**LAPORAN MODUL 4**

**BIG DATA ANALYTICS**



**Disusun oleh :**

Nama : Nurul Khofifah

NIM : 215410088

Kelas : IF 2

**PROGRAM STUDI INFORMATIKA**

**PROGRAM SARJANA**

**FAKULTAS TEKNOLOGI INFORMASI**

**UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA**

**YOGYAKARTA**

**2024**

**LAPORAN PRAKTIKUM BIG**

**DATA ANALYTICS**

**MODUL 4**

1. **TUJUAN PRAKTIKUM**

* Mampu melakukan Web Scraping

1. **DASAR TEORI**

Web scraping, web harvesting, web data extraction merupakan kegiatan yang

dilakukan untuk mengambil data tertentu secara semi-terstruktur dari sebuah halaman

website (wikipedia). Terdapat beberapa cara, Parsing HTML adalah salah satu Teknik

yang banyak digunakan untuk melakukan web scraping. Parsing HTML melakukan

ekstraksi dengan memisahkan tag, link maupun text/data. Bahasa program yang dipakai

bisa berbagai macam, pada materi ini digunakan bahasa Python dengan memanfaatkan

Modul BeautifulSoup4 dan parser lxml. Hal dasar yang harus diketahui adalah:

* Pahami tag-tag yang ada, seperti h1, h2, p, a, table, tr, td, dan lain-lain
* Pahami penggunaan class dan id, baik di tag layout maupun pengelompokan (div, span)
* Pahami konsep hirarki (tree) dari struktur HTML Untuk melihat struktur hirarki
* dengan mudah gunakan fasilitas inspect element di browser

Langkah :

* Install dahulu python
* Install BeautifulSoup → scraper

o conda install beautifulsoup4

* Install lxml → parser

o conda install lxml

o Catatan: BeautifulSoup memiliki parser bawaan “html.parser” yg tidak harus diinstall

* Install requests

o conda install requests → get web page

* Script Dasar :

from bs4 import BeautifulSoup

file\_html = open(“webku.html”,’r’)

page = BeautifulSoup(file\_html,”lxml”)

Jika baca dari Live Web

from bs4 import BeautifulSoup

import requests

url = “https://www.detik.com”

webpage = requests.get(url).text

page = BeautifulSoup(webpage, “lxml”)

Object dalam BeautifulSoup:

o Untuk akses atribut nama, dengan tag.name

o Untuk atribut misalnya : border, class atau id, dengan tag[‘border] atau tag[‘class’], dll

o Untuk berbagai attribut yang lain bisa juga dengan tag.attrs

Penggunaan atribut “text” akan mengambil isi text saja, dengan mengabaikan tag-tag yang ada Untuk mencari suatu yang lebih spesifik di halaman web, misalnya tag tertentu atau tag dengan nama class tertentu, dapat digunakan find atau find\_all

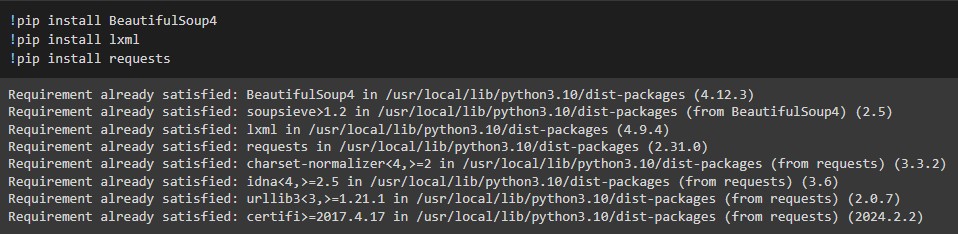
• Find → mengambil satu data pertama yang match

• Find\_all → mengambil semua data yang match

1. **PEMBAHASAN LISTING**

**Praktik**

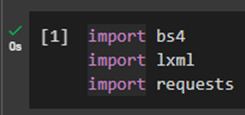
1. Install module pandas dan xlrd, dengan menjalankan :



Pembahasan :

Berikut adalah proses installasi progra yang dibutuhkan untuk melanjutkan program.

