LAPORAN

RENCANA TUGAS MANDIRI (RTM) Ke-II

MATA KULIAH BIG DATA KELAS B

“WEB SCRAPING”

DISUSUN OLEH:

Aulia Nur Fitriani (21083010051)

DOSEN PENGAMPU:

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

PROGRAM STUDI SAINS DATA

FAKULTAS ILMU KOMPUTER

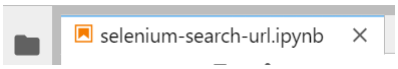
UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” JAWA TIMUR

2023

TUGAS 2

Mengimplementasikan web scraping/web crawling dari suatu situs tertentu dengan studi kasus menggunakan sumber data tertentu menggunakan bahasa pemrograman Python dan mendokumentasikan langkah demi langkah prosesnya. Berikut adalah langkah-langkah nya:

1. Membuat file ipynb bernama selenuim-search-url.ipynb



2. Install package BeautifulSoup4, Selenium, dan Webdriver-Manager

```
[2]: pip instal BeautifulSoUp4

Persyaratan sudah terpenuhi: BeautifulSoUp4 di c:\users\user\anaconda3\lib\site-packages (4.9.3)
Persyaratan sudah terpenuhi: saringan sup>1,2; python_version >= "3.0" di c:\users\user\anaconda3\lib\site-packages (dari BeautifulSoUp4) (2.0.1)
Catatan: Anda mungkin perlu me-restart kernel untuk menggunakan paket yang diperbarui.

[3]: pip instal selenium

Persyaratan sudah terpenuhi: selenium di c:\users\user\anaconda3\lib\site-packages (4.8.2)Catatan: Anda mungkin perlu me-restart kernel untuk menggunakan p
aket yang diperbarui.
KESALAHAN: Setelah Oktober 2020 Anda mungkin mengalami kesalahan saat memasang atau memperbarui paket. Ini karena pip akan mengubah caranya menyelesaikan k
onflik ketergantungan.

Kami menyarankan Anda menggunakan --use-feature=2020-resolver untuk menguji paket Anda dengan resolver baru sebelum menjadi default.

permintaan 2.24.0 membutuhkan urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1, tetapi Anda akan memiliki urllib3 1.26.14 yang tidak kompatibel.

Persyaratan sudah terpenuhi: trio~=0.17 di c:\users\user\anaconda3\lib\site-packages (dari Selenium) (0.22.0)
Persyaratan sudah terpenuhi: trio-websocket~=0.9 di c:\users\user\anaconda3\lib\site-packages (dari Selenium) (0.9.2)
Mengumpulkan urllib3[socks]~=1.26
  Menggunakan cache urllib3-1.26.14-py2.py3-none-any.whl (140 kB)
Persyaratan sudah terpenuhi: certifi>=2021.10.8 di c:\users\user\anaconda3\lib\site-packages (dari Selenium) (2022.12.7)
Persyaratan sudah terpenuhi: attrs>=19.2.0 di c:\users\user\anaconda3\lib\site-packages (dari trio~=0.17->Selenium) (20.3.0)
Persyaratan sudah terpenuhi: sniffio di c:\users\user\anaconda3\lib\site-packages (dari trio~=0.17->selenium) (1.3.0)
Persyaratan sudah terpenuhi: cffi>=1.14; os_name == "nt" dan execution_name != "pypy" di c:\users\user\anaconda3\lib\site-packages (dari trio~=0.17->seleni
um) (1.14.3)
Persyaratan sudah terpenuhi: async-generator>=1.9 di c:\users\user\anaconda3\lib\site-packages (dari trio~=0.17->Selenium) (1.10)
Persyaratan sudah terpenuhi: exceptiongroup>=1.0.0rc9; python_version < "3.11" di c:\users\user\anaconda3\lib\site-packages (dari trio~=0.17->selenium) (1.
```

