get(url)
        elems = driver.find_elements("tag name", "p")

        paragraphs = [] #menyimpan teks kedalam sublist dari list data, sehingga menjadi dua dimensi
        for elem in elems:
            if len(elem.text.split()) >= 5: #filterisasi penghapusan baris < 5
                paragraphs.append(clean_paragraph(elem.text))
        data.append(paragraphs)

    driver.quit()
    return data
```

```
name_url = "hasil-search-url.txt"
urls = read_urls(name_url)
output_cleaning = scraping() #hasilnya disimpan dalam variabel output_cleaning
```

kode script pembersihan (cleaning) menggunakan notebook

Pada kode *script* di atas, dilakukan proses serangkaian pembersihan (*cleaning*) berdasarkan paragraf dari setiap *url*-nya. Dengan melakukan library seperti pada kode sebelumnya dan penghubungan ke *google chrome* dengan opsi tanpa GUI. Terdapat fungsi *read_urls ( )* yang dipergunakan untuk membaca dan menyimpan dua *url* berita dari file "hasil-search-url.txt" yang akan disimpan dalam daftar. Terdapat fungsi *clean_paragraph ( )* dipergunakan untuk membersihkan setiap paragraf yang memiliki karakter khusus, seperti penghapusan pola angka dalam kurung siku , "\n", "\",dan tanda kutip ganda.

Dan, terdapat fungsi utama *scraping ( )* yang dipergunakan untuk melakukan proses pembersihan dari hasil *scraping* dengan mencari elemen-elemen paragraf "p" dari setiap *url*. Kemudian, paragraf tersebut akan dibersihkan dengan menggunakan fungsi *cleaning_paragraph* dan hasilnya akan disimpan terlebih dahulu dalam *list* paragraph sebagai *sub list* data. Setelah melakukan pemroses semua paragraf, *list* paragraphs tersebut ditambahkan ke dalam *list* utama data berdasarkan setiap *url* yang diambil. Hasilnya akan disimpan dalam variabel *output_cleaning*.

```
#menampilkan hasilnya

output_cleaning
```

```
[['Ir. H. Joko Widodo adalah Presiden ke-7 Republik Indonesia yang mulai menjabat sejak 20 Oktober  Lahir di Surakarta, Jawa Tengah, pada 21 Juni 1961, Joko Widodo pertama kali ter
jun ke pemerintahan sebagai Wali Kota Surakarta (Solo) pada 28 Juli 2005 hingga 1 Oktober ',
  'Selepas itu, Joko Widodo menjabat sebagai Gubernur DKI Jakarta pada 15 Oktober 2012 sebelum terpilih sebagai Presiden Republik Indonesia pada Pemilihan Presiden (Pilpres)  Saat
Pilpres tersebut Joko Widodo terpilih bersama pasangannya, Jusuf Kalla.',
  'Dalam Pilpres 2019, Joko Widodo kembali terpilih sebagai Presiden Republik Indonesia untuk masa jabatannya yang kedua. Kali ini, Joko Widodo didampingi oleh Wakil Presiden K.H.
Ma'ruf Amin dan dilantik pada 20 Oktober 2019 untuk masa jabatan 2019 hingga 2024 mendatang.',
  'Pembangunan infrastruktur menjadi program prioritas di masa kepemimpinannya yang pertama. Pembangunan yang dilakukan secara merata hingga ke daerah terluar Indonesia ini dilakuk
an untuk mengejar ketertinggalan Indonesia dalam sektor ini dibandingkan negara-negara lain.',
  'Program prioritas tersebut dibarengi dengan program berupa bantuan sosial seperti Kartu Indonesia Pintar (KIP), Kartu Indonesia Sehat (KIS), hingga Program Keluarga Harapan (PK
H). Selain itu, sejak awal masa jabatannya, Joko Widodo juga mengupayakan reforma agraria dengan salah satunya melakukan percepatan penerbitan sertifikat hak atas tanah untuk mengu
rangi terjadinya sengketa lahan oleh karena ketiadaan sertifikat.',
  'Di masa jabatannya yang kedua, Joko Widodo mengalihkan fokus pemerintahan pada pembangunan dan peningkatan kapasitas sumber daya manusia Indonesia untuk dapat bersaing dengan ne
gara-negara lainnya. Adapun program pembangunan infrastruktur masih terus dilanjutkan bersamaan dengan itu.'],
 ["Ir. H. Joko Widodo (Indonesia: [dʒɔkɔ widɔdɔ]; lahir 21 Juni 1961), lebih dikenal sebagai Jokowi, adalah presiden Indonesia ke-7 yang mulai menjabat sejak 20 Oktober  Terpilih d
alam pemilu tahun 2014, Jokowi menjadi presiden Indonesia pertama yang bukan berasal dari elite politik atau militer Indonesia. Ia terpilih bersama Wakil Presiden Jusuf Kalla dan k
embali terpilih bersama Wakil Presiden Ma'ruf Amin pada tahun  Sebelumnya ia menjabat sebagai Wali Kota Surakarta dan Gubernur DKI Jakarta.",
  'Jokowi mengawali karier politiknya sebagai wali kota Surakarta, sejak 28 Juli 2005 hingga 1 Oktober 2012, didampingi F.X. Hadi Rudyatmo sebagai wakil wali kota. Dua tahun menjal
ani periode keduanya menjadi Wali Kota Surakarta, Jokowi ditunjuk oleh partainya, Partai Demokrasi Indonesia Perjuangan (PDI-P), untuk bersaing dalam Pilkada Jakarta 2012 berpasang
an dengan Basuki Tjahaja Purnama.',
```

output kode script pembersihan (cleaning) menggunakan notebook

Dilakukan proses pemanggilan variabel *output_cleaning* dan akan menghasilkan sebuah *list* dua dimensi yang berisi paragraf-paragraf yang telah diambil dari dua *url* teratas yang telah di *scrape* dan di *cleaning*. Dimana, setiap *sublist* di dalam *list* utama mewakili setiap paragraf dari satu *url*. Untuk membuktikan apakah *output* yang dihasilkan benar dapat dilakukan proses pengaksesan indeks dalam struktur data list dua dimensi.

3.2.1 Pengaksesan *Url* 1 Paragraf 2

