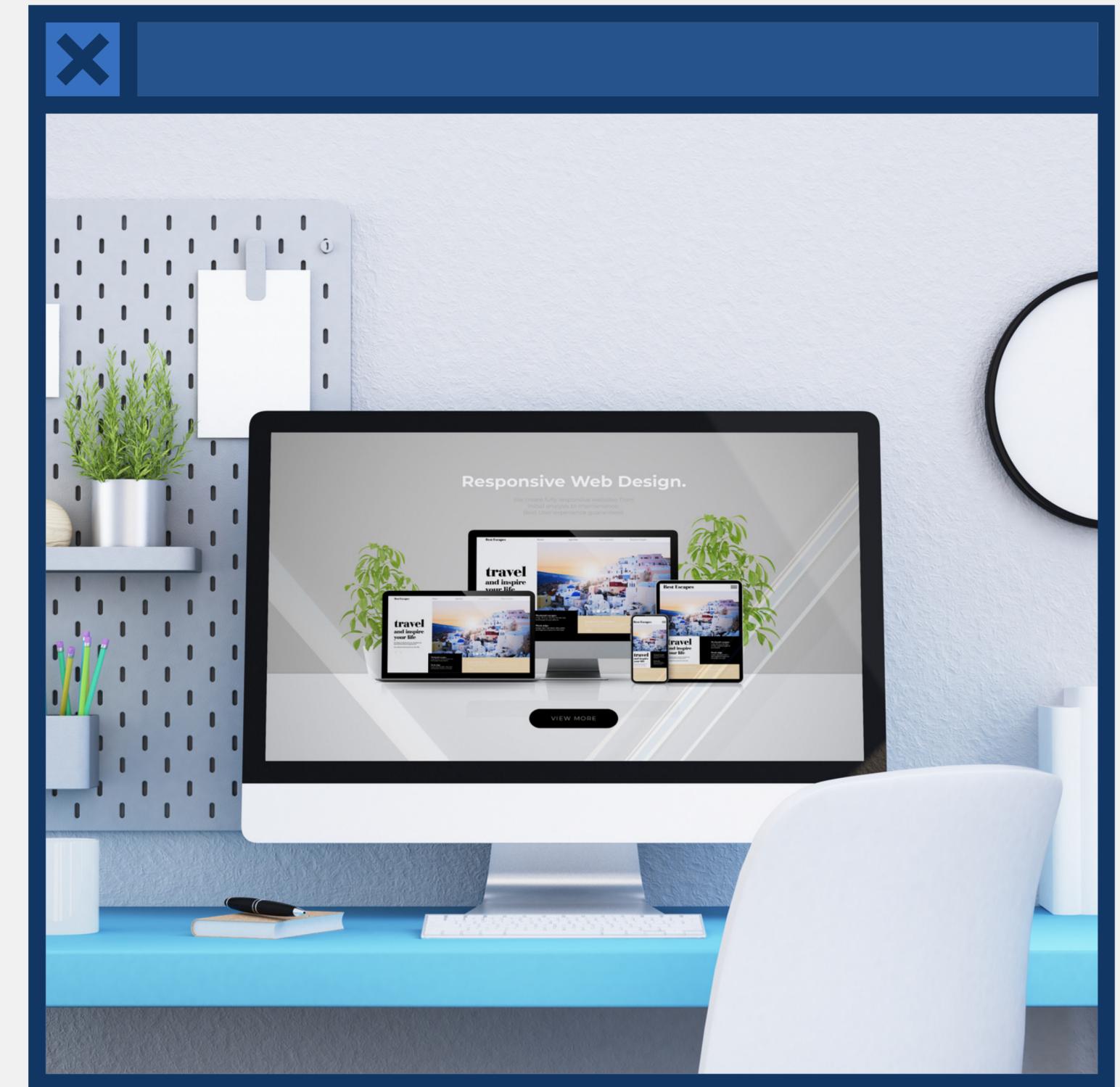


TEXT MINING & WEB SCRAPING

Bagas Wibowo - 2023



<https://github.com/voxeu/TextMining>



Apa itu *Text Mining* dan *Web Scraping*?

Web Scraping adalah teknik ekstraksi data dari berbagai situs web secara otomatis

Text Mining adalah teknik ekstraksi data dan informasi dari sumber data yang berupa teks secara otomatis



Apa Tujuannya?

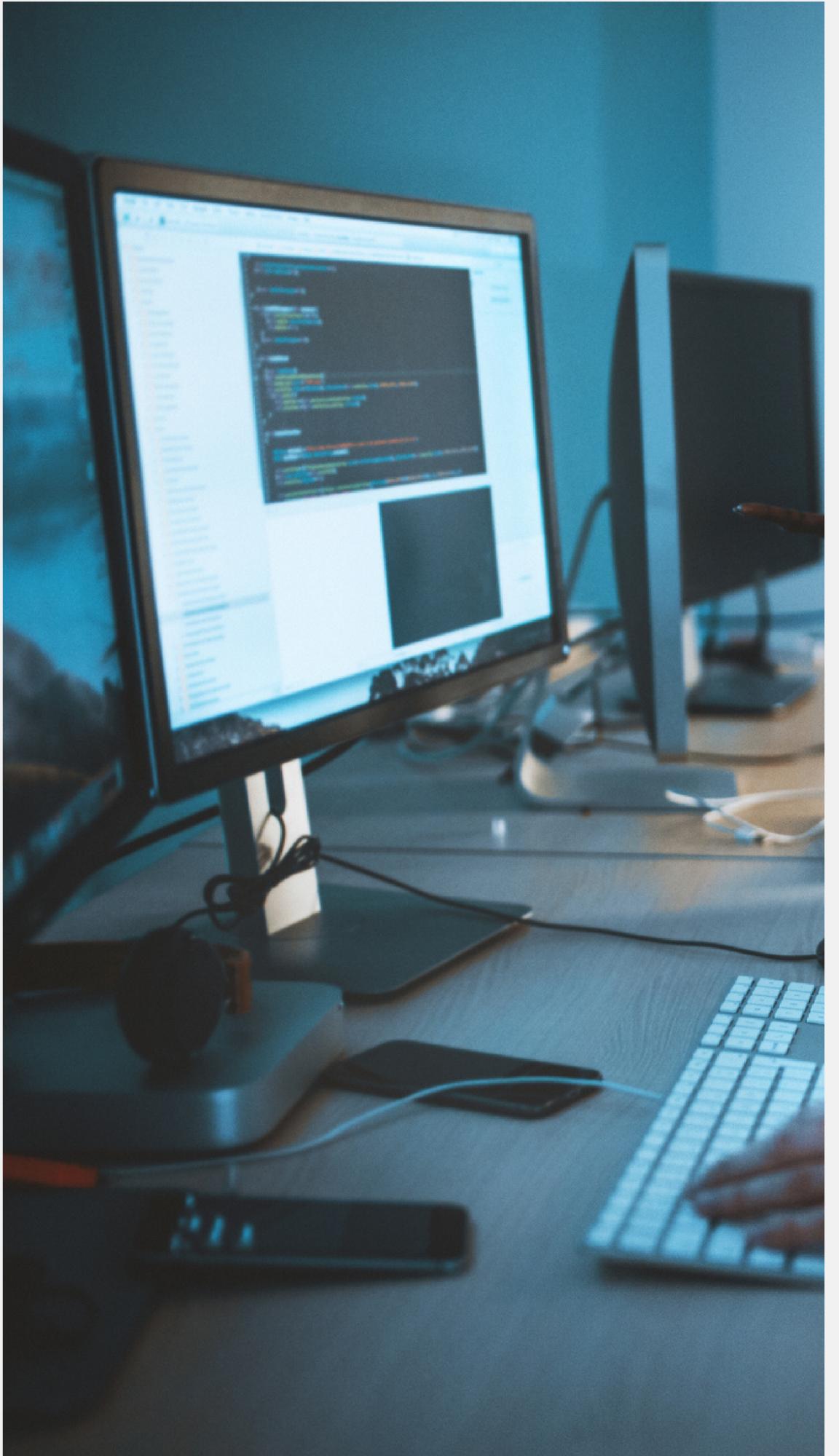
- Ekstraksi dan koleksi Data
- Ekstraksi dan koleksi Informasi
- Menemukan Wawasan dan Pengetahuan baru
- Otomasi





Apakah ada batasan atau limitasi?

- *Web Scraping vs. Web Crawling*
- *Skill vs. Device Specification*
- *Resource Scarcity*
- *Validity of Web Data*
- *Utilitarianism vs. Ethics*



Bagaimana caranya?

01 APLIKASI DAN TOOLS

Melakukan ekstraksi data dan informasi menggunakan aplikasi dan tools seperti WebScrapper dan Microsoft Power BI

02 PROGRAMMING

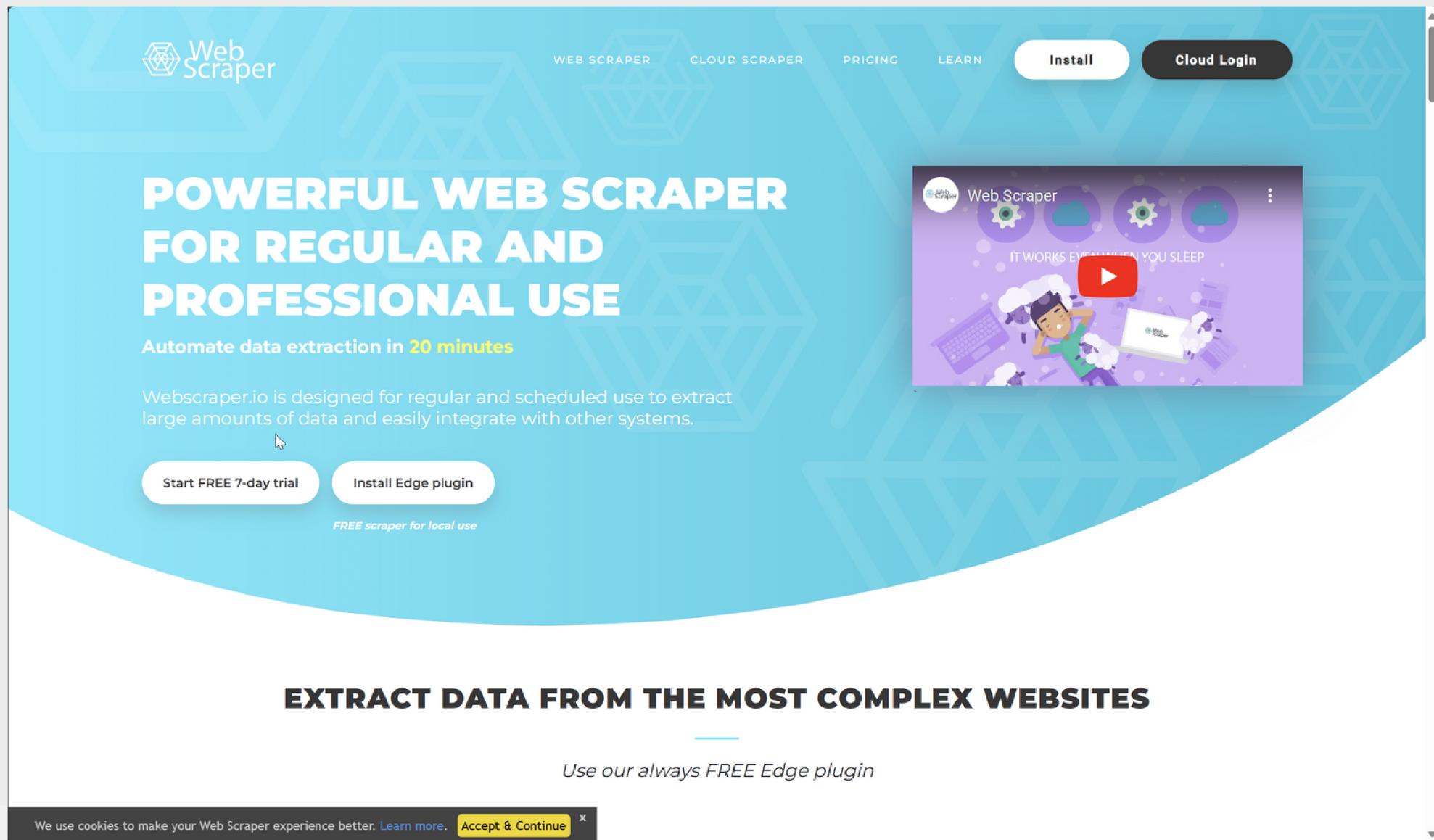
Melakukan ekstraksi data dan informasi menggunakan bahasa pemrograman seperti Python dan platform *Data Science* seperti Anaconda

03 API

Melakukan ekstraksi data dan informasi menggunakan API (*Application Programming Interface*) seperti Instagrapi, HikerAPI Tikapi, dan ChatPDF API.

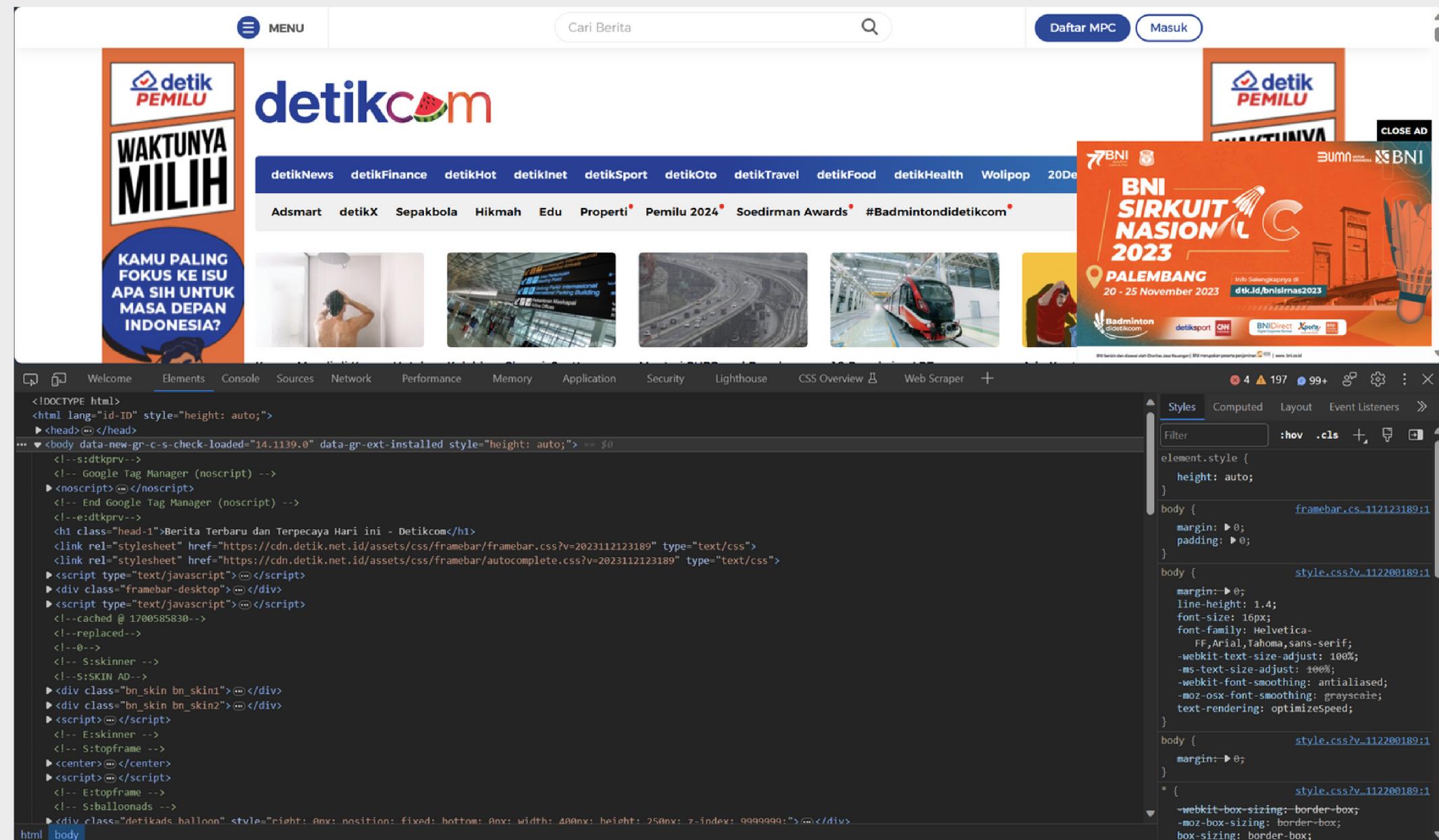
01 WEB SCRAPING: WEBSRAPER

Menggunakan tools extension browser: WebScraping untuk otomasi scrape data.



