

Kuliah 2 STA1381 – Pengantar Sains Data

Septian Rahardiantoro



Outline

- Review: Web Scraping
- Ilustrasi Web Scraping
- Pembagian Topik Projek Kelompok

Review: Web Scraping

Review: Web Scraping

- Web scraping adalah proses pengambilan data dari suatu website
- Web scraper adalah program (perangkat lunak) yang secara otomatis dapat mengekstrak data-data yang diperlukan dari satu atau lebih laman web (web pages) yang dijadikan target, kemudian menyimpan data-data tersebut kedalam suatu sistem database untuk dimanipulasi sesuai kebutuhan.
- Terdapat 2 metode web scraping :
 - Manual : metode dengan menyalin data dengan cara copy-paste dari sebuah website
 - Otomatis: metode yang menggunakan koding, aplikasi, atau extension browser

Cara Kerja Web Scraper

- Masukkan satu atau lebih URL untuk dilakukan scraping
- Scraper akan memuat seluruh kode HTML untuk halaman tersebut. Web Scraper yang lebih canggih akan memberikan lebih banyak data termasuk elemen CSS dan Javascript.
- Setelah itu, scraper akan mulai mengekstrak data pada halaman atau data tertentu yang dipilih pengguna untuk dijalankan. Jadi pengguna harus memilih data spesifik yang ingin diperoleh dari suatu halaman.
 - Contoh: data produk dan harga dari suatu halaman e-commerce.
- Web Scraper akan mengekstraksi semua data yang sudah dikumpulkan ke dalam format yang lebih mudah dipahami oleh pengguna, seperti format csv atau excel.

Manfaat Web Scraping

- 1. Memantau harga suatu produk yang ada di pasaran.
- 2. Mencari informasi dari perusahaan lain
- 3. Mendapatkan Leads
- 4. Membandingkan Ulasan dalam Jumlah Besar

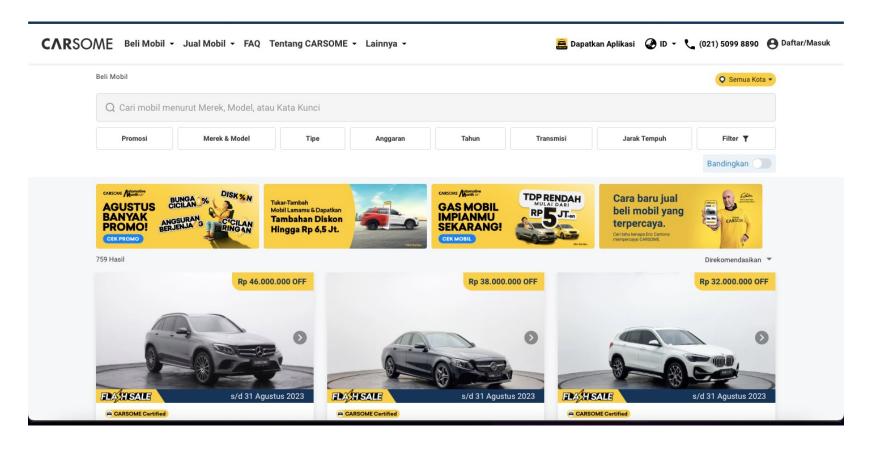
Teknik-teknik Web Scraping

- 1. Menyalin data secara manual
- 2. Menggunakan regular expression
- 3. Parsing HTML
- 4. Menganalisa DOM
- 5. Menggunakan Xpath
- 6. Menggunakan Google Sheet

Ilustrasi Web Scraping

1. Penerapan Web Scraper Untuk Mengambil Data Mobil Bekas

https://www.carsome.id/beli-mobil-bekas



```
library(rvest)

url <- "https://www.carsome.id/beli-mobil-bekas"
card <- read_html(url) %>% html_elements('.mod-card')
card %>%
  html_element('.mod-card__high-tag-list') %>%
  html_elements('span') %>%
  html_text() %>%
  paste(collapse = ',')
```

