

LAPORAN
RENCANA TUGAS MANDIRI (RTM) Ke-II
MATA KULIAH BIG DATA KELAS B
“WEB SCRAPING”



DISUSUN OLEH:

Aulia Nur Fitriani (21083010051)

DOSEN PENGAMPU:

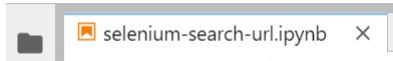
Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

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 selenium-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) (1.10)
```

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

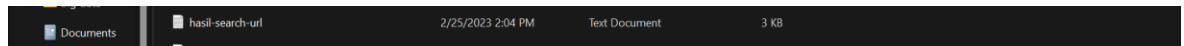
```
selenium-search-url.py
1 dari bs4 impor BeautifulSoup
2 dari webdriver impor selenium
3 dari webdriver_manager . chrome impor ChromeDriverManager
4 chrome_options = driver web . ChromeOptions ()
5 chrome_options . add_argument ( "--tanpa kepala" )
6 driver = driver web . Chrome ( ChromeDriverManager (). pasang (), chrome_options = chrome_options )
7 # Permintaan untuk mendapatkan tautan
8 permintaan = 'Adi Hidayat'
9 #Links = [] # Memulai daftar kosong untuk menangkap hasil akhir
10 # Tentukan jumlah halaman di pencarian google, setiap halaman berisi 10 #Link
11 n_halaman = 10
12
13 untuk halaman dalam rentang ( 1 , n_pages ):
14     url = "http://www.google.com/search?q=" + query + "&start=" + str ( halaman - 1 ) * 10
15     sopir . dapatkan ( url )
16     sup = BeautifulSoup ( driver .page_source , 'html.parser ' )
17     # sup = BeautifulSoup(r.text, 'html.parser')
18
19     cari = sup . find_all ( 'div' , class_ = "yuRuf" )
20     untuk h dalam pencarian :
21         #Links.append(ha.get('href'))
22         cetak ( h.a.teks )
23         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 :

```
Adi Hidayat - Wikipedia bahasa Indonesia, ensiklopedia bebashttps://id.wikipedia.org > wiki > Adi_Hidayat
https://id.wikipedia.org/wiki/Adi_Hidayat
Profil Adi Hidayat - VIVAhttps://www.viva.co.id > siapa > read > 1041-adi-hidayat
https://www.viva.co.id/siapa/read/1041-adi-hidayat#:~:text=Selama%20di%20Indonesia%2C%20ia%20juga,Ustaz%20Adi%20Hidayat%20Lc%2C%20MA.
Ini Profil Ustadz Adi Hidayat Sang Cendekiawan Muhammadiyahhttps://www.tvonenews.com > berita > nasional > 51651-i...
https://www.tvonenews.com/berita/nasional/51651-ini-profil-ustadz-adi-hidayat-sang-cendekiawan-muhammadiyah?page=all#:~:text=Diketahui%20ustadz%20Adi%20Hidayat,Kabar Gembira, Ustadz Adi Hidayat Bangun Pesantren di Serang Bantenhttps://langit7.id > read > kabar-gembira-ustadz-adi-hidayat...
https://langit7.id/read/28579/1/kabar-gembira-ustadz-adi-hidayat-bangun-pesantren-di-serang-banten-1673607688#:~:text=LANGIT7.ID%2C%20Serang%20%2D%20Ustadz,Ra
Pengalaman Mengikuti Kajian Ust. Adi Hidayat di BANDUNG | Foclasshttps://www.foclass.com > 2018/06 > pengalaman-mengik...
https://www.foclass.com/2018/06/pengalaman-mengikuti-kajian-ust-adi-hidayat-bandung.html#:~:text=Adi%20Hidayat%20yang%20diadakan%20di,tematik%20rutin%20diadak
Adi Hidayat Official - YouTubehttps://www.youtube.com > AdiHidayatOfficial
https://www.youtube.com/c/AdiHidayatOfficial
Ust. Dr. Adi Hidayat, Lc., MA. - Quantum Akhyar Institutehttp://quantumakhyar.com > uah
http://quantumakhyar.com/uah/
Adi Hidayat (Official) (@adihidayatofficial) • Instagram photos ...https://www.instagram.com > adihidayatofficial
https://www.instagram.com/adihidayatofficial/
Profil Adi Hidayat - VIVAhttps://www.viva.co.id > siapa > read > 1041-adi-hidayat
https://www.viva.co.id/siapa/read/1041-adi-hidayat
Kumpulan Berita USTAZ ADI HIDAYAT Terbaru Hari Inihttps://www.suara.com > tag > ustaz-adi-hidayat
https://www.suara.com/tag/ustaz-adi-hidayat
Berita dan Informasi Ustadz Adi Hidayat Terkini dan ... - Detik.comhttps://www.detik.com > tag > ustaz-adi-hidayat
https://www.detik.com/tag/ustaz-adi-hidayat
```