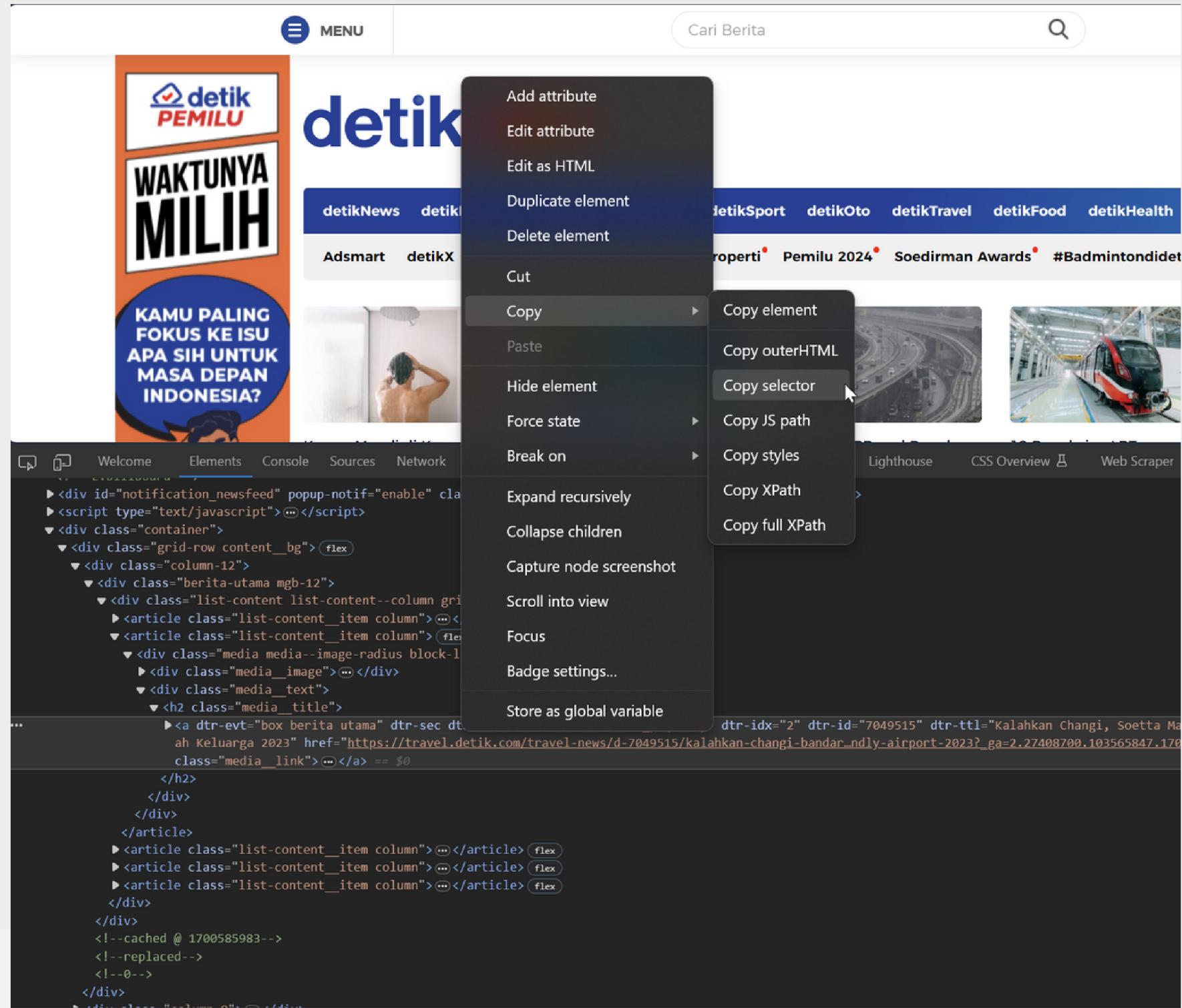
01 WEB SCRAPING: WEBSRAPER

berkenalan dengan
**F12 - INSPECT
ELEMENT**



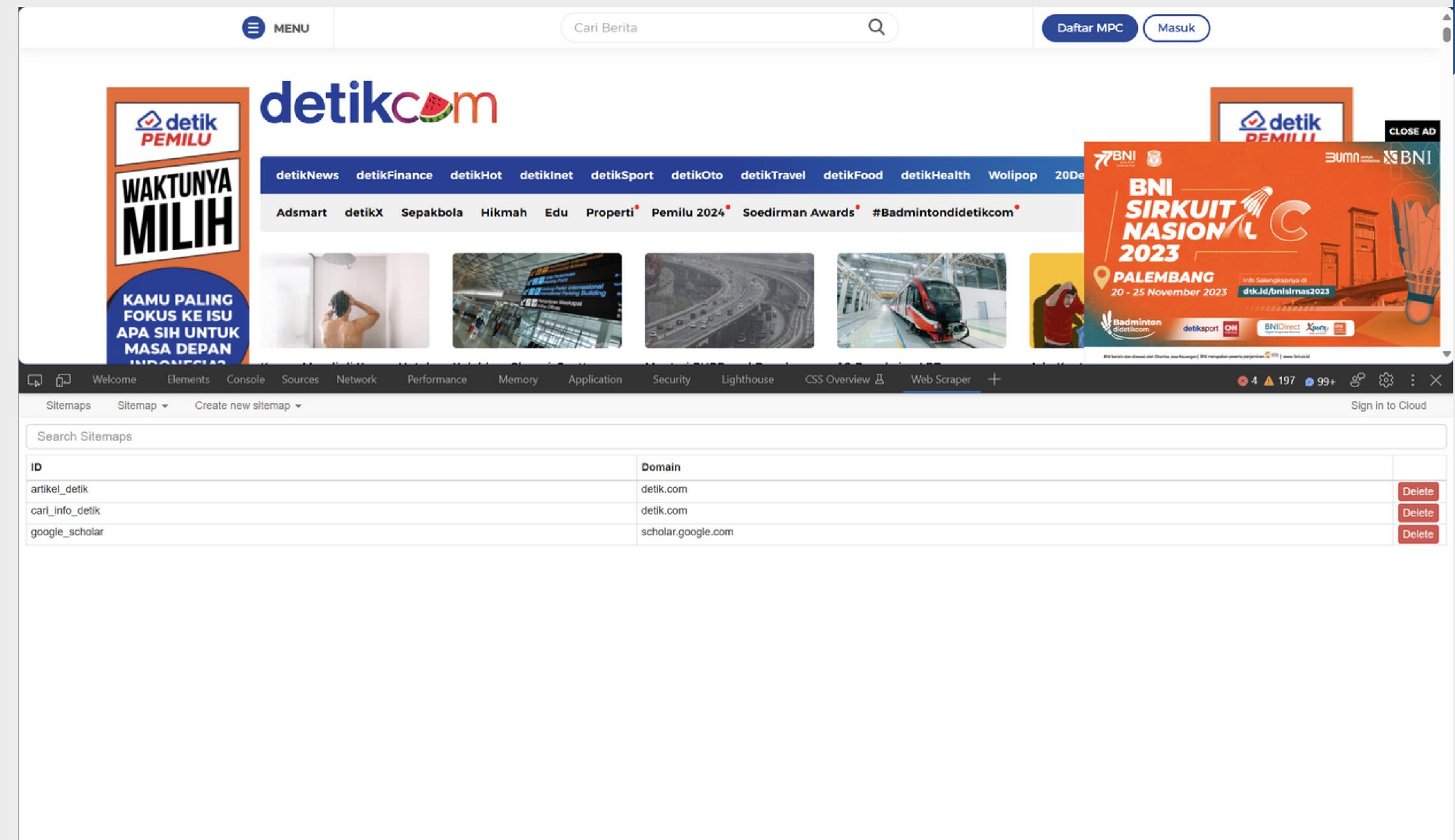
01 WEB SCRAPING: WEBSRAPER

berkenalan dengan
CSS SELECTOR



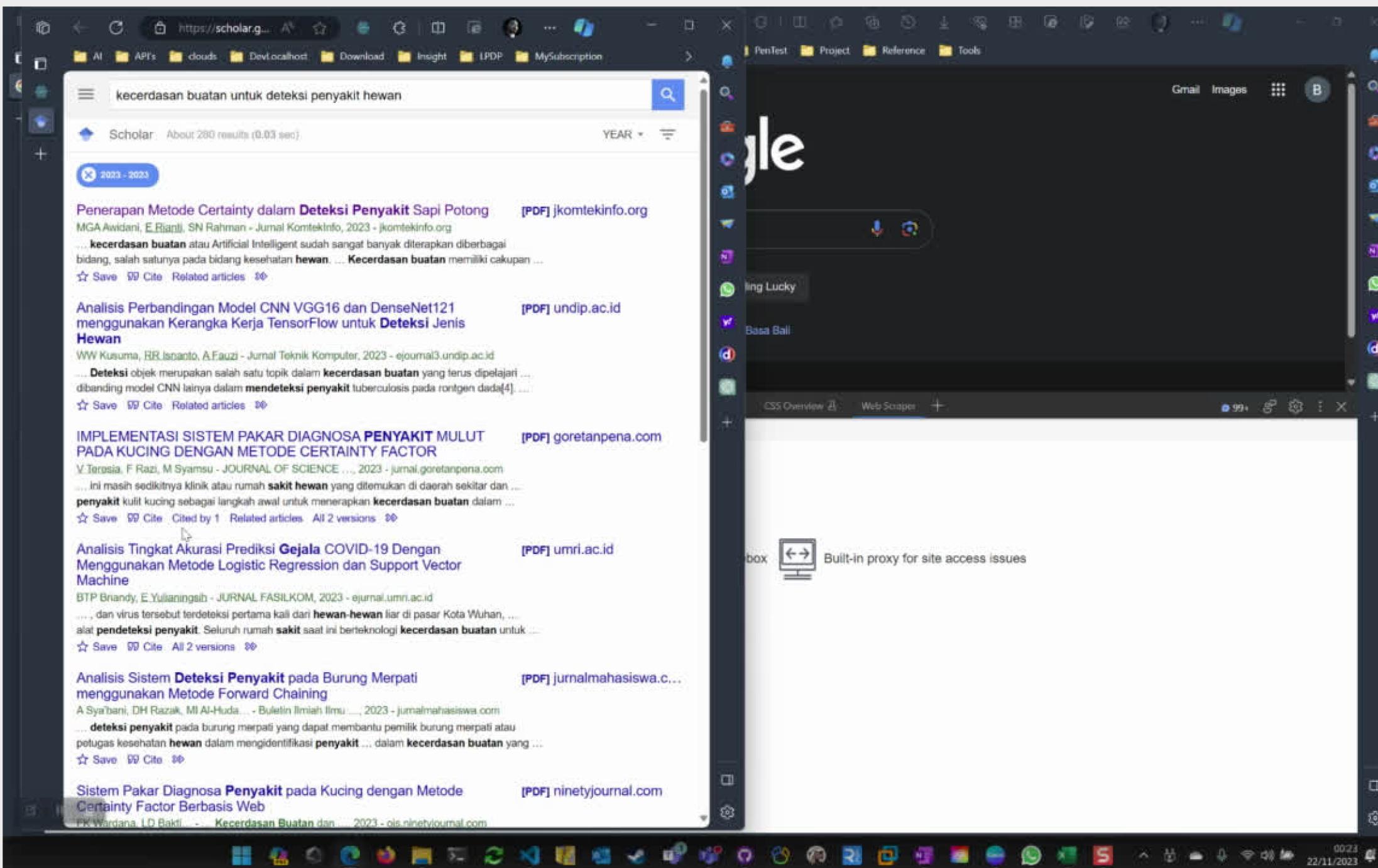
01 WEB SCRAPING: WEBSRAPER

berkenalan dengan
WEBSRAPER



01 WEB SCRAPING: WEBSRAPER

otomasi dengan
WEBSRAPER

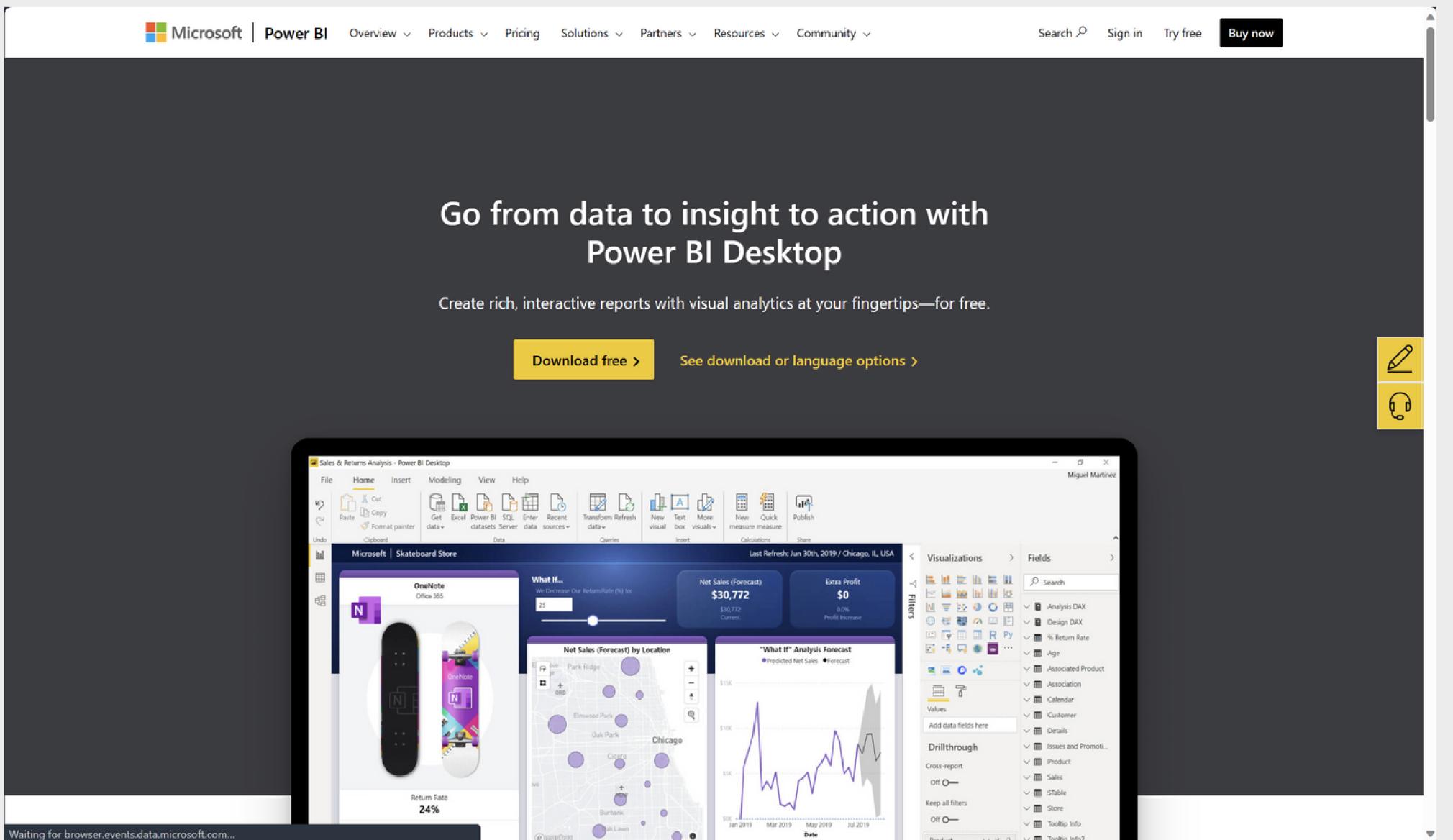


01 WEB SCRAPING: WEBSRAPER

powerful untuk otomasi scrape data publik
(tingkat kompleksitas dan kustomisasi terbatas)

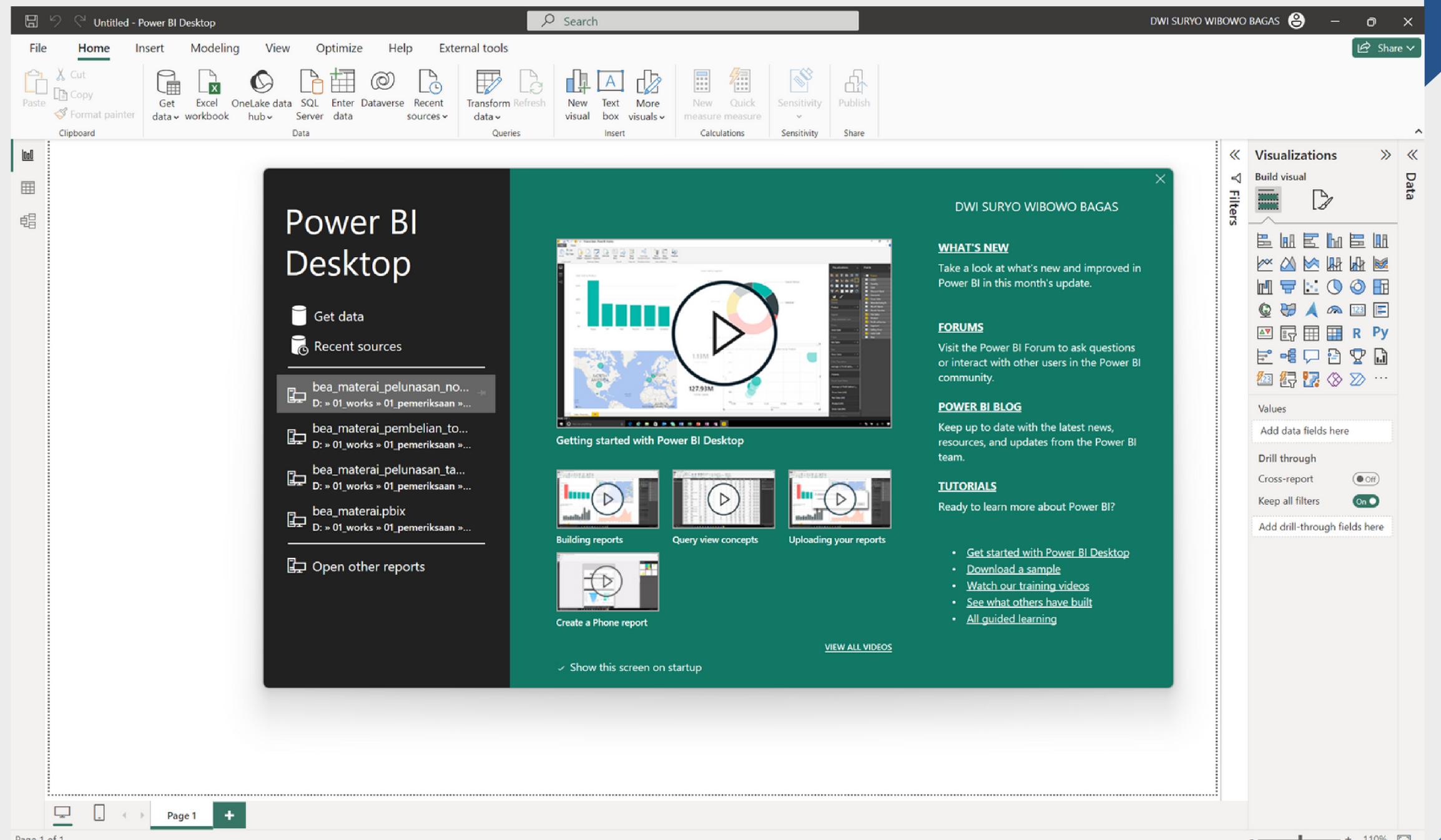
02 WEB SCRAPING: POWER BI