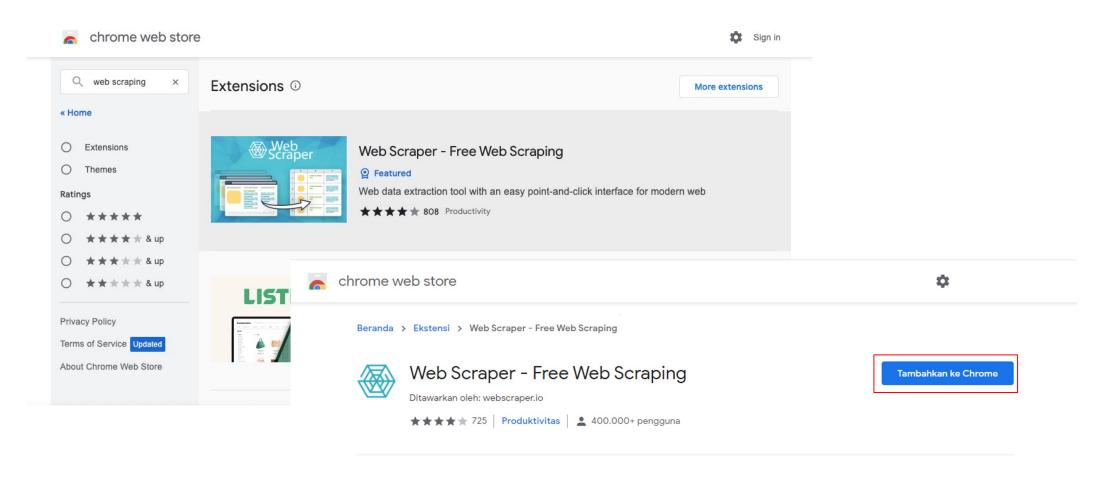
[1] "Plat Genap, Panoramic Sun Roof, Paddle Shift, Plat Genap, 360 Camera View, Sun Roof, Plat Ganjil, Panoramic Sun Roof, Electric Parking Brake, Plat Genap, Sports Mode, Map Navigator, Plat Genap, Sun Roof, Electric Parking Brake, Plat Ganjil, Jarak Tempuh Rendah, Electric Parking Brake, Plat Ganjil, Retract Mirror, Kamera Mundur, Plat Ganjil, Retract Mirror, Kamera Mundur, Plat Ganjil, Retract Mirror, Auto Cruise Control, Plat Ganjil, Retract Mirror, Sports Mode, Plat Genap, Jarak Tempuh Rendah, 360 Camera View, Plat Genap, Jarak Tempuh Rendah, Retract Mirror, Plat Genap, Auto Retract Mirror, Auto Cruise Control, Plat Ganjil, Auto Retract Mirror, Panoramic Sun Roof, Plat Ganjil, Jarak Tempuh Rendah, Electric Parking Brake, Plat Ganjil, Jarak Tempuh Rendah, Paddle Shift, Plat Ganjil, Sports Mode, Kamera Mundur, Plat Ganjil, Auto Cruise Control, Paddle Shift"