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

```
selenium-search-url.ipynb x selenium-search-url.py x Python 3 C
[1]: pip install newspaper3k
Requirement already satisfied: newspaper3k in c:\users\user\anaconda3\lib\site-packages (0.2.8)
Requirement already satisfied: feedfinder2>=0.4 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (0.0.4)
Requirement already satisfied: Pillow>=3.3.0 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (8.0.1)
Requirement already satisfied: BeautifulSoup4>=4.4.1 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (4.9.3)
Requirement already satisfied: lxml>=3.6.0 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (4.6.1)
Requirement already satisfied: jieba3k>=0.35.1 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (0.35.1)
Requirement already satisfied: python-dateutil>=2.5.3 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (2.8.1)
Requirement already satisfied: tinyssegmenter>=0.3 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (0.3)
Requirement already satisfied: cssselect>=0.9.2 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (1.2.0)
Requirement already satisfied: nltk>=3.2.1 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (3.5)
Requirement already satisfied: requests>=2.10.0 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (2.24.0)
Requirement already satisfied: tldextract>=2.0.1 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (3.4.0)
Requirement already satisfied: PyYAML>=3.11 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (5.3.1)
Requirement already satisfied: feedparser>=5.2.1 in c:\users\user\anaconda3\lib\site-packages (from newspaper3k) (6.0.10)
Requirement already satisfied: six in c:\users\user\anaconda3\lib\site-packages (from feedfinder2>=0.0.4->newspaper3k) (1.15.0)
Requirement already satisfied: soupsieve>1.2; python_version >= "3.0" in c:\users\user\anaconda3\lib\site-packages (from BeautifulSoup4>=4.4.1->newspaper3k) (2.0.1)
Requirement already satisfied: regex in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.2.1->newspaper3k) (2020.10.15)
Requirement already satisfied: tqdm in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.2.1->newspaper3k) (4.50.2)
Requirement already satisfied: joblib in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.2.1->newspaper3k) (0.17.0)
Requirement already satisfied: click in c:\users\user\anaconda3\lib\site-packages (from nltk>=3.2.1->newspaper3k) (7.1.2)
Requirement already satisfied: dnspython>=2.5 in c:\users\user\anaconda3\lib\site-packages (from requests>=2.10.0->newspaper3k) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\user\anaconda3\lib\site-packages (from requests>=2.10.0->newspaper3k) (2022.12.7)
Requirement already satisfied: urllib3>=1.25.0,!=1.25.1,<1.26,>=1.21.1 in c:\users\user\anaconda3\lib\site-packages (from requests>=2.10.0->newspaper3k) (1.25.11)
Requirement already satisfied: chardet<4,>=3.0.2 in c:\users\user\anaconda3\lib\site-packages (from requests>=2.10.0->newspaper3k) (3.0.4)
Requirement already satisfied: filelock>=3.0.8 in c:\users\user\anaconda3\lib\site-packages (from tldextract>=2.0.1->newspaper3k) (3.0.12)
Requirement already satisfied: requests-file>=1.4 in c:\users\user\anaconda3\lib\site-packages (from tldextract>=2.0.1->newspaper3k) (1.5.1)
Requirement already satisfied: private Windows.
```

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

```
selenium-search-url.pyb selenium-search-url.py Python 3 C
[2]: from newspaper import Article

article = Article('https://www.cnnindonesia.com/nasional/20210524180922-20-646404/ustaz-adi-hidayat-serahkan-rp143-m-untuk-palestina-ke-mui', 'id')
article.download()
article.parse()

print("Judul Artikel : ", article.title)
print("Author : ", article.authors)
print("ISI ARTIKEL : \n", article.text)

Judul Artikel : Ustaz Adi Hidayat Serahkan Rp14,3 M untuk Palestina ke MUI
Author : ['Cnn Indonesia', 'Https', 'Www.Facebook.Com Cnnindonesia']
ISI ARTIKEL :
--
Ustadz Adi Hidayat menyerahkan dana kemanusiaan sebesar Rp14,3 miliar dari rakyat Indonesia untuk Palestina lewat Majelis Ulama Indonesia (MUI). Rencananya, dana akan digunakan untuk pembangunan rumah sakit di Kota Hebron, Palestina.

"Kami terpenggil berdasarkan amanat Undang-Undang Dasar 1945 untuk sama-sama berkontribusi dalam menolong sesama. Kami menyerahkan satu juta dolar AS untuk membantu Palestina melalui MUI," kata Adi Hidayat saat menyerahkan bantuan di Kantor MUI Pusat, Jakarta, Senin (24/5).

Menurut UAH -- panggilan karib ustaz Adi Hidayat -- pihaknya membuka penggalangan dana sejak 16 hingga 22 Mei. Selama enam hari itu telah terkumpul dana kemanusiaan sekitar Rp30 miliar dari masyarakat Indonesia.

ADVERTISEMENT SCROLL TO RESUME CONTENT

Angka itu akan disalurkan dalam tiga kategori prioritas yakni untuk berbagai kebutuhan mendesak, pembangunan rumah sakit/infrastruktur, dan program pendidikan bagi rakyat Palestina.
```