Menggunakan tools aplikasi: Power BI Desktop untuk otomasi scrape data.



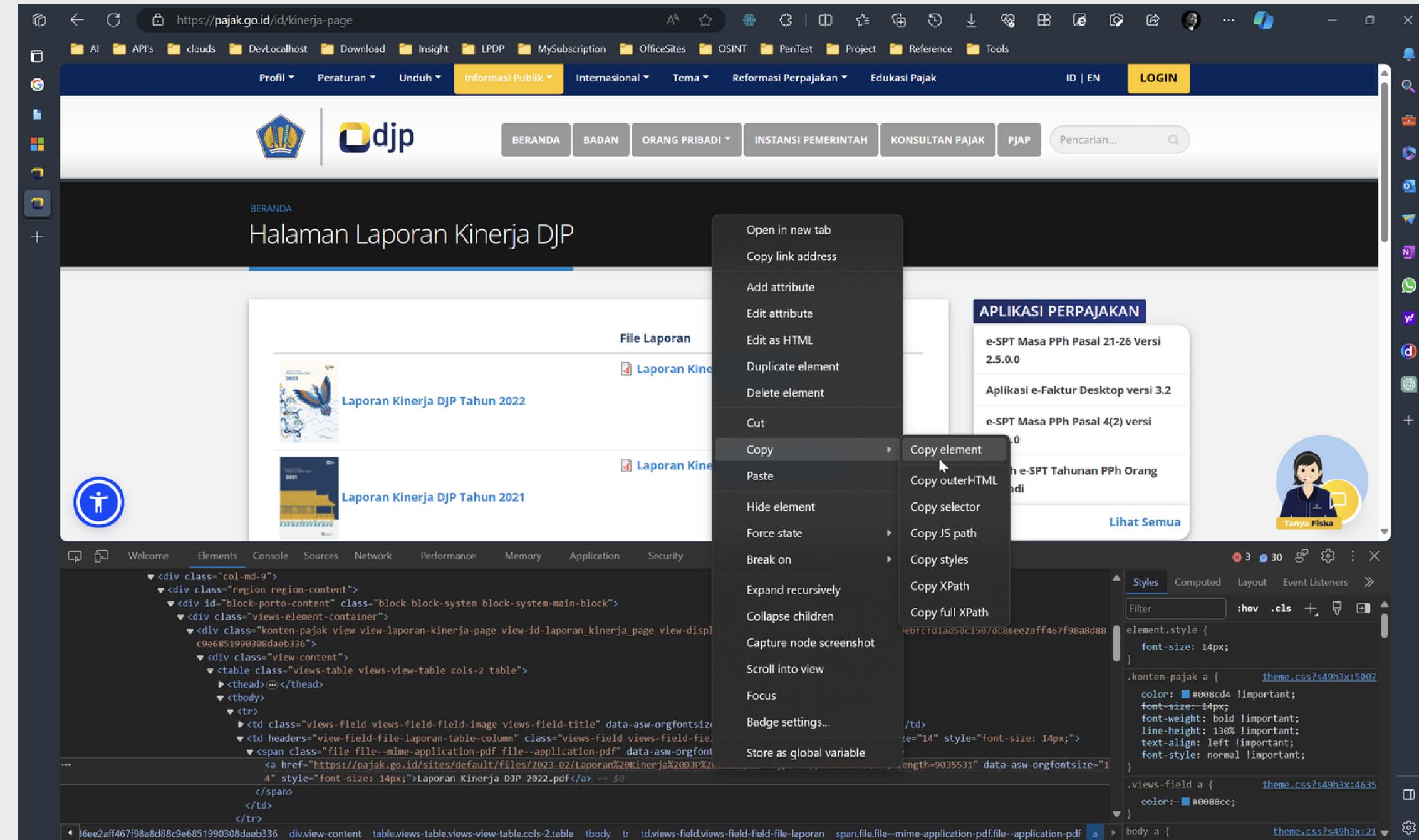
02 WEB SCRAPING: POWER BI

berkenalan dengan
Power BI Desktop



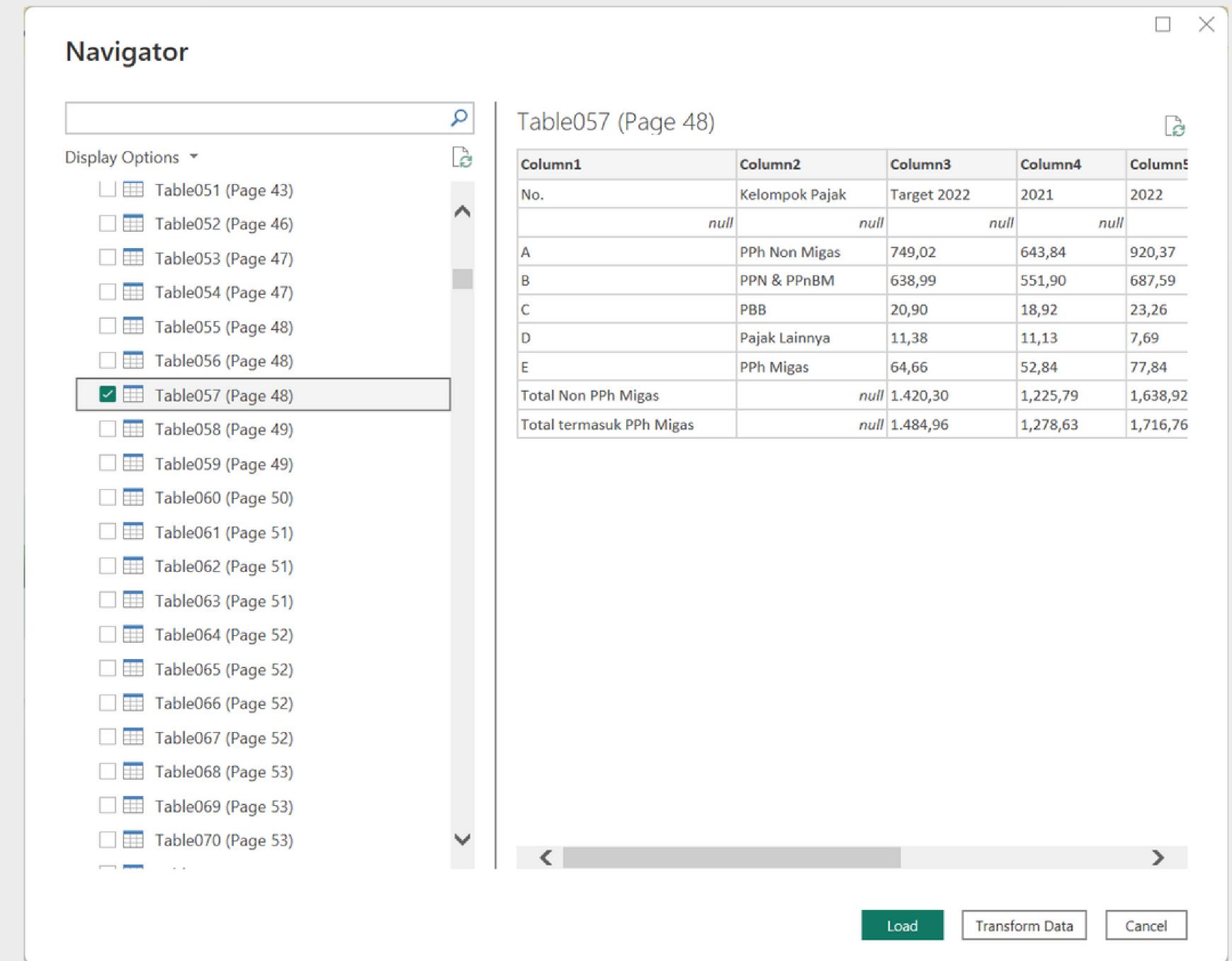
02 WEB SCRAPING: POWER BI

scrape PDF dari Web
dengan
Power BI Desktop



02 WEB SCRAPING: POWER BI

scrape PDF dari Web
dengan
Power BI Desktop



The screenshot shows the Power BI Desktop Navigator interface. On the left, a list of tables from a PDF document is displayed, with 'Table057 (Page 48)' selected. On the right, the contents of 'Table057 (Page 48)' are shown as a tabular dataset.