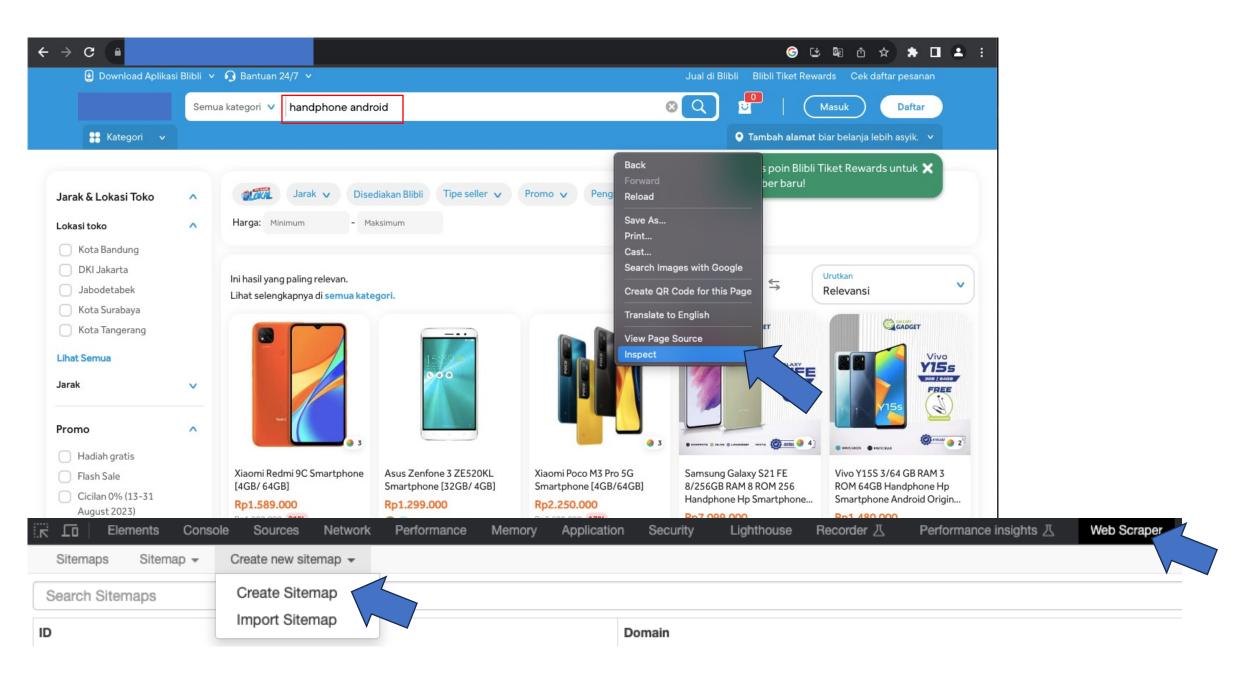
```
sapply(card,
      function(card) {
        card %>%
          html element('.mod-card high-tag-list') %>%
          html elements('span') %>%
          html text() %>%
          paste(collapse = ',')
      })
 [1] "Plat Genap, Panoramic Sun Roof, Paddle Shift"
 [2] "Plat Genap, 360 Camera View, Sun Roof"
 [3] "Plat Ganjil, Panoramic Sun Roof, Electric Parking Brake"
 [4] "Plat Genap, Sports Mode, Map Navigator"
 [5] "Plat Genap, Sun Roof, Electric Parking Brake"
 [6] "Plat Ganjil, Jarak Tempuh Rendah, Electric Parking Brake"
 [7] "Plat Ganjil, Retract Mirror, Kamera Mundur"
 [8] "Plat Ganjil, Retract Mirror, Kamera Mundur"
 [9] "Plat Ganjil, Retract Mirror, Auto Cruise Control"
[10] "Plat Ganjil, Retract Mirror, Sports Mode"
[11] "Plat Genap, Jarak Tempuh Rendah, 360 Camera View"
[12] "Plat Genap, Jarak Tempuh Rendah, Retract Mirror"
[13] "Plat Genap, Auto Retract Mirror, Auto Cruise Control"
[14] "Plat Ganjil, Auto Retract Mirror, Panoramic Sun Roof"
[15] "Plat Ganjil, Jarak Tempuh Rendah, Electric Parking Brake"
[16] "Plat Ganjil, Jarak Tempuh Rendah, Paddle Shift"
[17] "Plat Ganjil, Sports Mode, Kamera Mundur"
[18] "Plat Ganjil, Auto Cruise Control, Paddle Shift"
```

Dan seterusnya ...

2. Penerapan Web Scraper Untuk Mengambil Data Harga HP di ecommerce

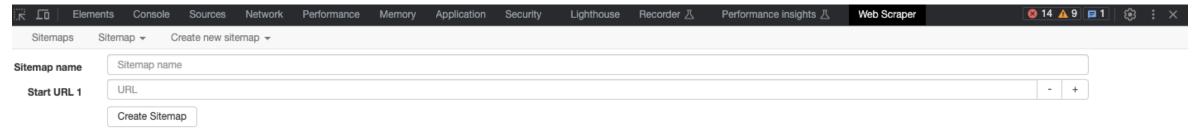
Dengan menggunakan Web Scraper (ekstensi pada google chrome)