3. Kemudian membuat file py dengan nama selenium-search-url dan membuat script code seperti dibawah ini.

selenium-search-url.ipynb | selenium-search-url.py

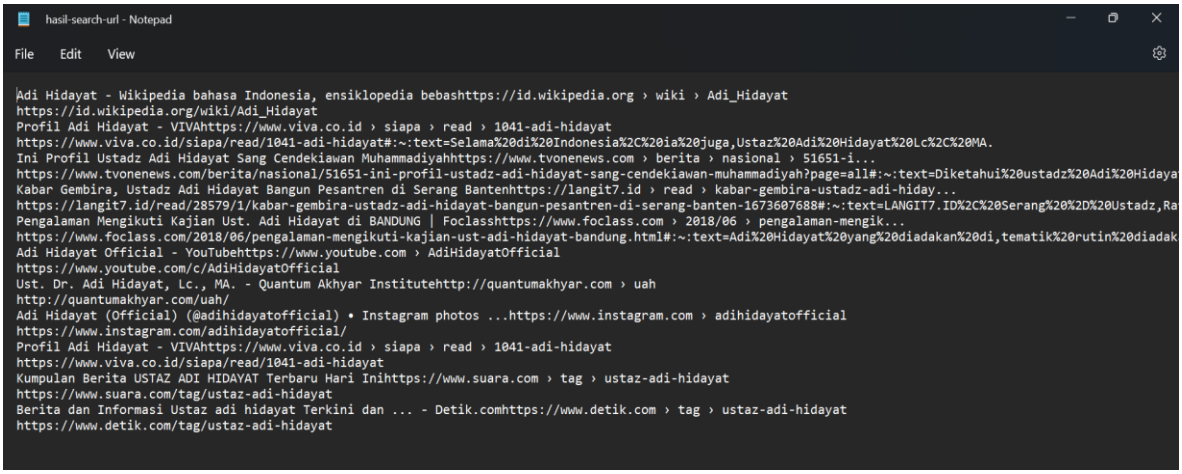
```
dari  bs4  impor  BeautifulSoup
dari  webdriver impor selenium
dari  webdriver_manager . chrome  impor  ChromeDriverManager
chrome_options  =  driver web . ChromeOptions ()
chrome_options . add_argument ( "--tanpa kepala" )
driver  =  driver web . Chrome ( ChromeDriverManager (). pasang (), chrome_options = chrome_options )
# Permintaan untuk mendapatkan tautan
permintaan  =  'Adi Hidayat'
#Links = [] # Memulai daftar kosong untuk menangkap hasil akhir
# Tentukan jumlah halaman di pencarian google, setiap halaman berisi 10 #link
n_halaman  =  10

untuk  halaman  dalam  rentang ( 1 , n_pages ):
    url  =  "http://www.google.com/search?q="  +  query  +  "&start="  +  str (( halaman  -  1 ) *  10 )
    sopir . dapatkan ( url )
    sup  =  BeautifulSoup ( driver .page_source , 'html.parser ' )
    # sup = BeautifulSoup(r.text, 'html.parser')

    cari  =  sup . find_all ( 'div' , class_ = "yuRUbf" )
    untuk  h  dalam  pencarian :
        #Links.append(haget('href'))
        cetak ( h.a.teks ) _ _ _ _
        print ( h .a . get ( ' href' ))
```