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

```
[3]: from collections import Counter # untuk menghitung jumlah kemunculan setiap elemen pada sebuah iterable (seperti list, tuple, atau string)
import re # berfungsi untuk memanipulasi string menggunakan ekspresi reguler
# regular expressions

article = article.text
article = article.lower() # digunakan agar tidak ada huruf kapital

# menggunakan fungsi re.sub() untuk substitusi var string
article2 = re.sub(r"[a-zA-Z\s]", "", article)
bagi = article.split() # membuatnya menjadi list
frekuensi = Counter(bagi) # memberikan seberapa sering kata itu muncul

print('kata yang muncul lebih dari 5 kali akan disimpan ke dalam file AdiHidayat.txt')

with open("AdiHidayat.txt", "w") as file: # menyimpan hasil output dlm format .txt
    for kalimat in article2: # menggunakan looping for
        file.write(f"{kalimat}") # menyimpan hasil cleaning
        file.write(f"{frekuensi}\n") # menyimpan hasil dari pembagian per kata dalam list
        break
    for kata, frekuensi in frekuensi.items():
        if frekuensi >= 5: # apabila kata dalam list muncul lebih dari sama dengan 5 kali maka
            file.write(f"{kata}:{frekuensi} kali \n") # akan tersimpan dalam file .txt dan akan memunculkan berapa kali kata tersebut muncul

kata yang muncul lebih dari 5 kali akan disimpan ke dalam file AdiHidayat.txt
```

- Berikut output yang dihasilkan dari code script diatas, dan dapat dihasilkan kata yang muncul 5 kali atau leboh dari 5 kali,.

```
AdiHidayat - Notepad
File Edit View

bangunan runtuh bisa dibangun kembali rumahrumah hancur bisa dibangun kembali tapi mental yang hancur dan pendidikan yang terbelakang tidak mudah untuk bisa
berbarengan dengan itu mui juga akan menyerahkan bantuan dana tahap satu untuk pembangunan rumah sakit indonesia hebron rsih sebesar rp miliar kepada wali k
selanjutnya donasi yang dikumpulkan adi hidayat bersama dengan donasi yang telah berhasil digalang panitia pembangunan rsih mui pusat akan diserahkan secara
pembangunan rsih palestina ini telah dicanangkan mui sejak awal bantuan kemanusiaan itu mendapatkan dukungan dari wali kota hebron tayser anu sneineh dan t
dalam mou tersebut disepakati kerja sama pembangunan rsih mui pusat akan menyediakan dan mengupayakan biayanya sebesar rp miliar dari donasi kemanusiaan sel
setelah penyerahan simbolik tahap satu ini penggalangan dana untuk pembangunan rsih masih akan terus dilakukan penyerahannya dilakukan secara bertahap beker
dukungan mui tidak hanya diwujudkan dalam bentuk bantuan kemanusiaan saja tetapi juga dukungan politis melalui pemerintah ri agar terus menjadi garda terdepan
embali,': 2, 'hancur': 2, 'juga': 2, 'tahap': 2, 'kepada': 2, 'tayser': 2, 'pusat': 2, 'secara': 2, 'simbolik': 2, '2020': 2, 'dukungan': 2, 'dilakukan': 2
, 'berbagai': 1, 'sakit/infrastruktur': 1, 'memenuhi': 1, 'menyalurkan': 1, 'umat': 1, 'rp10,2': 1, 'lembaga': 1, 'international': 1, 'networking': 1, 'for
awal': 1, 'mendapatkan': 1, 'anu': 1, 'sneineh': 1, 'yordania': 1, '20': 1, 'januari': 1, 'tersebut': 1, 'disepakati': 1, 'kerja': 1, 'rsih.': 1, 'mengupaya
adi:6 kali
dana:7 kali
kemanusiaan:6 kali
miliar:5 kali
indonesia:5 kali
untuk:10 kali
akan:8 kali
pembangunan:7 kali
di:6 kali
dalam:6 kali
bantuan:5 kali
mui:7 kali
telah:5 kali
dan:6 kali
yang:7 kali
```

- 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: cython>=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: cython>=0.10 in c:\users\user\anaconda3\lib\site-packages (from matplotlib) (0.10.0)
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 wordcloud, 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 bingkai 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()
```



Activate Windows
Go to Settings to activate Windows.

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()
```