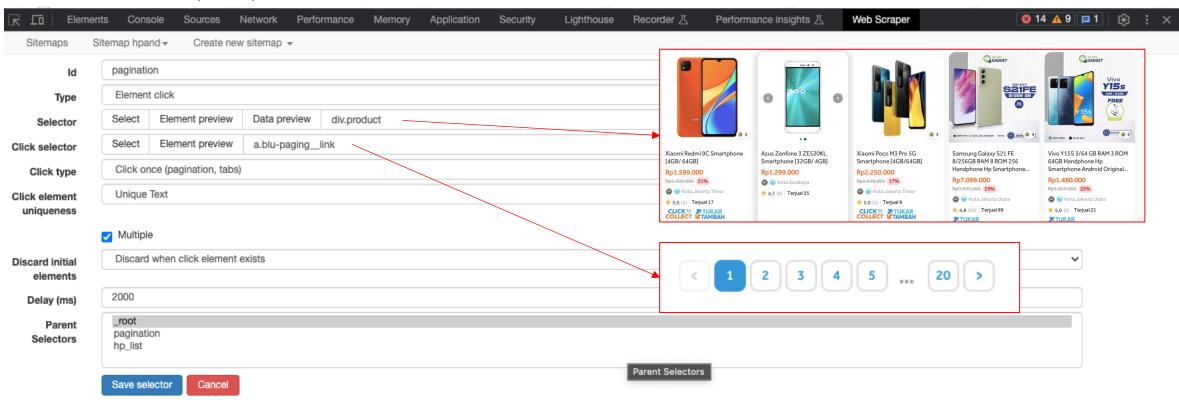


Create Sitemap

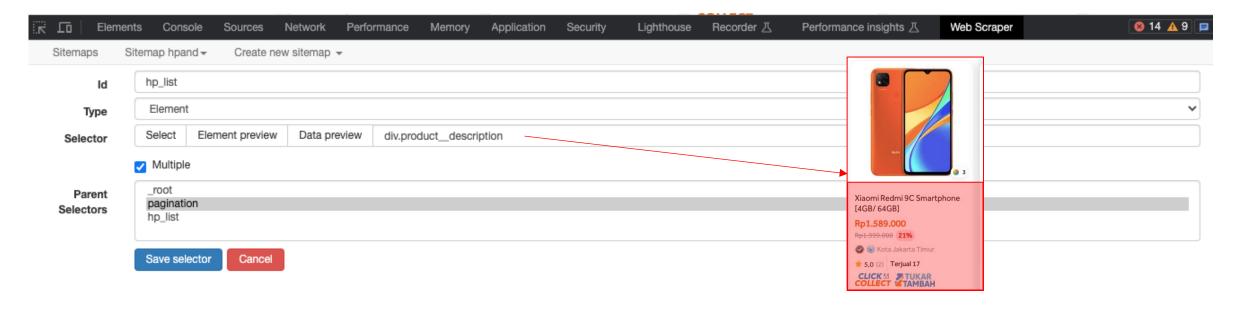
Isikan: Sitemap name dan Start URL 1



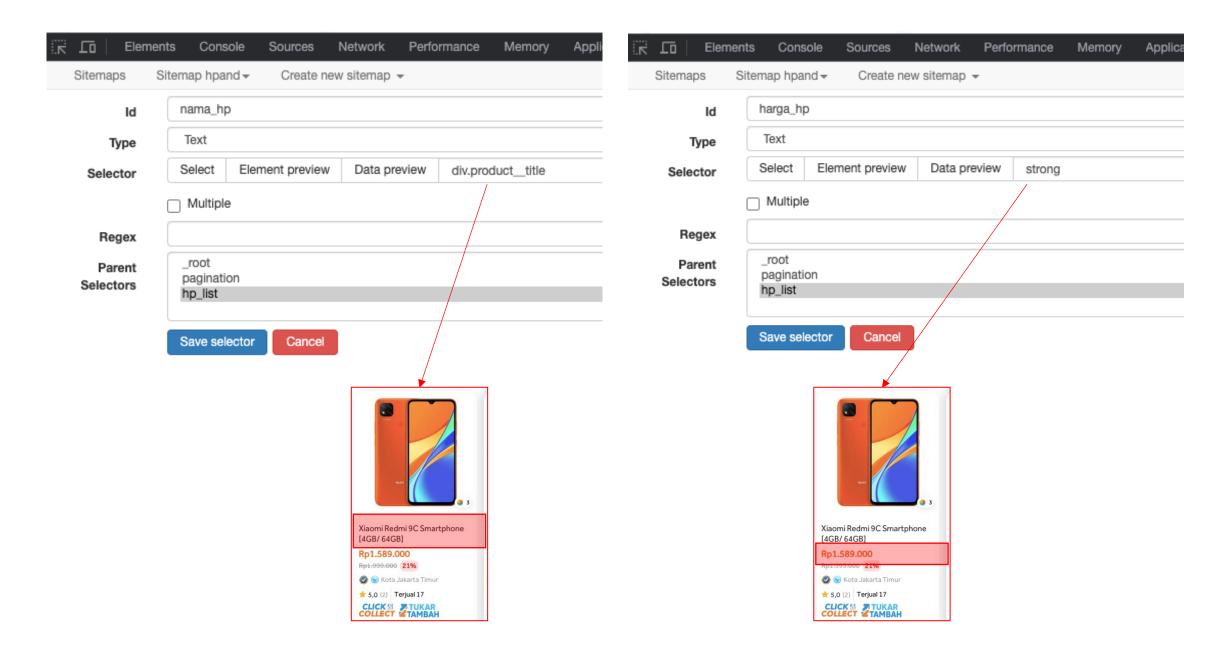
Add new selector (pada parent root)