4. Kemudian jalankan syntax berikut untuk memunculkan output pada file dengan nama hasil-search-url.txt

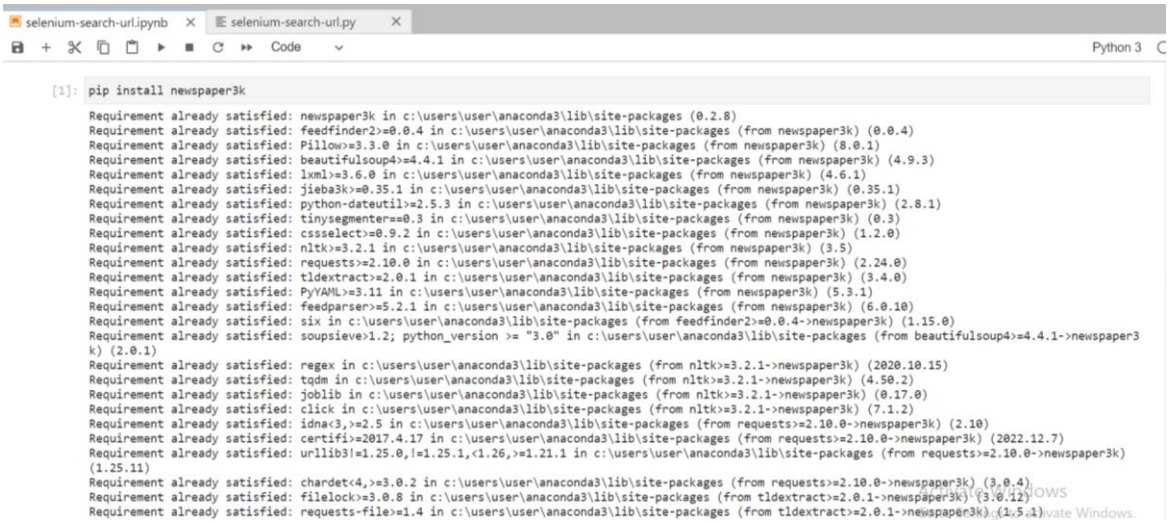
```
[6]: !python selenium-search-url.py > hasil-search-url.txt

selenium-search-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-search-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-search-url.py", line 22, in <module>
    print(h.a.text)
  File "C:\Users\USER\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 characters in position 19-21: character maps to <undefined>
```

5. Jika berhasil, akan muncul file bernama hasil-search-url.txt pada folder yang sama dengan file ipynb dan py diatas.

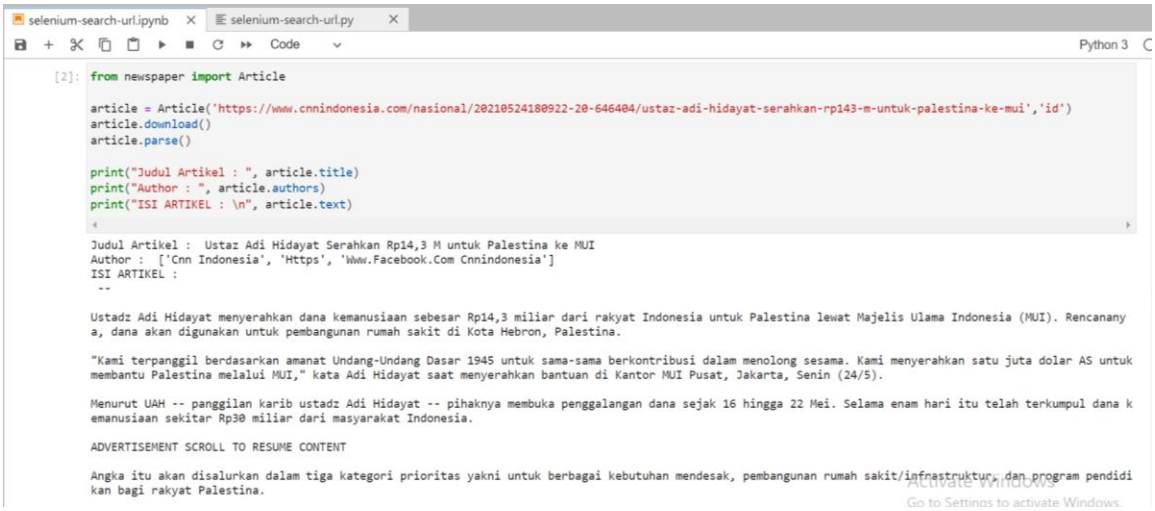
6. Isi dari file hasil-search-url.txt nya seperti berikut :