```
#melakukan penampilan output url ke-1 paragraf ke-2
#dengan indeks dimulai dari 0

output_cleaning[0][1]
```

```
'Selepas itu, Joko Widodo menjabat sebagai Gubernur DKI Jakarta pada 15 Oktober 2012 sebelum terpilih sebagai Presiden Republik Indonesia pada Pemilihan Presiden (Pilpres)  Saat Pi
lpres tersebut Joko Widodo terpilih bersama pasangannya, Jusuf Kalla.'
```

output url 1 paragraf 2

Pada kode *script* di atas dilakukan proses pengaksesan indeks *url* ke-satu dan paragraf ke-dua, dengan menggunakan sintaksi output_cleaning[0][1] dapat diketahui bahwa indeks dimulai dari angka 0 dan *output_cleaning* merupakan variabel yang menyimpan hasil pemrosesan. Maka dapat ditampilkan hasil seperti gambar diatas, dimana hasil tersebut memiliki kecocokan dengan berita pada *url* satu yakni https://www.presidenri.go.id/presiden-joko-widodo/ . Seperti pada potongan gambar dibawah ini.

output potongan berita url 1

3.2.2 Pengaksesan *Url* 2 Paragraf 1

```
#melakukan penampilan output url ke-2 paragraf ke-1
#dengan indeks dimulai dari 0

output_cleaning[1][0]
```

```
"Ir. H. Joko Widodo (Indonesia: [dʒoko widodo]; lahir 21 Juni 1961), lebih dikenal sebagai Jokowi, adalah presiden Indonesia ke-7 yang mulai menjabat sejak 20 Oktober  Terpilih dal
an pemilu tahun 2014, Jokowi menjadi presiden Indonesia pertama yang bukan berasal dari elite politik atau militer Indonesia. Ia terpilih bersama Wakil Presiden Jusuf Kalla dan kem
bali terpilih bersama Wakil Presiden Ma'ruf Amin pada tahun  Sebelumnya ia menjabat sebagai Wali Kota Surakarta dan Gubernur DKI Jakarta."
```

output url 2 paragraf 1

Pada kode *script* di atas dilakukan proses pengaksesan indeks *url* ke-dua dan paragraf ke-satu dengan menggunakan sintaksi output_cleaning[1][0] dapat diketahui bahwa indeks dimulai dari angka 0 dan *output_cleaning* merupakan variabel yang menyimpan hasil pemrosesan. Maka dapat ditampilkan hasil seperti gambar diatas, dimana hasil tersebut memiliki kecocokan dengan berita pada *url* satu yakni https://id.wikipedia.org/wiki/Joko_Widodo. Seperti pada potongan gambar dibawah ini.

output potongan berita url 2

Kesimpulan :

Pada penugasan *web scraping* dengan kata kunci "Joko Widodo" ini saya telah melakukan proses pengembangan kode *script* dengan menggunakan beberapa *library* seperti Selenium dan BeautifulSoup untuk melakukan proses *web scraping*, seta *library* Regex untuk melakukan manipulasi *string*. Dengan melakukan proses perolehan judul berita dan *url*, proses ektraksi teks berita, dan prses pembersihan (*cleaning*) yang dilakukan untuk penghapusan *slice n* di dua *url* teratas dalam *file* "hasil-search-url.txt" dibaris terlebih dahulu kemudian dilakukan penghapusan karakter lebih lanjut seperti "\n, \, ^, [9]" diparagraf.

Hal ini dilakukan, ketika hanya melakukan penghapusan dibaris saja masih terdapat karakter yang tidak diinginkan. Hasilnya akan disimpan dalam sebuah *list* dua dimensi dengan gambaran [['url 1 paragraf 1', 'url 1 paragraf 2', ..], ['url 2 paragraf 1', 'url 2 paragraf 2', …]] dimana setiap *sublist* di dalam *list* utama mewakili setiap paragraf dari satu *url*. Dengan menggunakan *python file* dan *notebook* untuk tujuan pendapatan *output* yang sama, namun cara pengaksesan yang berbeda. Pada *python file* hasilnya akan disimpan dalam file txt dan pada *notebook* hasilnya akan disimpan dalam suatu variabel saja. Oleh karena itu, pada *notebook* memungkinkan untuk melakukan proses penampilan pada indeks tertentu.

Link google drive berisi *word* yang belum di *compress*, *kode script*, dan *outputnya* :
https://drive.google.com/drive/folders/1yYjH2CYSAw0sDs3fhlsmPFgL1Hm_bBXn?usp=sharing