Column1	Column2	Column3	Column4	Column5
No.	Kelompok Pajak	Target 2022	2021	2022
A	PPh Non Migas	749,02	643,84	920,37
B	PPN & PPnBM	638,99	551,90	687,59
C	PBB	20,90	18,92	23,26
D	Pajak Lainnya	11,38	11,13	7,69
E	PPh Migas	64,66	52,84	77,84
Total Non PPh Migas		null	1,420,30	1,225,79
Total termasuk PPh Migas		null	1,484,96	1,278,63
				1,716,76

At the bottom of the Navigator window, there are three buttons: 'Load' (green), 'Transform Data' (white), and 'Cancel' (white).

02 WEB SCRAPING: POWER BI

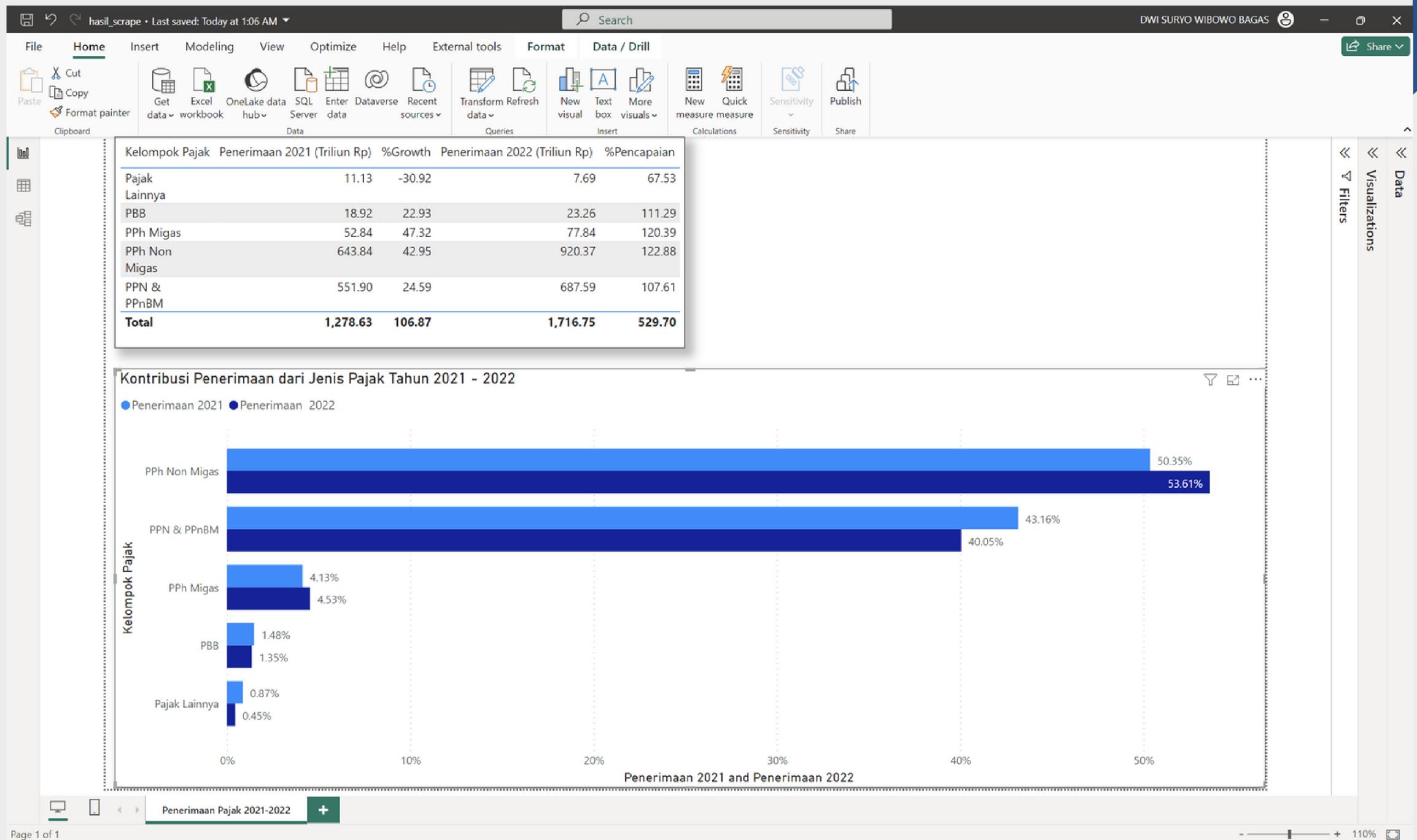
ETL dengan
Power BI Desktop

The screenshot shows the Power BI Desktop interface with the Power Query Editor open. The main area displays a table titled "Target dan Realisasi Penerimaan Pajak" with 5 rows and 9 columns. The columns are labeled: No., Kelompok Pajak, 1.2 Target 2022, 1.2 2021, 1.2 2022, 1.2 % Growth, 1.2 % Growth_1, and 1.2 %. The data includes various tax categories like PPh Non Migas, PPN & PPnBM, PBB, Pajak Lainnya, and PPh Migas, along with their respective target and actual values for the years 2022, 2021, and 2022, and growth percentages. Above the table, a formula bar shows the query transformation: `= Table.TransformColumnTypes(#"Filtered Rows",{{"Target#(lf)2022", type number}, {"2021", type number}, {"2022", type number}, {"%(lf)Growth", type number}})`. To the right of the table, the "APPLIED STEPS" pane is visible, showing the sequence of steps taken: "Source", "Navigation", "Promoted Headers", "Filtered Rows", and "Changed Type". The "Changed Type" step is currently selected.

No.	Kelompok Pajak	1.2 Target 2022	1.2 2021	1.2 2022	1.2 % Growth	1.2 % Growth_1	1.2 %
1	PPh Non Migas	749.02	643.84	920.37	14.76	42.95	
2	PPN & PPnBM	638.99	551.9	687.59	22.56	24.59	
3	PBB	20.9	18.92	23.26	-9.68	22.93	
4	Pajak Lainnya	11.38	11.13	7.69	63.84	-30.92	
5	PPh Migas	64.66	52.84	77.84	59.99	47.32	

02 WEB SCRAPING: POWER BI

Visualisasi dengan
Power BI Desktop



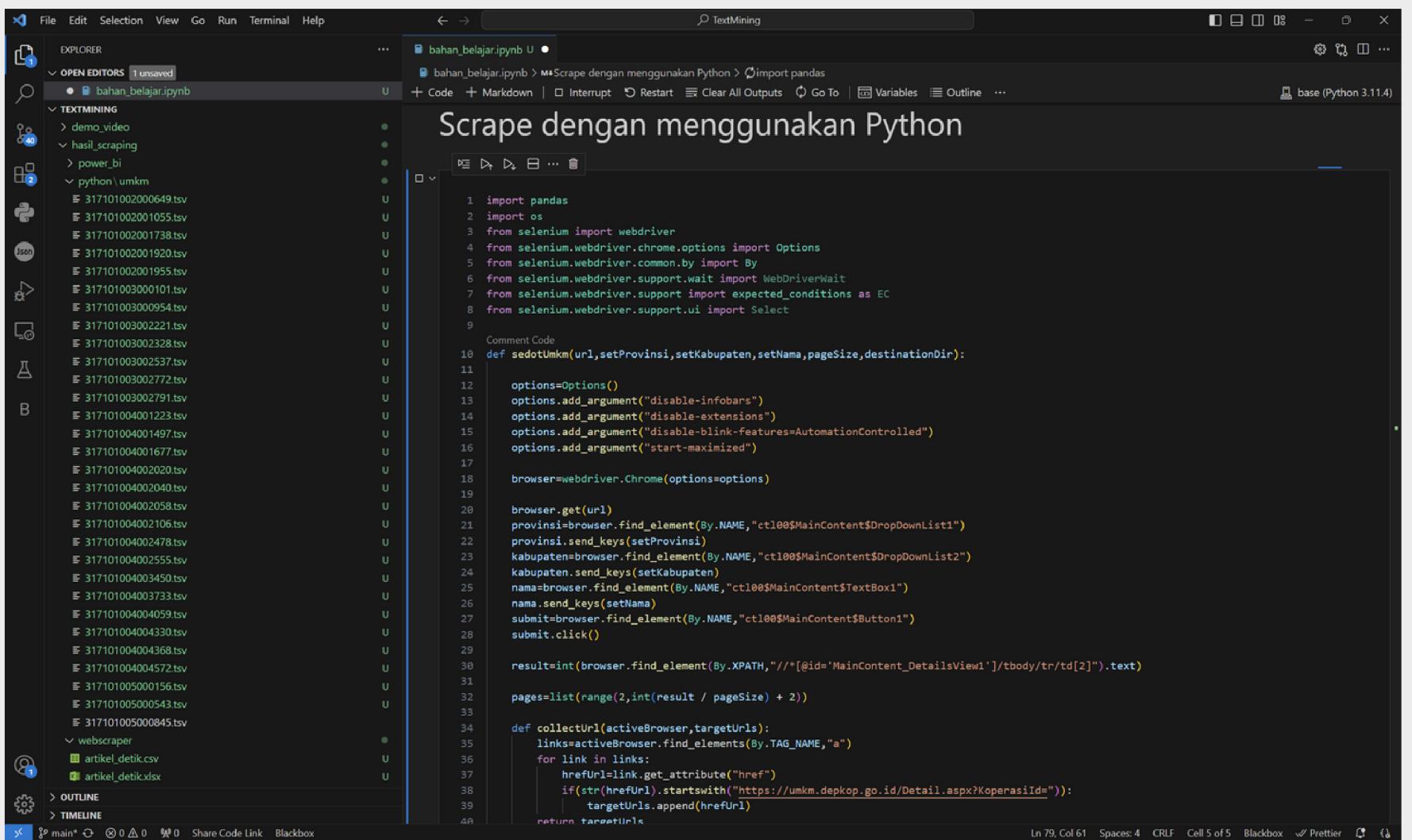
02 WEB SCRAPING: POWER BI

**powerful untuk otomasi ETL dan Visualisasi Big Data
dengan format tabular**

*(terbukti untuk pengolahan 1.2 Milyar Baris Data
tetapi haus resources)*

03 WEB SCRAPING: PYTHON

Menggunakan bahasa pemrograman Python dengan library Selenium dan Pandas untuk otomasi scrape data.



The screenshot shows a Jupyter Notebook interface with the title "Scrape dengan menggunakan Python". The code cell contains the following Python script:

```
import pandas
import os
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.common.by import By
from selenium.webdriver.support import wait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.support.ui import Select

Comment Code
def sedotUmkm(url,setProvinsi,setKabupaten,setName,pageSize,destinationDir):
    options=Options()
    options.add_argument("disable-infobars")
    options.add_argument("disable-extensions")
    options.add_argument("disable-blink-features=AutomationControlled")
    options.add_argument("start-maximized")

    browser=webdriver.Chrome(options=options)

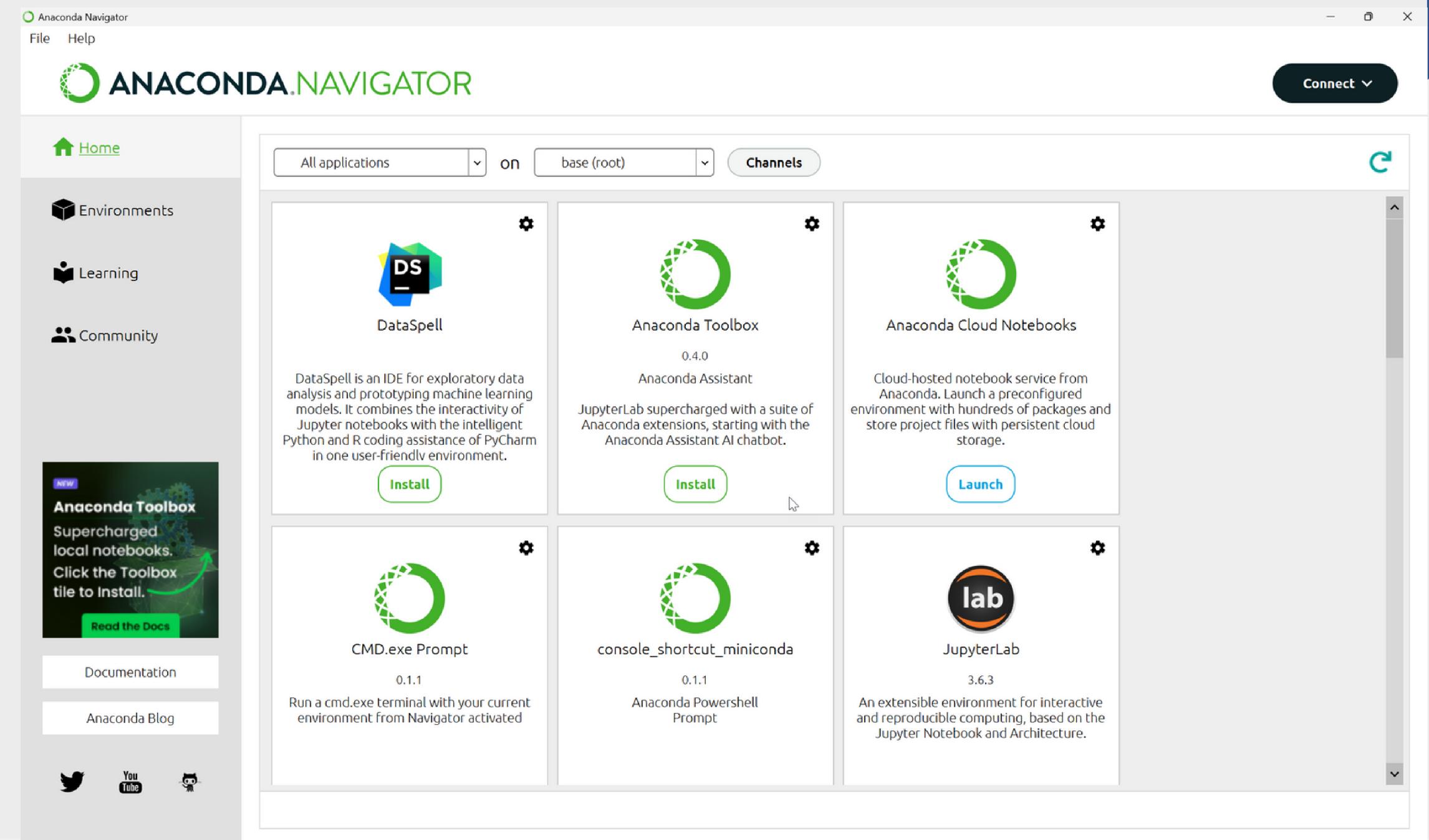
    browser.get(url)
    provinsi=browser.find_element(By.NAME,"ctl00$MainContent$DropDownList1")
    provinsi.send_keys(setProvinsi)
    kabupaten=browser.find_element(By.NAME,"ctl00$MainContent$DropDownList2")
    kabupaten.send_keys(setKabupaten)
    nama=browser.find_element(By.NAME,"ctl00$MainContent$TextBox1")
    nama.send_keys(setName)
    submit=browser.find_element(By.NAME,"ctl00$MainContent$Button1")
    submit.click()

    result=int(browser.find_element(By.XPATH,"//*[@id='MainContent_DetailsView1']/tbody/tr/td[2]").text)
    pages=list(range(2,int(result / pageSize) + 2))

    def collectUrl(activeBrowser,targetUrls):
        links=activeBrowser.find_elements(By.TAG_NAME,"a")
        for link in links:
            hrefurl=link.get_attribute("href")
            if(str(hrefurl).startswith("https://umkm.depker.go.id/Detail.aspx?KoperasiId=")):
                targetUrls.append(hrefurl)
    return targetUrls
```

