MODUL PRAKTIKUM 4

Web Srapping

1. Tujuan

Setelah menyelesaikan modul ini, anda diharapkan dapat :

1. Mahasiswa mampu menginstall tools scrapping
2. Mampu menginstall membuat program scrapping
3. Mampu menyimpan data hasil scrapping pada database

2. Dasar Teori

Web scraping adalah teknik untuk mengambil data dari website secara otomatis menggunakan program atau algoritma tertentu. Proses ini dilakukan dengan cara mengekstrak informasi yang terdapat pada halaman website dan menyimpannya dalam format yang dapat diolah, seperti file CSV atau Excel.

Proses web scraping biasanya dilakukan dengan menggunakan perangkat lunak tertentu atau bahasa pemrograman seperti Python. Beberapa langkah dalam melakukan web scraping antara lain:

1. Menentukan website dan data yang ingin diambil.
2. Menentukan metode dan teknik yang akan digunakan dalam web scraping.
3. Memulai program dan mengakses website yang dituju.
4. Mengambil data yang diinginkan dari halaman web menggunakan teknik seperti parsing atau regular expression.
5. Menyimpan data yang sudah diambil dalam format yang sesuai.

Web scraping dapat digunakan untuk berbagai tujuan, seperti untuk analisis data, riset pasar, atau pengumpulan informasi untuk kepentingan bisnis atau akademik.

Namun, penting untuk diingat bahwa pengambilan data dari website yang dilakukan tanpa izin atau melanggar aturan dari website tersebut dapat menjadi pelanggaran hak cipta dan privasi.

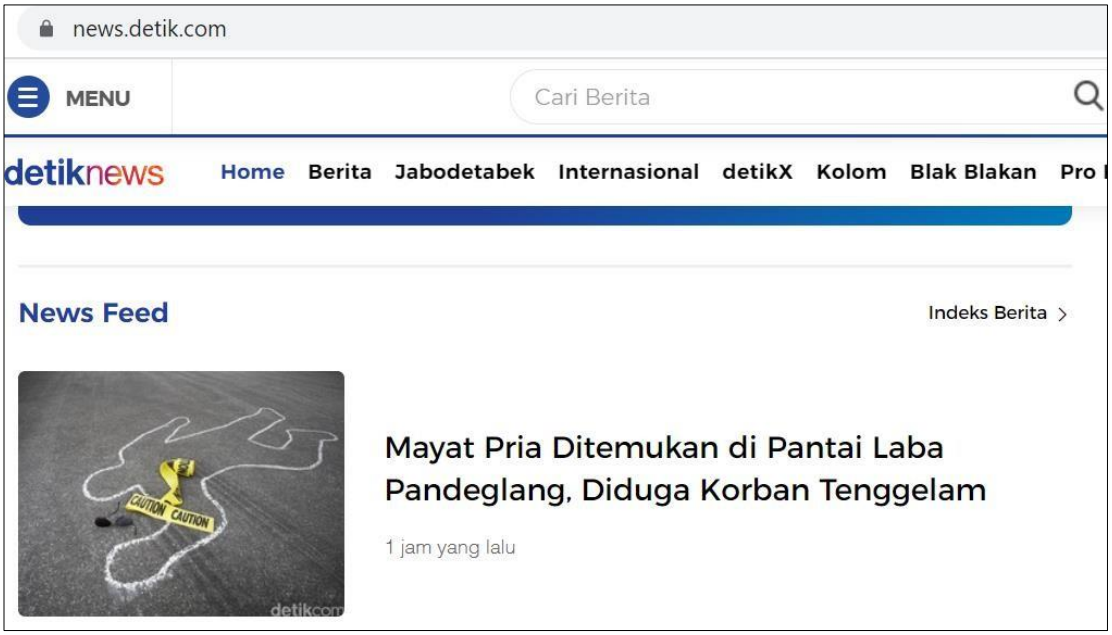
3. Daftar Alat dan Bahan

1. Personal Computer (Min. 4Core Processor, 8GB RAM)
2. Python, MySQL

4. Langkah Kerja

4.1 Menentukan website dan data yang ingin diambil.

Data dapat bersumber dari website umum seperti portal berita detik.com, kompas.com, petamaps.com dan tribunnews.com. Sumber data yang digunakan sangat tergantung dari kebutuhan data yang diinginkan. Pada kasus ini, sebagai contoh yaitu pengambilan data judul berita dari portal news.detik.com.