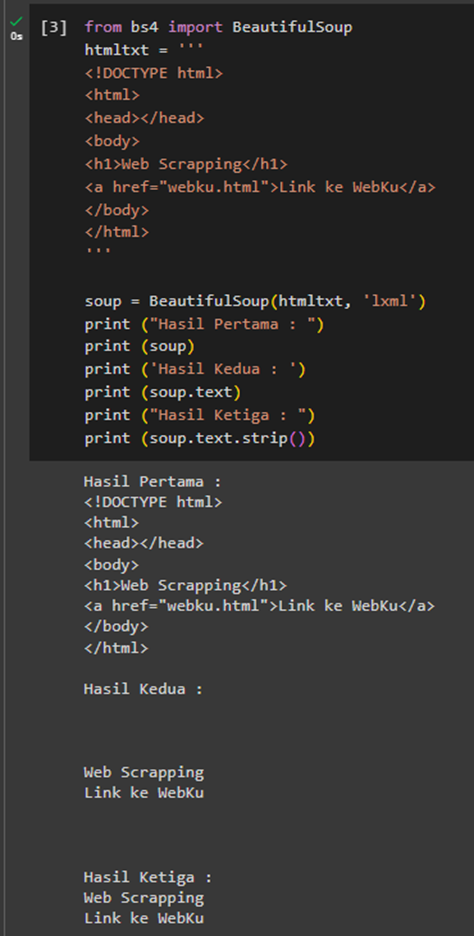
1. Memastikan installasi berhasil, masuk ke REPL python

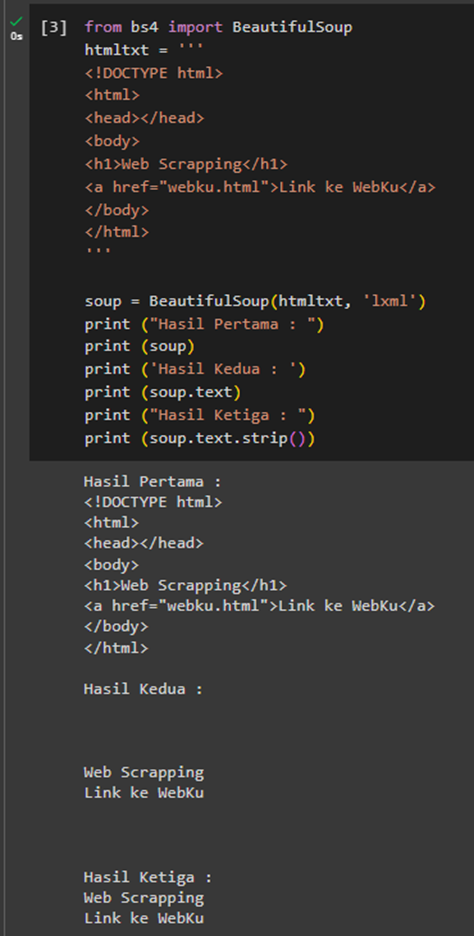


Pembahasan :

Berikut adalah proses pengecekan hasil installasi.

1. Bukalah Jupyter Notebook, ketikkan script berikut ini :



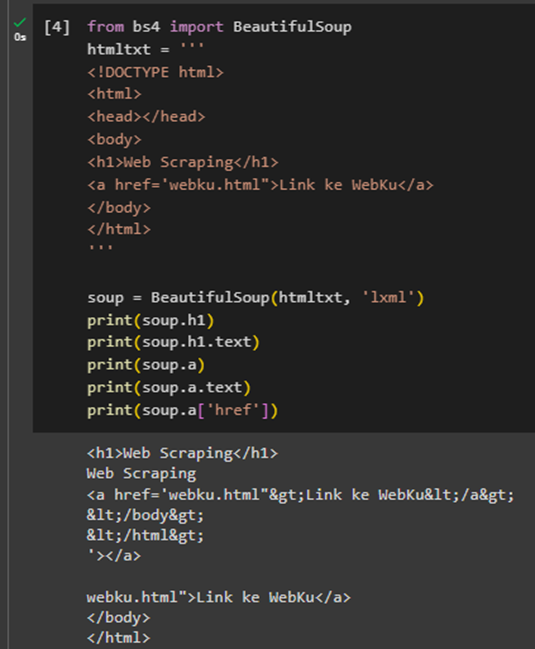


Pebahasan :

Berikut adalah program yang digunakan untuk mengambil data pada web.