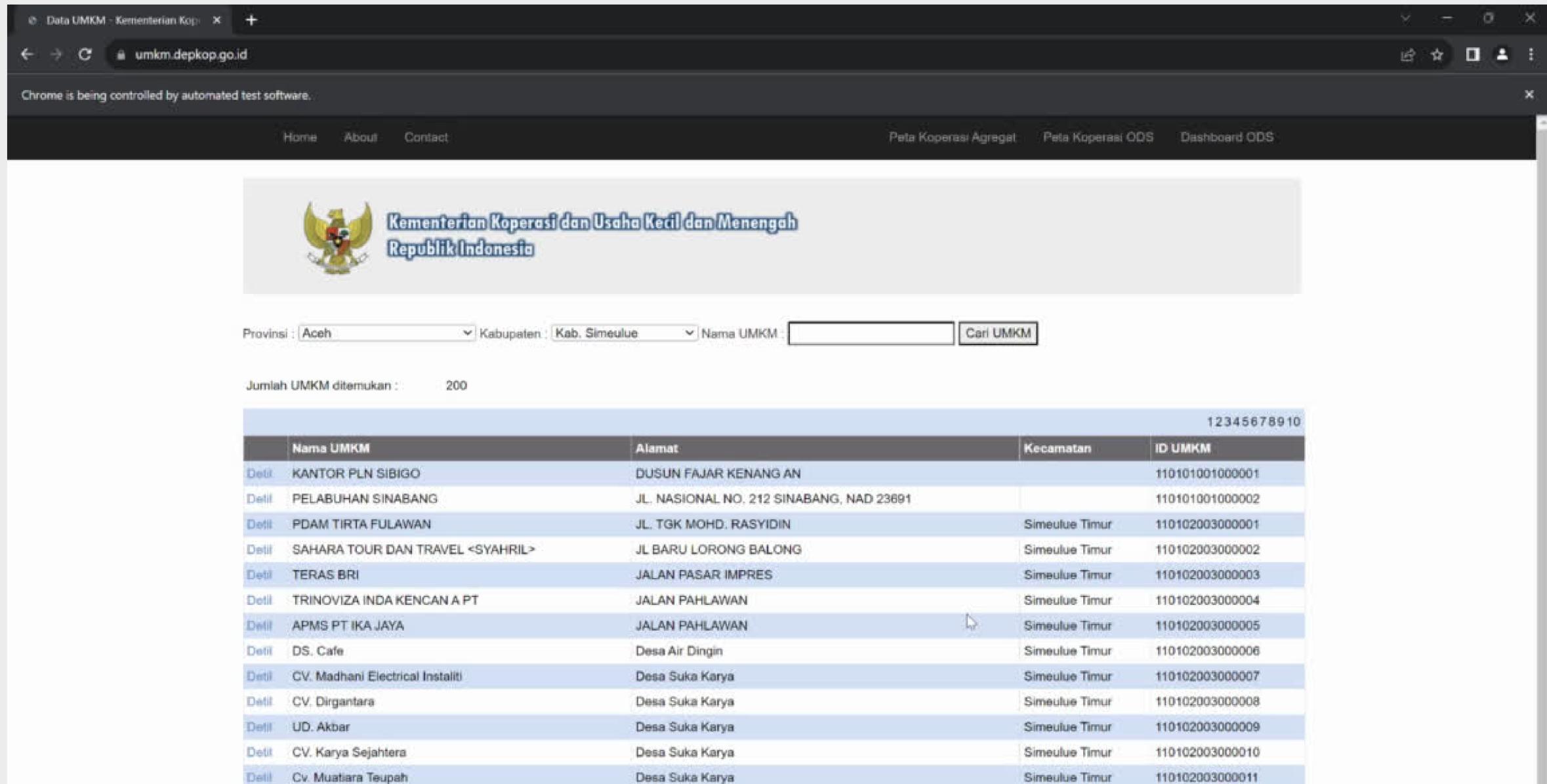
03 WEB SCRAPING: PYTHON

berkenalan dengan
Python dan Anaconda



03 WEB SCRAPING: PYTHON

otomasi dengan
PYTHON



The screenshot shows a web browser window with the title "Data UMKM - Kementerian Koperasi dan Usaha Kecil dan Menengah Republik Indonesia". The URL in the address bar is "umkm.depkop.go.id". A message at the top of the page reads "Chrome is being controlled by automated test software.". The page header includes links for "Home", "About", "Contact", "Peta Koperasi Agregat", "Peta Koperasi ODS", and "Dashboard ODS". Below the header, there is a logo of the Indonesian Garuda and the text "Kementerian Koperasi dan Usaha Kecil dan Menengah Republik Indonesia". A search form allows users to filter by "Provinsi" (Aceh), "Kabupaten" (Kab. Simeulue), and "Nama UMKM", with a "Cari UMKM" button. A message below the form states "Jumlah UMKM ditemukan : 200". A table displays 20 rows of UMKM data, each with columns for "Nama UMKM", "Alamat", "Kecamatan", and "ID UMKM". The data is as follows:

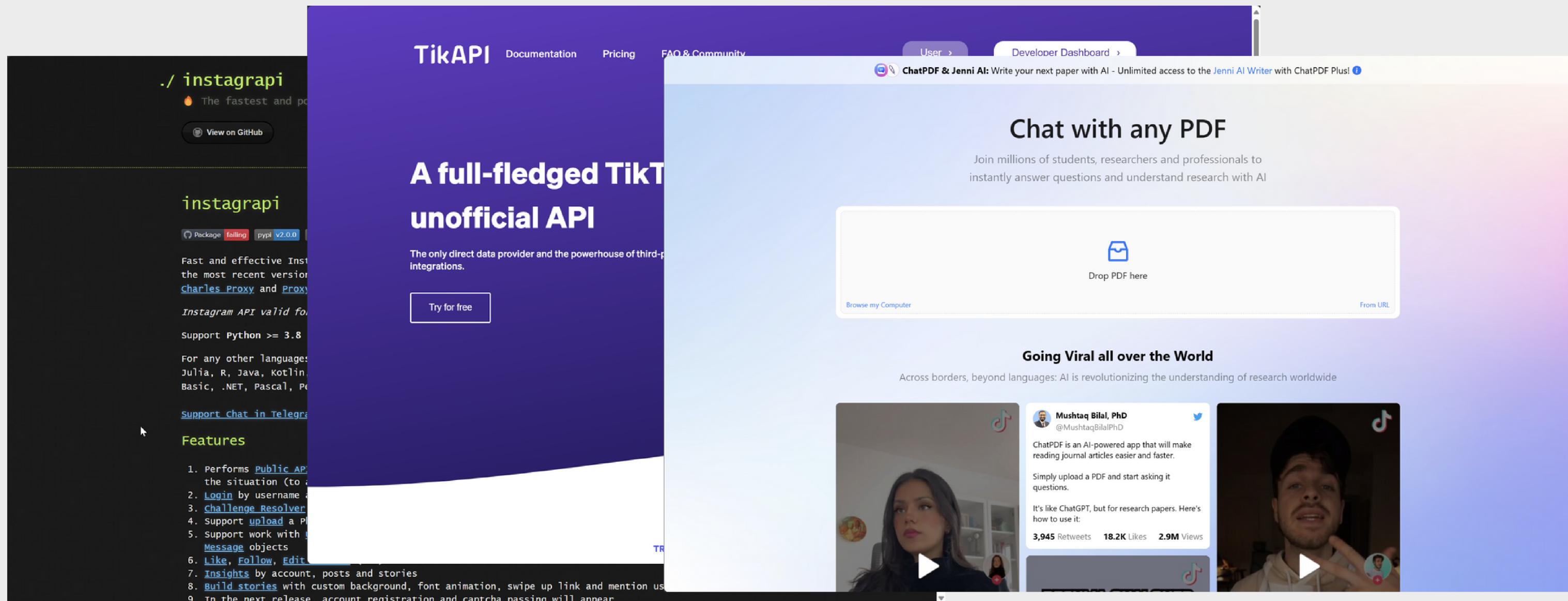
	Nama UMKM	Alamat	Kecamatan	ID UMKM
Detil	KANTOR PLN SIBIGO	DUSUN FAJAR KENANG AN		110101001000001
Detil	PELABUHAN SINABANG	JL. NASIONAL NO. 212 SINABANG, NAD 23691		110101001000002
Detil	PDAM TIRTA FULAWAN	JL. TGK MOHD. RASYIDIN	Simeulue Timur	110102003000001
Detil	SAHARA TOUR DAN TRAVEL <SYAHRIL>	JL BARU LORONG BALONG	Simeulue Timur	110102003000002
Detil	TERAS BRI	JALAN PASAR IMPRES	Simeulue Timur	110102003000003
Detil	TRINOVIZA INDIA KENCANA PT	JALAN PAHLAWAN	Simeulue Timur	110102003000004
Detil	APMS PT IKA JAYA	JALAN PAHLAWAN	Simeulue Timur	110102003000005
Detil	DS. Cafe	Desa Air Dingin	Simeulue Timur	110102003000006
Detil	CV. Madhani Electrical Instalasi	Desa Suka Karya	Simeulue Timur	110102003000007
Detil	CV. Dirgantara	Desa Suka Karya	Simeulue Timur	110102003000008
Detil	UD. Akbar	Desa Suka Karya	Simeulue Timur	110102003000009
Detil	CV. Karya Sejahtera	Desa Suka Karya	Simeulue Timur	110102003000010
Detil	Cv. Muatiara Teupah	Desa Suka Karya	Simeulue Timur	110102003000011

03 WEB SCRAPING: PYTHON

powerful untuk apapun

04 WEB SCRAPING: API

Menggunakan Application Programming Interface untuk otomasi scrape data.



04 WEB SCRAPING: API

powerful untuk production dan tujuan komersial

05 EXTRACT DATA: CSV DAN PDF

mengekstrak data
(teks) dari file **CSV**
dan **PDF**

The screenshot shows a Jupyter Notebook interface titled "TextMining". The notebook has one open cell named "bahan_belajar.ipynb". The code in the first cell is for extracting data from a CSV file:

```
1 import pandas as pd
2
3 # Muat dataset dari file CSV dengan menggunakan Pandas
4 file_sumber="./hasil_scraping/webscraper/artikel_detik.csv"
5 csvDataFrame=pd.read_csv(file_sumber)
6 csvDataFrame
```

The code in the second cell is for extracting data from a PDF file:

```
1 kontenDataFrame=csvDataFrame["konten"]
2 kontenDataFrame
```

The code in the third cell is for printing the extracted content:

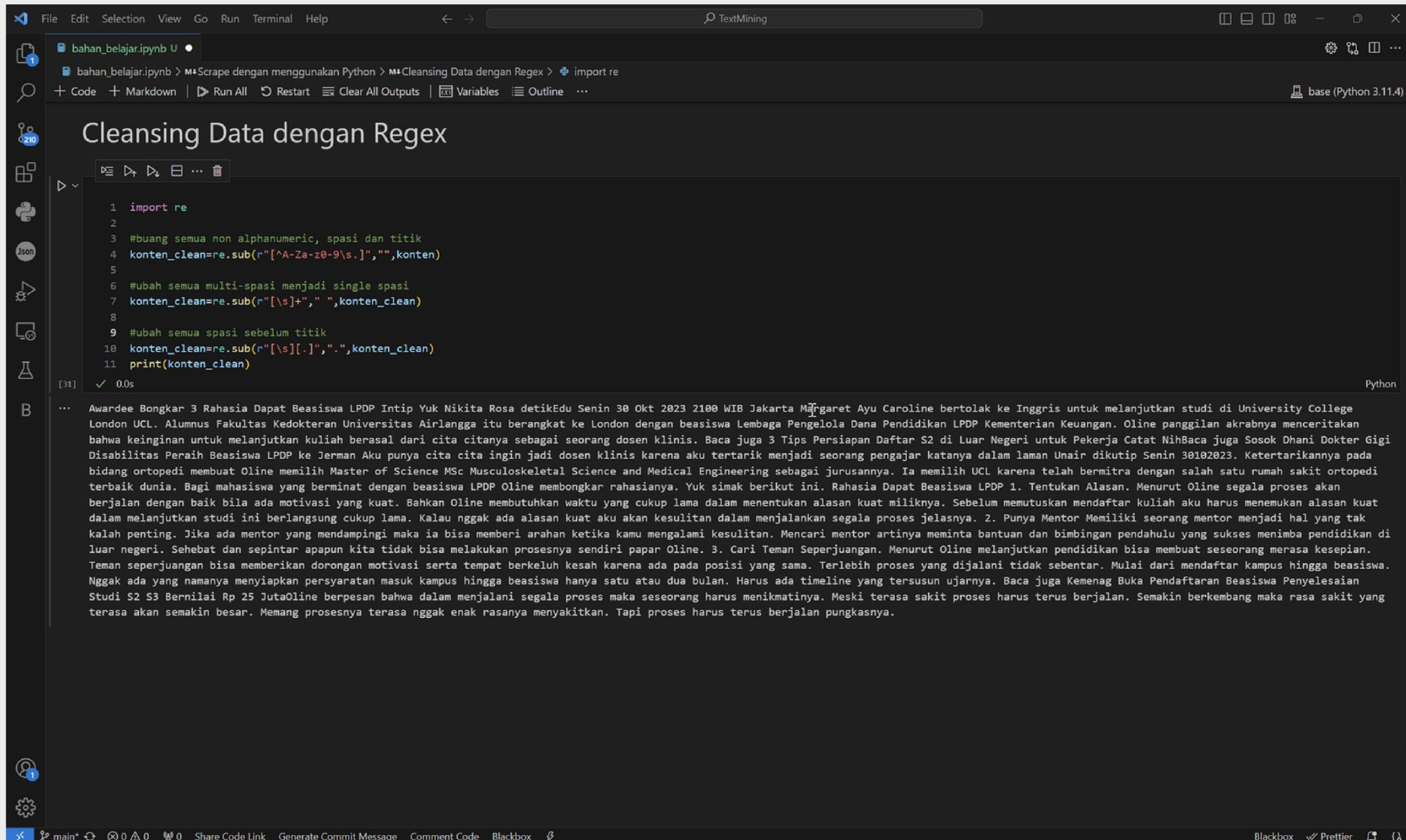
```
1 konten=kontenDataFrame[0]
2 print(konten)
```