4.2 Menentukan metode dan teknik yang akan digunakan dalam web scraping

Pada kasus ini metode yang digunakan ialah crapping dengan Regex menggunakan Python. Adapun beberapa library yang dibutuhkan yaitu:

1. urllib (bawaan python): untuk request http
2. BeautifulSoup (> `pip install beautifulsoup4`): untuk parsing
3. mysql.connector (> `pip install mysql-connector-python`): untuk koneksi database

4.3 Membuat program akses web sumber data

1. Import dependency dan melakukan request

```
import urllib.request
from bs4 import BeautifulSoup
import re
import numpy as np

url = 'https://news.detik.com/'

req = urllib.request.Request(url)
resp = urllib.request.urlopen(req)
respData = resp.read()
```

2. Melakukan seleksi komponen html

   Dimisalkan judul berita yang ditargetkan tersimpan dalam kode HTML berikut:

   Data dilihat bahwa judul berada pada posisi yang diapit oleh `<div class="ai_replace_title">` dan `</div>`, Maka seleksi komponen dapat dilakukan dengan cara:

```
…

judul = re.findall(r'<div
class="ai_replace_title">(.*?)</div>',str(respData)
)
```

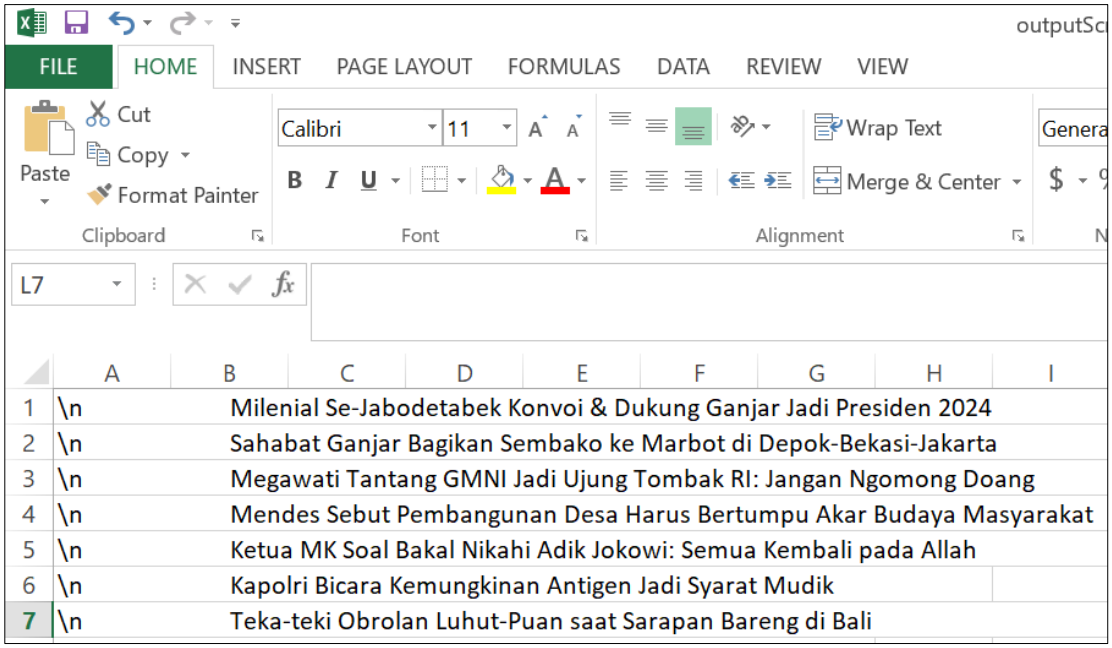
```
print(judul)
```

Regex digunakan untuk mencari dan menyeleksi pattern

3. Menyimpan hasil scrapping dalam data CSV

```
...

for eachP in range(len(judul)):
    print()

    get_judul                                                    =
BeautifulSoup(judul[eachP],"html.parser")
    data_judul = get_judul.get_text()

    print(data_judul)

    form = '"'+data_judul+'"
    output  =  open('outputScrap.csv',"a")  #tinggal
ubah nanti jadi csv
    output.write(str(form))
    output.write('\n')
    output.close()
```

4. Contoh hasil



4.4. Menyimpan data hasil scrapping pada database

```
import csv
import mysql.connector

# mysql connection
sql    =    mysql.connector.connect(host='localhost',
password='', user='root', database='newsdetik')
sqlCursor = sql.cursor()

with open('outputScrap.csv', newline='') as csvfile:
    spamreader = csv.reader(csvfile, delimiter=',',
quotechar='"')
    for row in spamreader:

        judul = row[0]

        sqlQuery   =   "INSERT   INTO   berita(judul)
VALUES('"+judul+"')"

        sqlCursor.execute(sqlQuery)
        sql.commit()

        print(sqlQuery)

sql.close()
```

5. Pertanyaan dan Tugas

1. Tiap mahasiswa mentukan 1 target sumber data yang berbeda
2. Buat scrapping untuk pengambilan data hingga tersimpan dalam database