1. Buatlah file baru dan copykan script program diatas, kemudian modifikasilah seperi

dibawah ini :

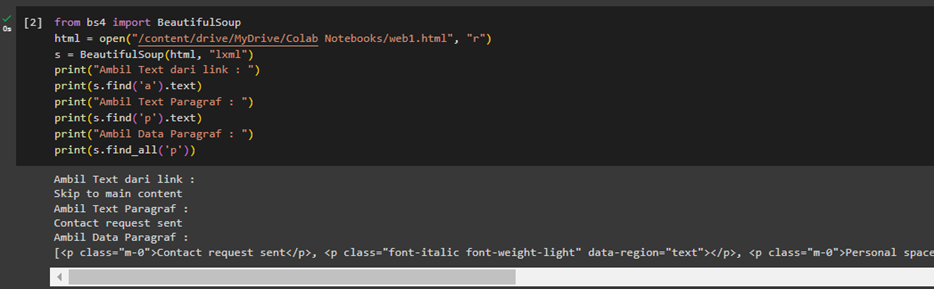


Pembahasan :

Berikut adalah program yang telah dimodifikasi yang digunakan untuk masuk ke dalam web.

1. Gunakan file web1.html, kemudian buatlah file baru dan ketikkan perintah seperti

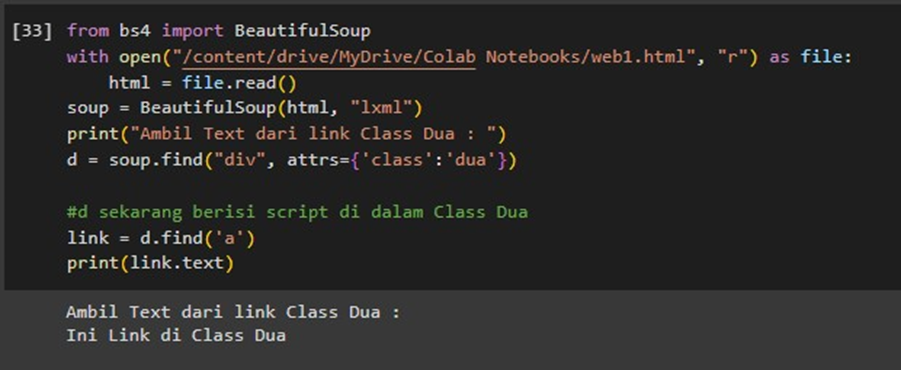
dibawah ini :



Pembahasan :

Berikut adalah program yang digunakan untuk mengambil link yang ada.

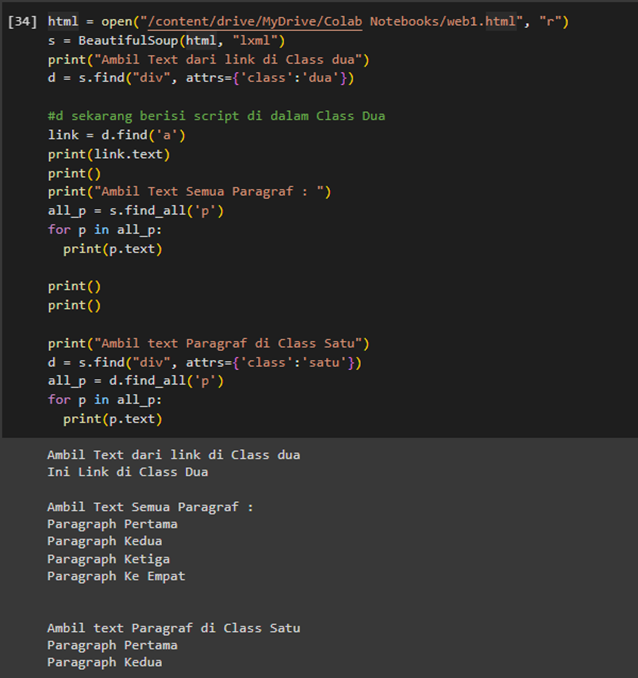
1. Buatlah file baru dan ketikkan script berikut ini