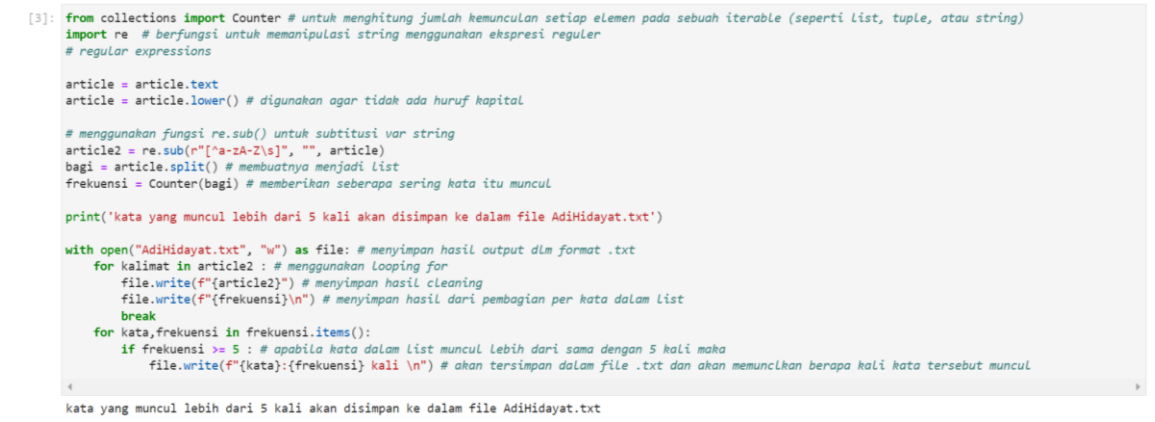
7. Selanjutnya untuk melakukan scraping pada website yang diinginkan adalah dengan cara menginstall package newspaper3k



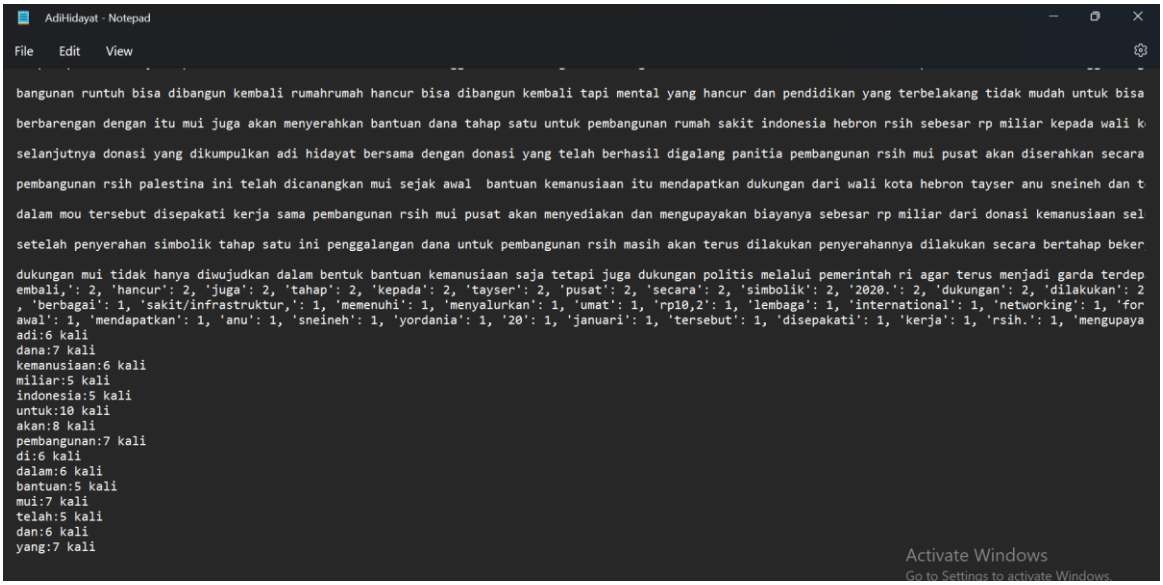
8. Setelah itu menjalankan code script berikut dan memasukkan link mana yang mau di scrapping dan saya memilih website dari CNN



9. Melakukan pembersihan data dengan kata munxul lebih dari 5 kali dan saya menyimpan pada file txt dengan nama AdiHidayat



10. Berikut output yang dihasilkan dari code script diatas, dan dapat dihasilkan kata yang muncul 5 kali atau leboih daroi 5 kali,.