The fourth cell is for extracting data from a PDF file:

```
1 from pypdf import PdfReader
2
3 reader = PdfReader("./sumber_data/Awardee Bongkar 3 Rahasia Dapat Beasiswa LPDP.pdf")
4 konten = ""
5 for page in reader.pages:
6     konten += page.extract_text() + "\n"
7 print(konten)
```

06 CLEANSING DATA: REGEX

membersihkan data
dari karakter yang
tidak akan digunakan



The screenshot shows a Jupyter Notebook interface with a dark theme. The title bar reads "File Edit Selection View Go Run Terminal Help" and "TextMining". The notebook file is "bahan_belajar.ipynb". The code cell contains the following Python script:

```
1 import re
2
3 #buang semua non alphanumeric, spasi dan titik
4 konten_clean=re.sub(r"[^A-Za-z0-9\s]","",konten)
5
6 #ubah semua multi-spasi menjadi single spasi
7 konten_clean=re.sub(r"\s+"," ",konten_clean)
8
9 #ubah semua spasi sebelum titik
10 konten_clean=re.sub(r"[\s][.]", ". ",konten_clean)
11 print(konten_clean)
```

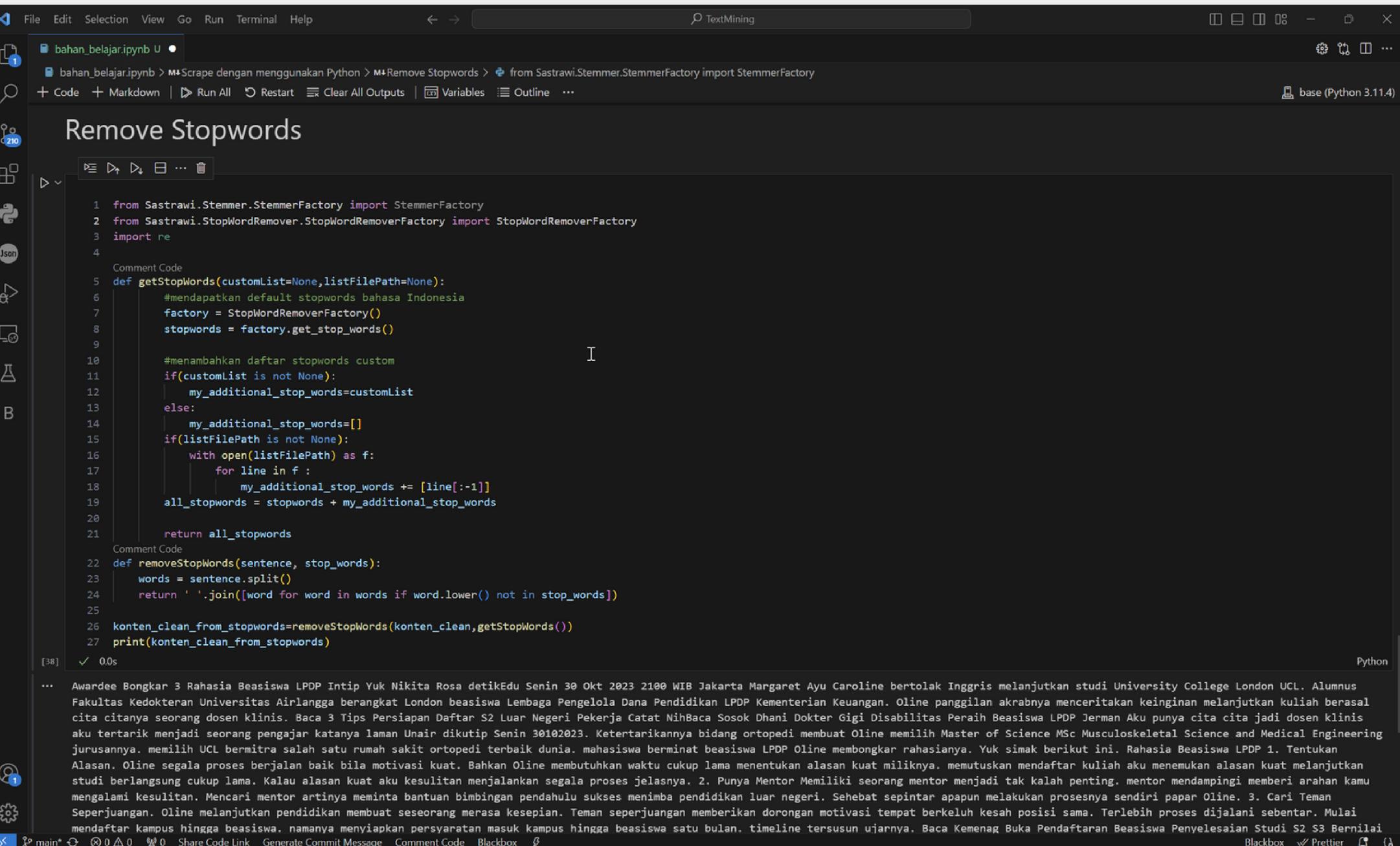
The output cell displays the cleaned text:

```
[31]: ✓ 0.0s
...
Awardee Bongkar 3 Rahasia Dapat Beasiswa LPDP Intip Yuk Nikita Rosa detikEdu Senin 30 Okt 2023 2100 WIB Jakarta Margaret Ayu Caroline bertolak ke Inggris untuk melanjutkan studi di University College London UCL. Alumnus Fakultas Kedokteran Universitas Airlangga itu berangkat ke London dengan beasiswa Lembaga Pengelola Dana Pendidikan LPDP Kementerian Keuangan. Online panggilan akrabnya menceritakan bahwa keinginan untuk melanjutkan kuliah berasal dari cita citanya sebagai seorang dosen klinis. Baca juga 3 Tips Persiapan Daftar S2 di Luar Negeri untuk Pekerja Catat NIH Baca juga Sosok Dhani Dokter Gigi Disabilitas Peraih Beasiswa LPDP ke Jerman Aku punya cita cita ingin jadi dosen klinis karena aku tertarik menjadi seorang pengajar katanya dalam laman Unair dikutip Senin 30102023. Ketertarikannya pada bidang ortopedi membuat Online memilih Master of Science MSc Musculoskeletal Science and Medical Engineering sebagai jurusannya. Ia memilih UCL karena telah bermitra dengan salah satu rumah sakit ortopedi terbaik dunia. Bagi mahasiswa yang berminat dengan beasiswa LPDP Online membongkar rahasianya. Yuk simak berikut ini. Rahasia Dapat Beasiswa LPDP 1. Tentukan Alasan. Menurut Online segala proses akan berjalan dengan baik bila ada motivasi yang kuat. Bahkan Online membutuhkan waktu yang cukup lama dalam menentukan alasan kuat miliknya. Sebelum memutuskan mendaftar kuliah aku harus menemukan alasan kuat dalam melanjutkan studi ini berlangsung cukup lama. Kalau nggak ada alasan kuat aku akan kesulitan dalam menjalankan segala proses jelasnya. 2. Punya Mentor Memiliki seorang mentor menjadi hal yang tak kalah penting. Jika ada mentor yang mendampingi maka ia bisa memberi arahan ketika kamu mengalami kesulitan. Mencari mentor artinya meminta bantuan dan bimbingan pendahulu yang sukses menimba pendidikan di luar negeri. Sehebat dan sepihak apapun kita tidak bisa melakukan prosesnya sendiri papar Online. 3. Cari Teman Seperjuangan. Menurut Online melanjutkan pendidikan bisa membuat seseorang merasa kesepian. Teman seperjuangan bisa memberikan dorongan motivasi serta tempat berkeluh kesah karena ada pada posisi yang sama. Terlebih proses yang dijalani tidak sebentar. Mulai dari mendaftar kampus hingga beasiswa. Nggak ada yang namanya siapkan persyaratan masuk kampus hingga beasiswa hanya satu atau dua bulan. Harus ada timeline yang tersusun ujarnya. Baca juga Kemenag Buka Pendaftaran Beasiswa Studi S2 S3 Bernilai Rp 25 Juta Online berpesan bahwa dalam menjalani segala proses maka seseorang harus menikmatinya. Meski terasa sakit proses harus terus berjalan. Semakin berkembang maka rasa sakit yang terasa akan semakin besar. Memang prosesnya terasa nggak enak rasanya menyakitkan. Tapi proses harus terus berjalan pungkasnya.
```

At the bottom, there are various notebook controls like "main*", "Share Code Link", "Generate Commit Message", "Comment Code", "Blackbox", and "Prettier".