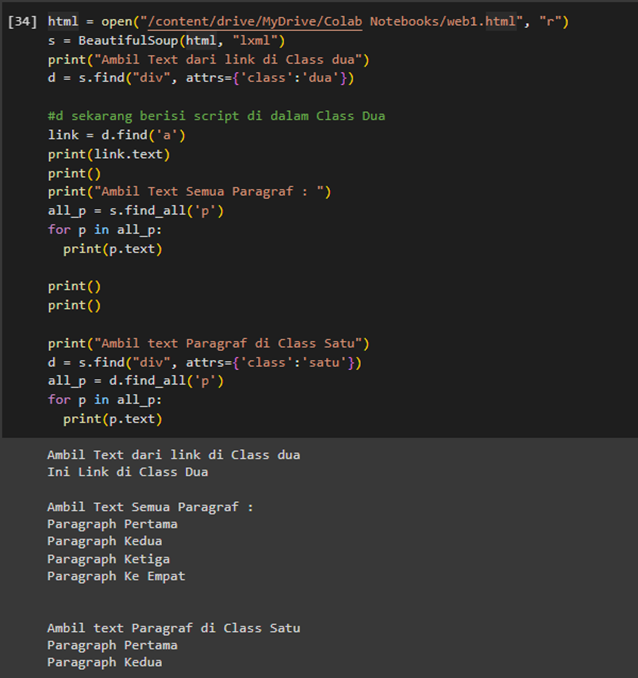


Pembahasan :

Berikut adalah program yang digunakan untuk mengambil link dari scrip ke 2.

1. Kemudian modifikasilah file dari no.7, tambahkan perintah seperti berikut ini

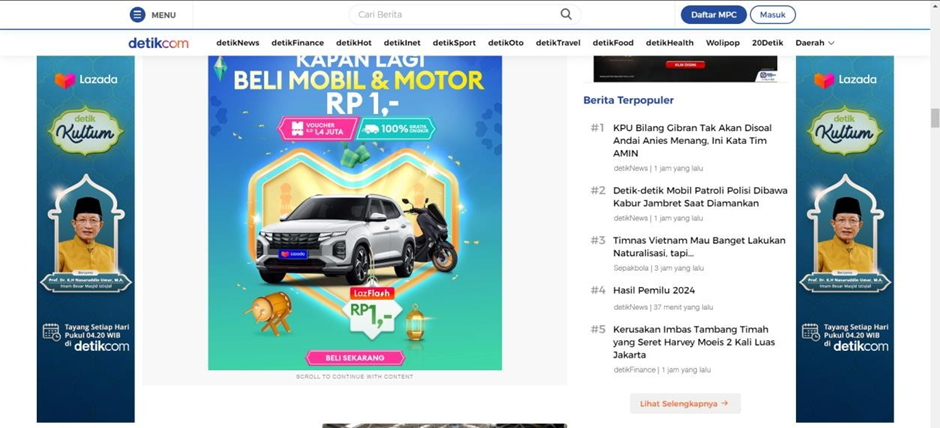




Pembahasan :

Berikut adalah program yang digunakan untuk mengambil teks pada bagian paragraf dari link yang tersedia.