11. Selanjutnya, membuat 2d nya dengan menginstall packages wordcloud dan matplotlib

```
[4]: pip install wordcloud

Collecting wordcloud
  Downloading wordcloud-1.8.2.2-cp38-cp38-win_amd64.whl (152 kB)
Requirement already satisfied: matplotlib in c:\users\user\anaconda3\lib\site-packages (from wordcloud) (3.3.2)
Requirement already satisfied: numpy>=1.6.1 in c:\users\user\anaconda3\lib\site-packages (from wordcloud) (1.19.2)
Requirement already satisfied: pillow in c:\users\user\anaconda3\lib\site-packages (from wordcloud) (8.0.1)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.4.7)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.8.1)
Requirement already satisfied: cycler>=0.10 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (0.10.0)
Requirement already satisfied: certifi>=2020.06.20 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2022.12.7)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\user\anaconda3\lib\site-packages (from matplotlib->wordcloud) (1.3.0)
Requirement already satisfied: six>=1.5 in c:\users\user\anaconda3\lib\site-packages (from python-dateutil>=2.1->matplotlib->wordcloud) (1.15.0)
Installing collected packages: wordcloud
Successfully installed wordcloud-1.8.2.2
Note: you may need to restart the kernel to use updated packages.

[5]: pip install matplotlib

Requirement already satisfied: matplotlib in c:\users\user\anaconda3\lib\site-packages (3.3.2)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (2.8.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (8.0.1)
Requirement already satisfied: numpy>=1.15 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (1.19.2)
Requirement already satisfied: cycler>=0.10 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (1.3.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (2.4.7)
Requirement already satisfied: certifi>=2020.06.20 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (2022.12.7)
Requirement already satisfied: six>=1.5 in c:\users\user\anaconda3\lib\site-packages (from python-dateutil>=2.1->matplotlib) (1.15.0)
Note: you may need to restart the kernel to use updated packages.
```


12. Untuk memunculkan worldcloud, menggunakan code script berikut ini

```
[17]: from wordcloud import WordCloud # library wordcloud
import matplotlib.pyplot as plt
import numpy as np
from PIL import Image # untuk memanipulasi gambar

teks = ''.join(bagi)
mask = np.array(Image.open("ustadzAdi.jpeg")) # memanipulasi gambar dgn array
def warna_frekuensi(word, font_size, position, orientation, random_state=None, **kwargs): # settings tampilan
    return tuple(np.random.randint(0, 255, 3))
wordcloud = WordCloud(width = 800, height = 800, background_color ='white', min_font_size = 10).generate(teks)

plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off") # menghilangkan binghai pada library matplotlib
plt.tight_layout(pad = 0) # menyesuaikan layout agar tidak tumpang tindih sehingga dapat dibaca

plt.show()
```

13. Untuk membuat diagram lingkaran, menggunakan code script berikut ini

```
[10]: daftar_data = bagi
```

```
[11]: import matplotlib.pyplot as plt
count_data = Counter(daftar_data)

# Mengambil nilai dan label dari setiap item pada dictionary count_data
values = count_data.values()
labels = count_data.keys()

# Membuat diagram lingkaran
plt.pie(values, labels=labels, autopct='%1.1f%%')

# Menampilkan diagram
plt.show()
```

14. Kemudian kita munculkan diagram batang, menggunakan code script berikut ini

```
[12]: plt.figure(figsize=(25, 10)) # mengatur ukuran gambar
plt.rcParams.update({'font.size': 10}) # mengatur ukuran font
plt.bar(labels, values)
plt.xticks(rotation=90) # mengatur agar font label x berotasi 90 derajat
plt.show()
```