07 CLEANSING DATA: STOPWORDS

membersihkan data
dari kata berulang
yang tidak signifikan
dalam pengolahan

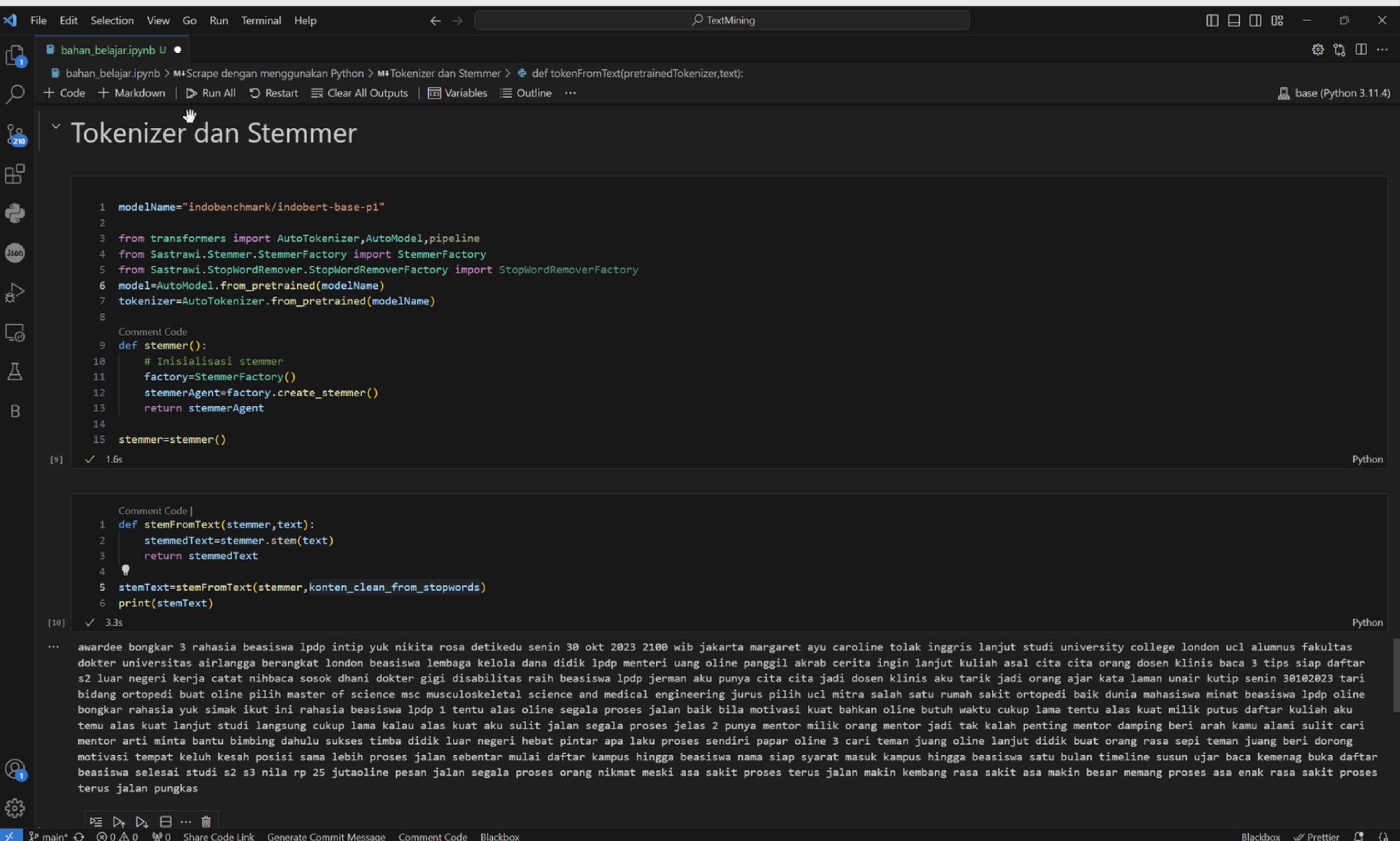


The screenshot shows a Jupyter Notebook interface with a dark theme. The title bar reads "TextMining". The left sidebar lists files: "bahan_belajar.ipynb" (the current notebook) and "base (Python 3.11.4)". The main area contains the following Python code:

```
File Edit Selection View Go Run Terminal Help
bahan_belajar.ipynb
bahan_belajar.ipynb > M+Scrape dengan menggunakan Python > M+Remove Stopwords > from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
+ Code + Markdown | Run All | Restart | Clear All Outputs | Variables | Outline ...
Remove Stopwords
1 from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
2 from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
3 import re
4
5 Comment Code
6 def getStopWords(customList=None,listFilePath=None):
7     #mendapatkan default stopwords bahasa Indonesia
8     factory = StopWordRemoverFactory()
9     stopwords = factory.get_stop_words()
10
11     #menambahkan daftar stopwords custom
12     if(customList is not None):
13         my_additional_stop_words=customList
14     else:
15         my_additional_stop_words=[]
16     if(listFilePath is not None):
17         with open(listFilePath) as f:
18             for line in f:
19                 my_additional_stop_words += [line[:-1]]
20
21     all_stopwords = stopwords + my_additional_stop_words
22
23     return all_stopwords
24
25 Comment Code
26 def removeStopWords(sentence, stop_words):
27     words = sentence.split()
28     return ' '.join([word for word in words if word.lower() not in stop_words])
29
30 konten_clean_from_stopwords=removeStopWords(konten_clean,getStopWords())
31 print(konten_clean_from_stopwords)
[38] 0.0s
...
Awardee Bongkar 3 Rahasia Beasiswa LPDP Intip Yuk Nikita Rosa detikEdu Senin 30 Okt 2023 2100 WIB Jakarta Margaret Ayu Caroline bertolak Inggris melanjutkan studi University College London UCL. Alumnus Fakultas Kedokteran Universitas Airlangga berangkat London beasiswa Lembaga Pengelola Dana Pendidikan LPDP Kementerian Keuangan. Oline panggilan akrabnya menceritakan keinginan melanjutkan kuliah berasal cita citanya seorang dosen klinis. Baca 3 Tips Persiapan Daftar S2 Luar Negeri Pekerja Catat NihBaca Sosok Dokter Gigi Disabilitas Peraih Beasiswa LPDP Jerman Aku punya cita cita jadi dosen klinis aku tertarik menjadi seorang pengajar katanaya laman Unair dikutip Senin 30102023. Ketertarikannya bidang ortopedi membuat Oline memilih Master of Science MSc Musculoskeletal Science and Medical Engineering jurusannya. memilih UCL bermitra salah satu rumah sakit ortopedi terbaik dunia. mahasiswa berminat beasiswa LPDP Oline membongkar rahasianya. Yuk simak berikut ini. Rahasia Beasiswa LPDP 1. Tentukan Alasan. Oline segala proses berjalan baik bila motivasi kuat. Bahkan Oline membutuhkan waktu cukup lama menentukan alasan kuat miliknya. memutuskan mendaftar kuliah aku menemukan alasan kuat melanjutkan studi berlangsung cukup lama. Kalau alasan kuat aku kesulitan menjalankan segala proses jelasknya. 2. Punya Mentor Memiliki seorang mentor menjadi tak kalah penting. mentor mendampingi memberi arahan kamu mengalami kesulitan. Mencari mentor artinya meminta bantuan bimbingan pendahulu sukses menimba pendidikan luar negeri. Sehebat sepintar apapun melakukan prosesnya sendiri papar Oline. 3. Cari Teman Seperjuangan. Oline melanjutkan pendidikan membuat seseorang merasa kesepian. Teman seperjuangan memberikan dorongan motivasi tempat berkeluh kesah posisi sama. Terlebih proses dijalani sebentar. Mulai mendaftar kampus hingga beasiswa. namanya menyiapkan persyaratan masuk kampus hingga beasiswa satu bulan. timeline tersusun ujarnya. Baca Kemenag Buka Pendaftaran Beasiswa Penyelesaian Studi S2 S3 Bernilai Blackbox ↗ Prettier ↗
```

08 TOKENIZER DAN STEMMER

membuat token dan
membentuk kata dasar
dari teks



```
1 modelName="indobenchmark/indobert-base-p1"
2
3 from transformers import AutoTokenizer,AutoModel,pipeline
4 from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
5 from Sastrawi.StopWordRemover.StopWordRemoverFactory import StopWordRemoverFactory
6 model=AutoModel.from_pretrained(modelName)
7 tokenizer=AutoTokenizer.from_pretrained(modelName)
8
9 Comment Code
10 def stemmer():
11     # Inisialisasi stemmer
12     factory=StemmerFactory()
13     stemmerAgent=factory.create_stemmer()
14     return stemmerAgent
15 stemmer=stemmer()

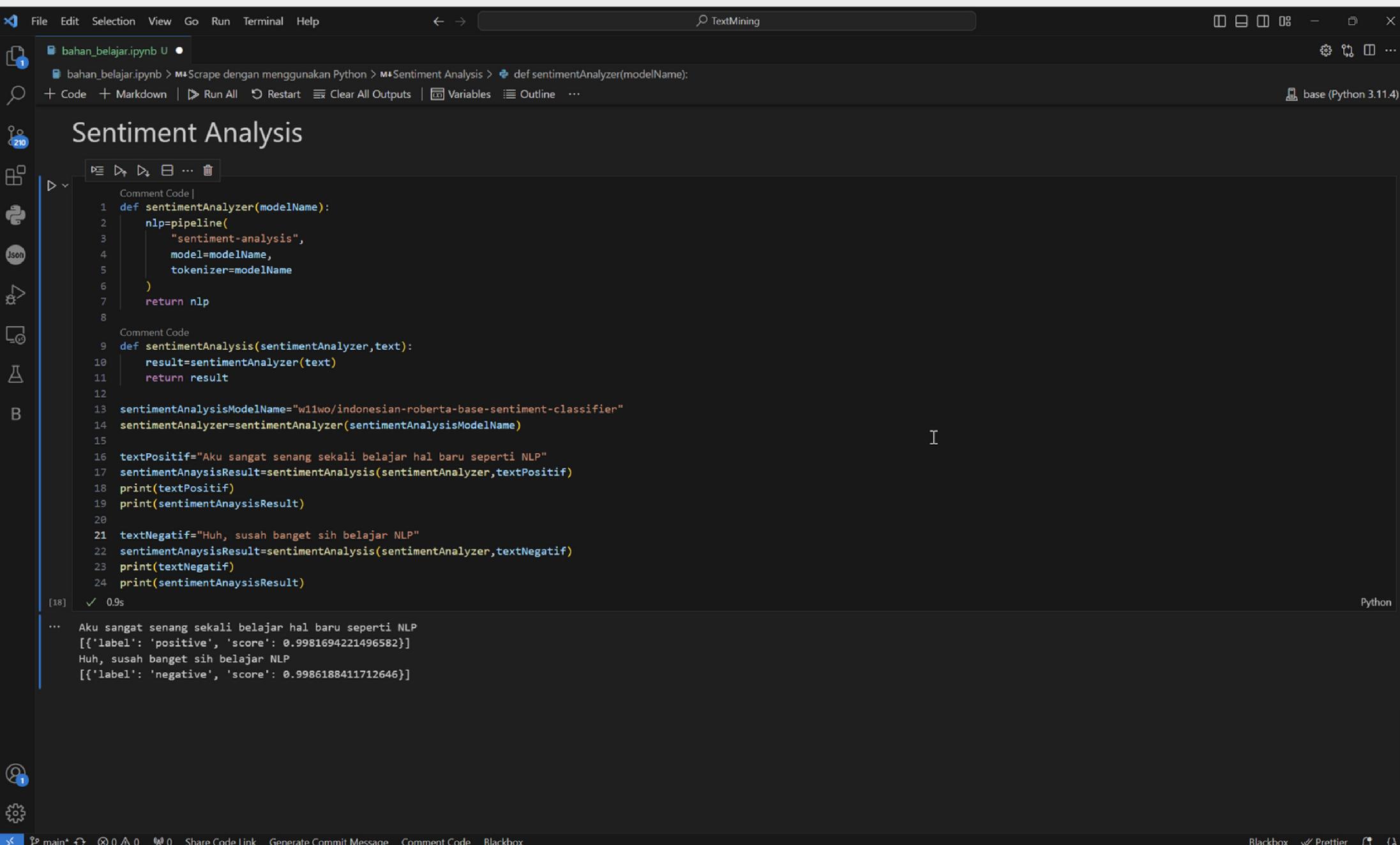
[9] ✓ 1.6s
```

```
Comment Code]
1 def stemFromText(stemmer,text):
2     stemmedText=stemmer.stem(text)
3     return stemmedText
4
5 stemText=stemFromText(stemmer,konten_clean_from_stopwords)
6 print(stemText)
[10] ✓ 3.3s
```

```
... awardee bongkar 3 rahasia beasiswa lpdp intip yuk nikita rosa detikedu senin 30 okt 2023 2100 wib jakarta margaret ayu caroline tolak inggris lanjut studi university college london ucl alumnus fakultas dokter universitas airlangga berangkat london beasiswa lembaga kelola dana didik lpdp menteri uang oline panggil akrab cerita ingin lanjut kuliah asal cita cita orang dosen klinis baca 3 tips siap daftar s2 luar negeri kerja catat nihbaca sosok dhani dokter gigi disabilitas raih beasiswa lpdp jerman aku punya cita cita jadi dosen klinis aku tarik jadi orang ajar kata laman unair kutip senin 30102023 tari bidang ortopedi buat oline pilih master of science msc musculoskeletal science and medical engineering jurus pilih ucl mitra salah satu rumah sakit ortopedi baik dunia mahasiswa minat beasiswa lpdp oline bongkar rahasia yuk simak ikut ini rahasia beasiswa lpdp 1 tentu alas oline segala proses jalan baik bila motivasi kuat bahkan oline butuh waktu cukup lama tentu alas kuat milik putus daftar kuliah aku temu alas kuat lanjut studi langsung cukup lama kalau alas kuat aku sulit jalan segala proses jelas 2 punya mentor milik orang mentor jadi tak kalah penting mentor damping beri arah kamu alami sulit cari mentor arti minta bantu bimbing dahulu sukses timbul didik luar negeri hebat pintar apa laku proses sendiri papar oline 3 cari teman juang oline lanjut didik buat orang rasa sepi teman juang beri dorongan motivasi tempat keluh kesah posisi sama lebih proses jalan sebentar mulai daftar kampus hingga beasiswa nama siap syarat masuk kampus hingga beasiswa satu bulan timeline susun ujar baca kemeneg buka daftar beasiswa selesai studi s2 s3 nilai rp 25 jutaonline pesan jalan segala proses orang nikmat meski asa sakit proses terus jalan pungkas
```

09 SENTIMENT ANALYSIS

melakukan prediksi
sentiment dari sebuah
teks



The screenshot shows a Jupyter Notebook interface with a dark theme. The title bar says 'bahar_belajar.ipynb'. The notebook has a single cell containing Python code for sentiment analysis. The code defines a function `sentimentAnalyzer` that creates an NLP pipeline with a specific model and tokenizer. It then defines a function `sentimentAnalysis` that takes a text input and uses the pipeline to analyze it. The code imports `nlp` from `transformers` and `TextMining` from `textmining`. It also includes two test cases: one for positive text ('Aku sangat senang sekali belajar hal baru seperti NLP') and one for negative text ('Huh, susah banget sih belajar NLP'). The output of the code shows the predicted sentiment and its score.

```
File Edit Selection View Go Run Terminal Help
bahar_belajar.ipynb > Scrape dengan menggunakan Python > Sentiment Analysis > def sentimentAnalyzer(modelName):
+ Code + Markdown | ▶ Run All ⚡ Restart ⌂ Clear All Outputs | ⏷ Variables ⏷ Outline ...
TextMining
base (Python 3.11.4)

Sentiment Analysis

Comment Code
1 def sentimentAnalyzer(modelName):
2     nlp=pipeline(
3         "sentiment-analysis",
4         model=modelName,
5         tokenizer=modelName
6     )
7     return nlp
8
9 Comment Code
10 def sentimentAnalysis(sentimentAnalyzer,text):
11     result=sentimentAnalyzer(text)
12     return result
13
14 sentimentAnalysisModelName="w11wo/indonesian-roberta-base-sentiment-classifier"
15 sentimentAnalyzer=sentimentAnalyzer(sentimentAnalysisModelName)
16
17 textPositif="Aku sangat senang sekali belajar hal baru seperti NLP"
18 sentimentAnaysisResult=sentimentAnalysis(sentimentAnalyzer,textPositif)
19 print(textPositif)
20 print(sentimentAnaysisResult)
21
22 textNegatif="Huh, susah banget sih belajar NLP"
23 sentimentAnaysisResult=sentimentAnalysis(sentimentAnalyzer,textNegatif)
24 print(textNegatif)
25 print(sentimentAnaysisResult)
[18] ✓ 0.9s
...
Aku sangat senang sekali belajar hal baru seperti NLP
[{'label': 'positive', 'score': 0.9981694221496582}]
Huh, susah banget sih belajar NLP
[{'label': 'negative', 'score': 0.9986188411712646}]
```