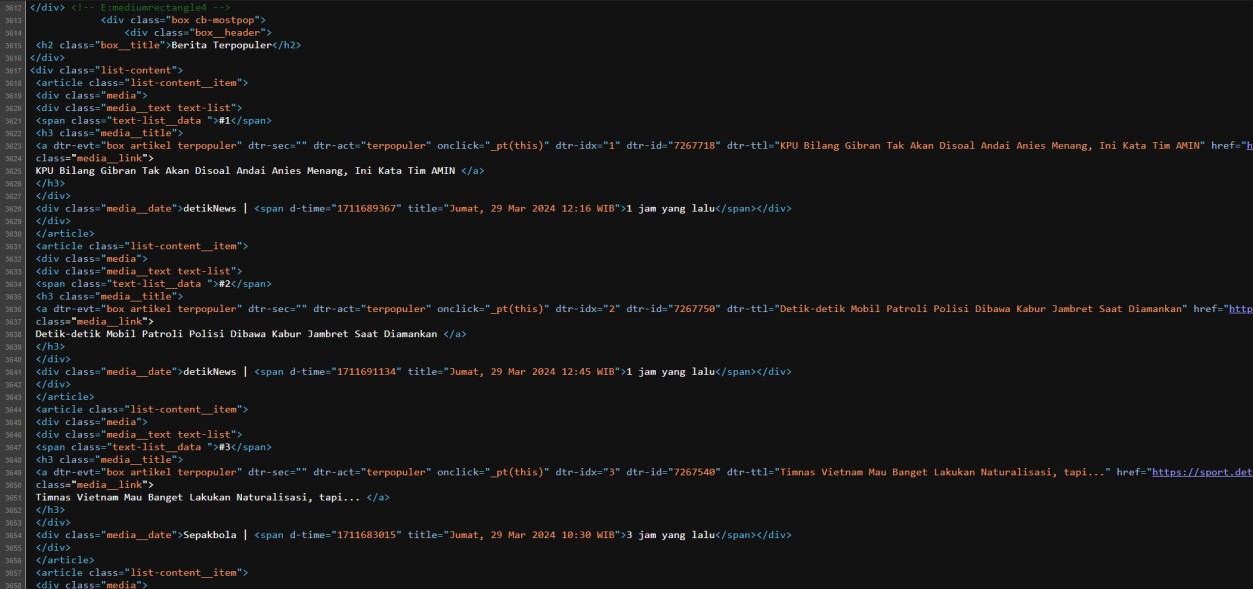
1. Bukalah Alamat Web: https://www.detik.com, inspect Element pada bagian “Terpopuler” :



Pembahasan :

Berikut adalah hasil dari web yang telah berhasil dibuka.

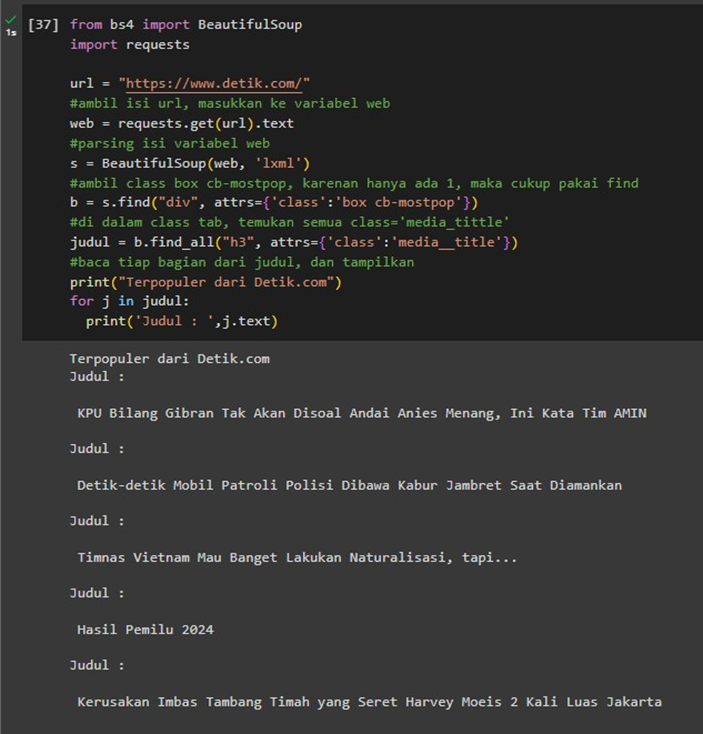
1. Klik kanan pada halaman tersebut, kemudian View Source



Pembahasan :

Berikutnya klik kanan pada halaman web untuk membuka scipt seperti pada gambar diatas.