10 POS (PART OF SPEECH) ANALYSIS

melakukan prediksi
posisi dari sebuah teks
dalam konteks kalimat



A screenshot of a Jupyter Notebook interface titled "bahan_belajar.ipynb". The code cell contains Python code for performing Part of Speech (POS) analysis using the malaya library. The code defines a function `posAnalyzer` which takes a model name and returns a transformer object. Another function `posAnalysis` takes a text and a posAnalyzer object, and returns the result. The code then specifies `modelName="bert"` and performs the analysis on `konten_clean_from_stopwords`. The output of the code is a large list of POS-tagged words from the text, showing the part of speech for each word.

```
File Edit Selection View Go Run Terminal Help ← → ⌘ TextMining
bahan_belajar.ipynb U
bahani_scrape dengan menggunakan Python > POS Analysis > import malaya
+ Code + Markdown | Run All Restart Clear All Outputs Variables Outline ...
base (Python 3.11.4)

POSAnalysis
import malaya
Comment Code
def posAnalyzer(modelName):
    model=malaya.pos.transformer(model=modelName)
    return model
Comment Code
def posAnalysis(posAnalyzer,text):
    result=posAnalyzer.predict(text)
    return result
modelName="bert"
posAnalyzer=posAnalyzer(modelName)
posAnalysis=posAnalysis(posAnalyzer,konten_clean_from_stopwords)
print(posAnalysis)
[22] 13.1s
...
[('Awardee', 'PROPN'), ('Bongkar', 'PROPN'), ('3', 'NUM'), ('Rahasia', 'PROPN'), ('Beasiswa', 'PROPN'), ('LPDP', 'PROPN'), ('Intip', 'PROPN'), ('Yuk', 'PROPN'), ('Nikita', 'PROPN'), ('Rosa', 'PROPN'), ('detikEdu', 'PROPN'), ('Senin', 'PROPN'), ('30', 'NUM'), ('Okt', 'PROPN'), ('2023', 'NUM'), ('2100', 'NUM'), ('WIB', 'PROPN'), ('Jakarta', 'PROPN'), ('Margaret', 'PROPN'), ('Caroline', 'PROPN'), ('bertolak', 'VERB'), ('Inggris', 'PROPN'), ('melanjutkan', 'VERB'), ('studi', 'NOUN'), ('University', 'PROPN'), ('College', 'PROPN'), ('London', 'PROPN'), ('UCL', 'PROPN'), ('.', 'PUNCT'), ('Alumnus', 'PROPN'), ('Fakultas', 'PROPN'), ('Kedokteran', 'PROPN'), ('Universitas', 'PROPN'), ('Airlangga', 'PROPN'), ('berangkat', 'VERB'), ('London', 'PROPN'), ('beasiswa', 'PROPN'), ('Lembaga', 'PROPN'), ('Pengelola', 'PROPN'), ('Dana', 'PROPN'), ('Pendidikan', 'PROPN'), ('LPDP', 'PROPN'), ('Kementerian', 'PROPN'), ('Keuangan', 'PROPN'), ('.', 'PUNCT'), ('Oline', 'PROPN'), ('panggilan', 'NOUN'), ('akranya', 'ADJ'), ('menceritakan', 'VERB'), ('keinginan', 'NOUN'), ('melanjutkan', 'VERB'), ('kuliah', 'NOUN'), ('berasal', 'VERB'), ('cita', 'NOUN'), ('citanya', 'NOUN'), ('seorang', 'DET'), ('dosen', 'NOUN'), ('klinis', 'ADJ'), ('.', 'PUNCT'), ('Baca', 'VERB'), ('3', 'NUM'), ('Tips', 'PROPN'), ('Persiapan', 'PROPN'), ('Daftar', 'PROPN'), ('S2', 'PROPN'), ('Luar', 'PROPN'), ('Negeri', 'PROPN'), ('Pekerja', 'PROPN'), ('Catat', 'VERB'), ('NihBaca', 'PROPN'), ('Sosok', 'PROPN'), ('Dhani', 'PROPN'), ('Dokter', 'PROPN'), ('Gigi', 'PROPN'), ('Disabilitas', 'PROPN'), ('Peraih', 'PROPN'), ('Beasiswa', 'PROPN'), ('LPDP', 'PROPN'), ('Jerman', 'PROPN'), ('Aku', 'PRON'), ('punya', 'NOUN'), ('cita', 'NOUN'), ('jadi', 'VERB'), ('dosen', 'NOUN'), ('aku', 'PRON'), ('tertarik', 'VERB'), ('menjadi', 'VERB'), ('seorang', 'DET'), ('pengajar', 'NOUN'), ('katanya', 'NOUN'), ('lamau', 'NOUN'), ('Unair', 'PROPN'), ('dikutip', 'VERB'), ('Senin', 'PROPN'), ('30102023', 'NUM'), ('.', 'PUNCT'), ('Ketertarikannya', 'PROPN'), ('bidang', 'NOUN'), ('ortopedi', 'CCONJ'), ('membuat', 'VERB'), ('Oline', 'PROPN'), ('memilih', 'VERB'), ('Master', 'PROPN'), ('of', 'PROPN'), ('Science', 'PROPN'), ('MSc', 'PROPN'), ('Musculoskeletal', 'PROPN'), ('Science', 'PROPN'), ('and', 'CCONJ'), ('Medical', 'PROPN'), ('Engineering', 'PROPN'), ('jurusannya', 'NOUN'), ('.', 'PUNCT'), ('memilih', 'VERB'), ('UCL', 'PROPN'), ('bermitra', 'VERB'), ('salah', 'DET'), ('satu', 'DET'), ('rumah', 'NOUN'), ('ortopedi', 'PROPN'), ('terbaik', 'ADJ'), ('dunia', 'NOUN'), ('.', 'PUNCT'), ('mahasiswa', 'NOUN'), ('berminat', 'ADJ'), ('beasiswa', 'NOUN'), ('LPDP', 'PROPN'), ('Oline', 'PROPN'), ('membongkar', 'VERB'), ('rahasiaya', 'NOUN'), ('.', 'PUNCT'), ('Yuk', 'PROPN'), ('simak', 'VERB'), ('berikut', 'VERB'), ('ini', 'DET'), ('Rahasia', 'PROPN'), ('Beasiswa', 'PROPN'), ('LPDP', 'PROPN'), ('1', 'NUM'), ('.', 'PUNCT'), ('Tentukan', 'ADV'), ('Alasan', 'NOUN'), ('.', 'PUNCT'), ('Oline', 'PROPN'), ('segala', 'DET'), ('proses', 'NOUN'), ('berjalan', 'VERB'), ('baik', 'ADJ'), ('bila', 'SCONJ'), ('motivasi', 'NOUN'), ('kuat', 'ADJ'), ('.', 'PUNCT'), ('Bahkan', 'SCONJ'), ('Oline', 'PROPN'), ('membutuhkan', 'VERB'), ('waktu', 'NOUN'), ('cukup', 'ADJ'), ('menentukan', 'VERB'), ('alasan', 'NOUN'), ('kuat', 'ADJ'), ('miliknya', 'NOUN'), ('.', 'PUNCT'), ('memutuskan', 'VERB'), ('mendaftan', 'VERB'), ('kuliah', 'NOUN'), ('aku', 'PRON'), ('menemukan', 'VERB'), ('alasan', 'NOUN'), ('kuat', 'ADJ'), ('melanjutkan', 'VERB'), ('studi', 'NOUN'), ('berlangsung', 'VERB'), ('cukup', 'ADV'), ('lama', 'ADJ'), ('.', 'PUNCT'), ('kalau', 'SCONJ'), ('alasan', 'NOUN'), ('kuat', 'ADJ'), ('aku', 'PRON'), ('kesulitan', 'NOUN'), ('menjalankan', 'VERB'), ('segala', 'DET'), ('proses', 'NOUN'), ('jelasnya', 'ADJ'), ('.', 'PUNCT'), ('2', 'NUM'), ('.', 'PUNCT'), ('Punya', 'PROPN'), ('Mentor', 'PROPN'), ('Memiliki', 'VERB'), ('seorang', 'DET'), ('mentor', 'NOUN'), ('menjadi', 'VERB'), ('tak', 'ADV'), ('kalah', 'VERB'), ('tenting', 'ADJ'), ('.', 'PUNCT'), ('mentor', 'NOUN'), ('mendampingi', 'VERB'), ('memberi', 'VERB'), ('arah', 'NOUN'), ('kamu', 'PRON')]
```

11 NER (NAMED ENTITY RECOGNITION) ANALYSIS

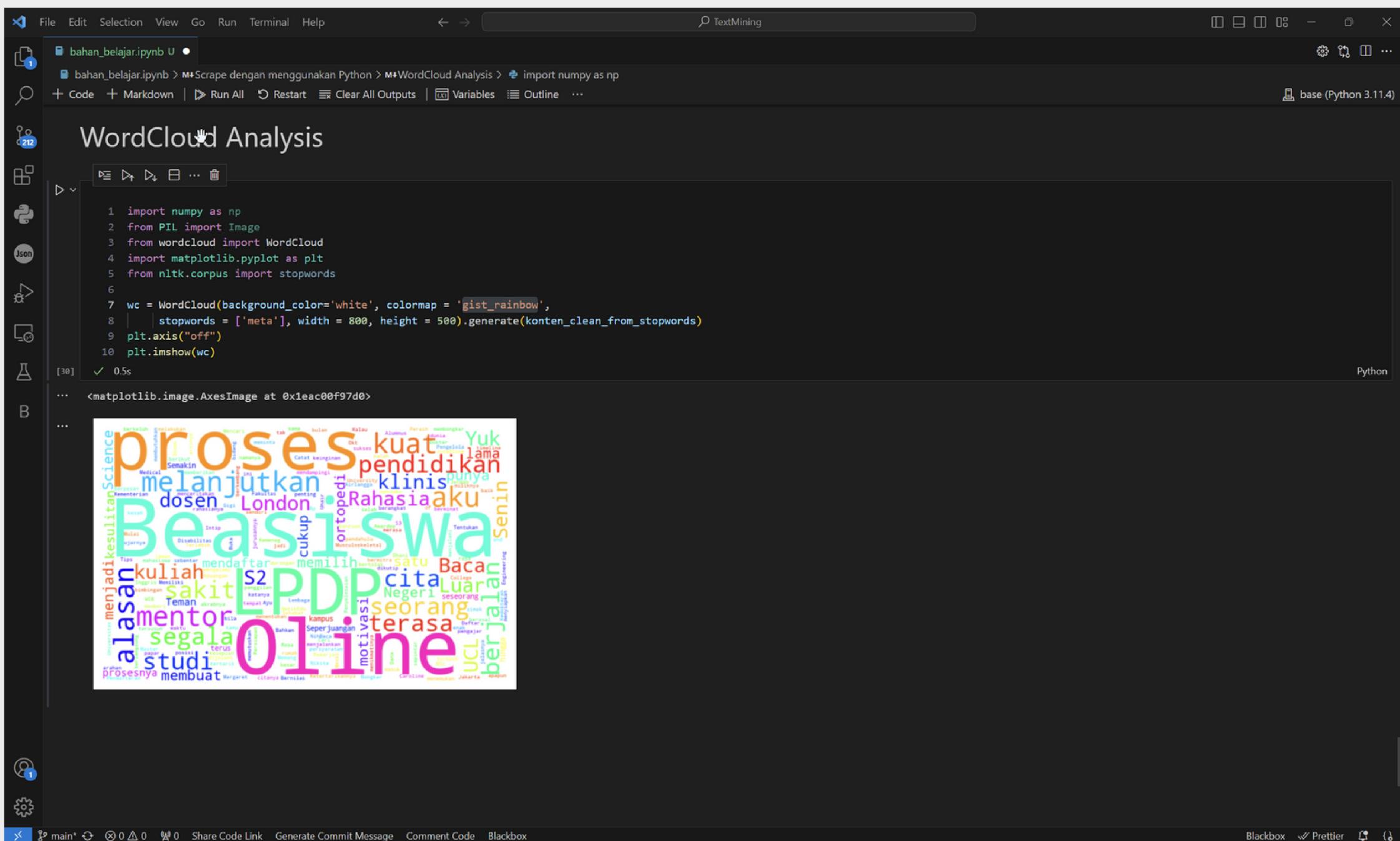
melakukan prediksi pengenalan entitas/subjek dari sebuah teks dalam konteks kalimat

The screenshot shows a Jupyter Notebook interface with the following details:

- File Bar:** File, Edit, Selection, View, Go, Run, Terminal, Help.
- Search Bar:** TextMining
- Toolbar:** Code, Markdown, Run All, Restart, Clear All Outputs, Variables, Outline, ...
- Bottom Status Bar:** base (Python 3.11.4)
- Section Header:** NER Analysis
- Code Cell:** Contains Python code for NER analysis using the malaya library. The code defines a `nerAnalyzer` function to load a model and a `nerAnalysis` function to predict entities in a text. It then initializes the `nerAnalyzer` with the "xlnet" model and prints the results.
- Output Cell:** Shows the execution time as 13.3s and the resulting JSON-like list of named entities extracted from the text.
- Bottom Status Bar:** Python

12 WORD CLOUD ANALYSIS

melakukan pengelompokan kata berdasarkan banyaknya kemunculan



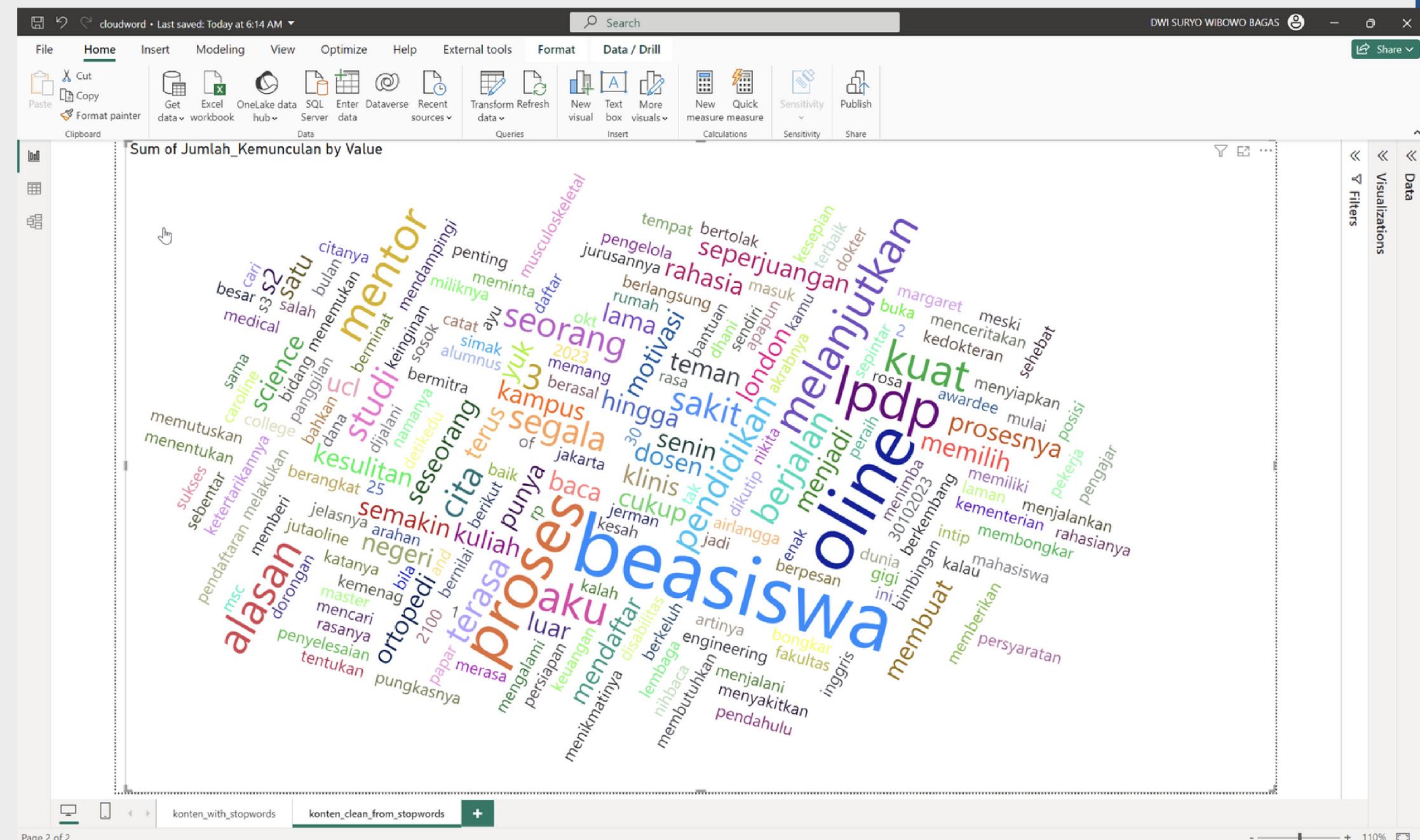
The screenshot shows a Jupyter Notebook interface titled "bahan_belajar.ipynb". The code cell contains Python code for generating a word cloud:

```
1 import numpy as np
2 from PIL import Image
3 from wordcloud import WordCloud
4 import matplotlib.pyplot as plt
5 from nltk.corpus import stopwords
6
7 wc = WordCloud(background_color='white', colormap = 'gist_rainbow',
8                 stopwords = ['meta'], width = 800, height = 500).generate(konten_clean_from_stopwords)
9 plt.axis("off")
10 plt.imshow(wc)
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
```

The output cell displays a colorful word cloud image with the following prominent words: proses, kuat, pendidikan, Yuk, melanjutkan, dosen, London, ortopedi, Rahasia, aku, Senin, Beasiswa, S2, Kuliah, sakit, cukaup, Baca, Luaran, Negeri, seorang, terasa, berjalan, Oline, mentor, segala, studi, membuat, motivasi, dan lain-lain.

12 WORD CLOUD ANALYSIS: POWER BI

melakukan pengelompokan kata berdasarkan banyaknya kemunculan



|||||

TERIMA KASIH