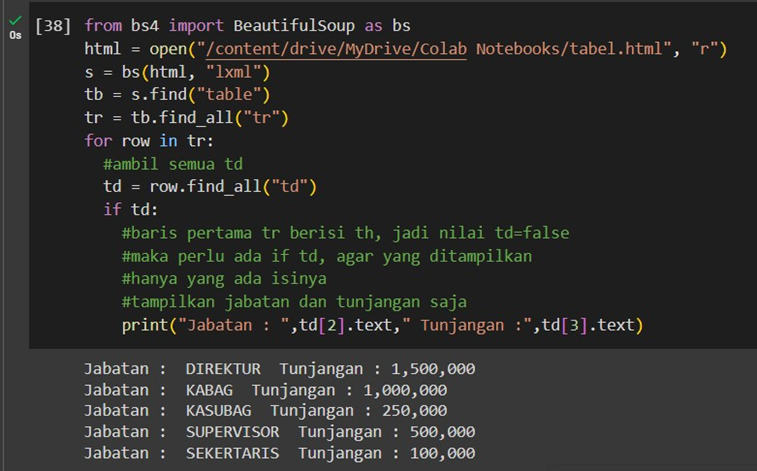
1. Buatlah file baru dan ketikkan perintah berikut ini



Pembahasan :

Berikut adalah progra yang digunakan untuk engambil data judul yang populer dari situs detik.com dan hasilnya seperti pada gambar.

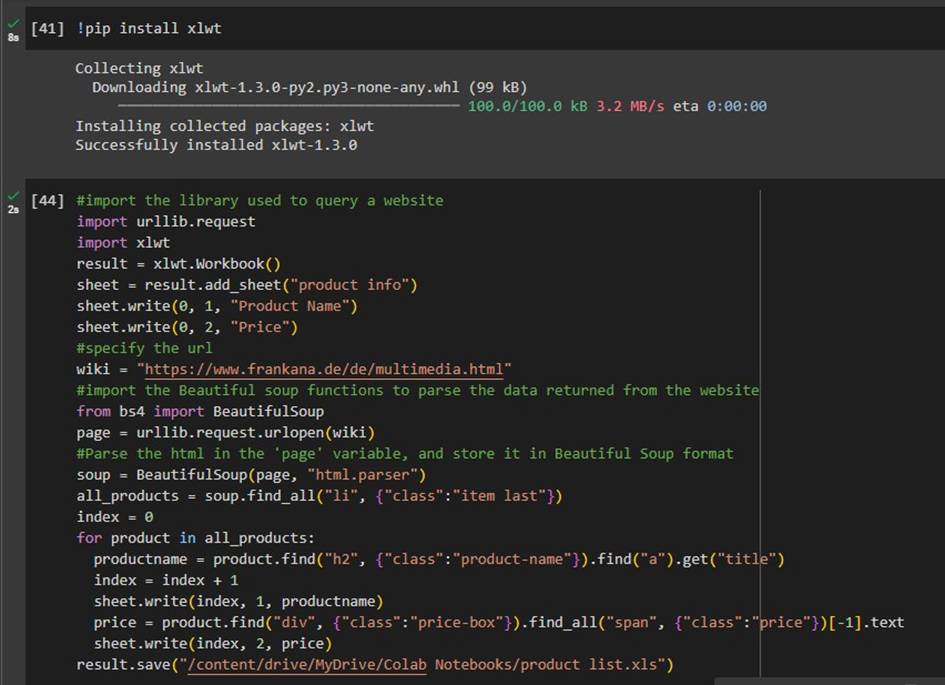
1. Gunakan file tabel.html lalu buat file baru, kemdian ketikkan script berikut ini



Pebahasan :

Berikut adalah program yang digunakan untuk mengambil data jabatan dan tujangan dari sebuah file yang tersedia.

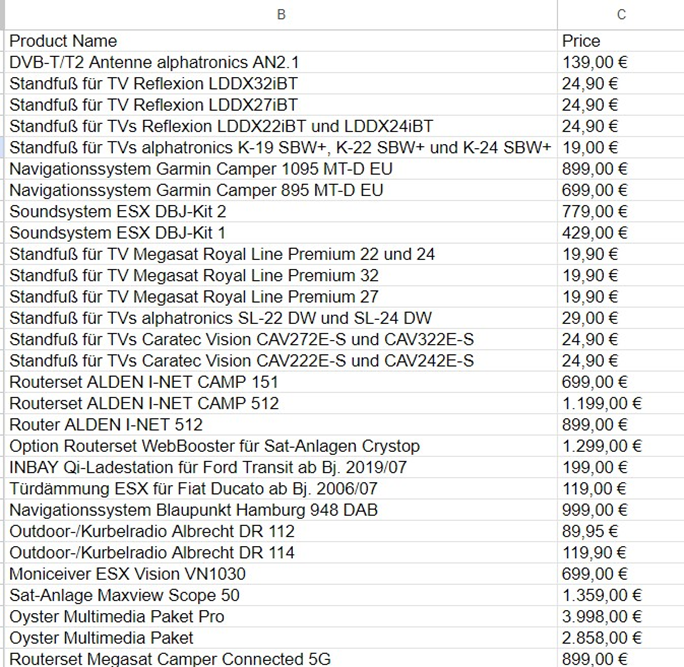
1. Kita buka contoh online shop <https://www.frankana.de/> dan Ketikkan perintah berikut ini:



Pembahasan :

Berikut adalah program yang digunakan untuk mengambil data produk dari online shop yang dituju.

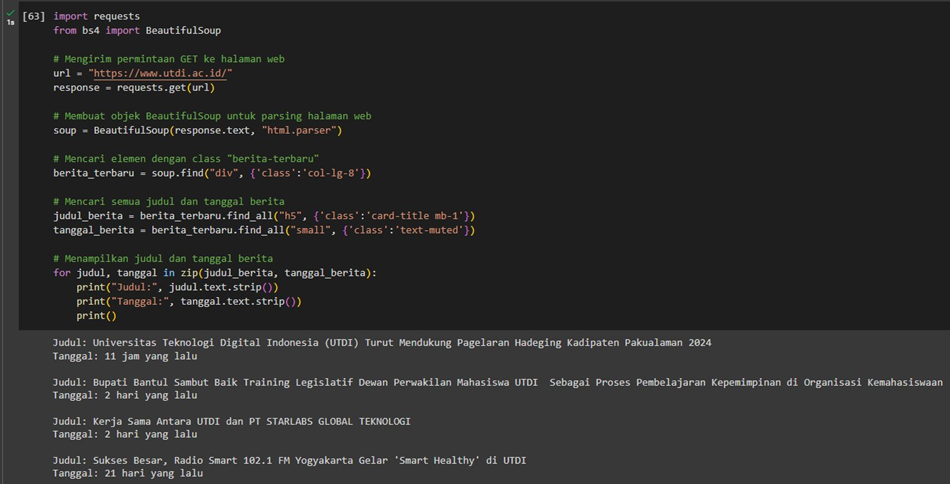
1. Hasilnya



Pembahasan :

Berikut adalah list hasil pengambilan data produk dari online shop yang tadi di gunakan.

1. **PEMBAHASAN TUGAS**
2. Bukalah Web https://www.utdi.ac.id/, kemudian lakukan scraping untuk “Berita Terbaru” (yang ditampilkan hanya Judul dan Tanggal).



Pembahasan :

Diatas adalah progra yang digunakan untuk mencari judul dan tanggal menguplod topik yang ada di dalam web tersebut.

1. **KESIMPULAN**

Pengumpulan data merupakan tahap kunci dalam penelitian yang memungkinkan penyelidik untuk mengumpulkan informasi yang diperlukan untuk analisis dan interpretasi lebih lanjut. Dengan menggunakan berbagai metode pengumpulan data yang sesuai, seperti survei, wawancara, atau observasi, para peneliti dapat memperoleh pemahaman yang lebih dalam tentang fenomena yang mereka teliti. Data yang dikumpulkan dengan cermat dan secara sistematis membantu memvalidasi temuan, mengidentifikasi pola atau tren, dan pada akhirnya, menghasilkan kesimpulan yang dapat dipercaya. Dengan demikian, pengumpulan data menjadi pondasi yang kuat bagi penelitian yang berarti dan relevan.