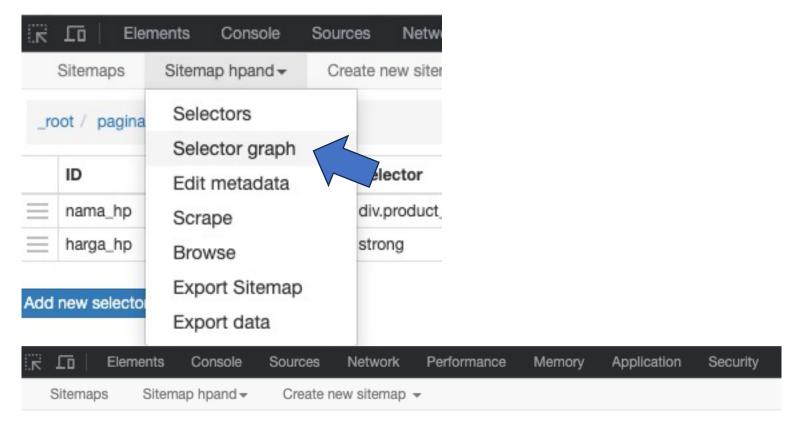


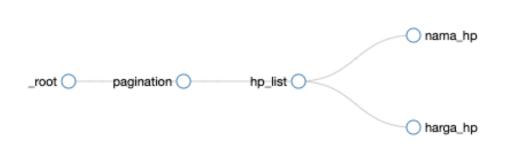
Add new selector (pada parent pagination)

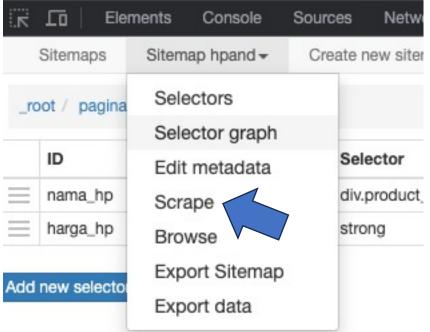


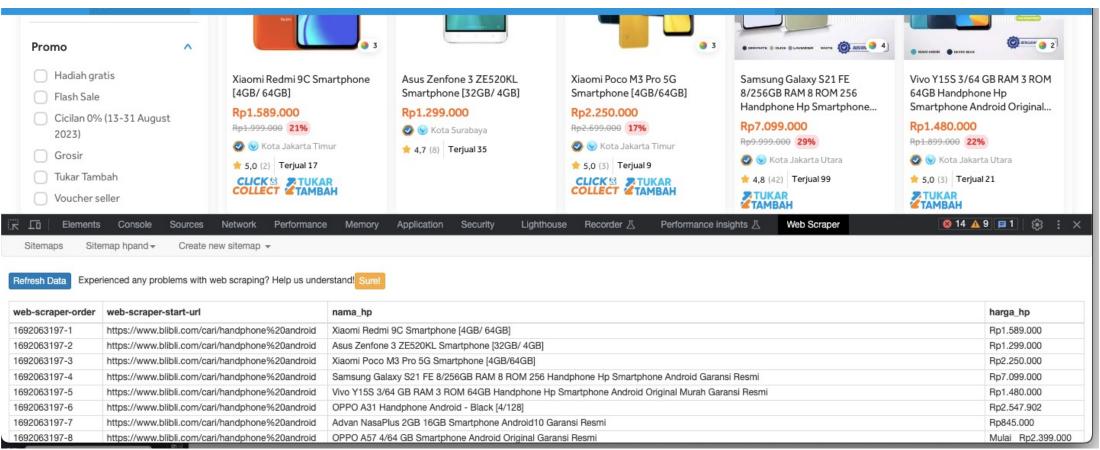
Add 2 buah new selector (pada parent hp_list)

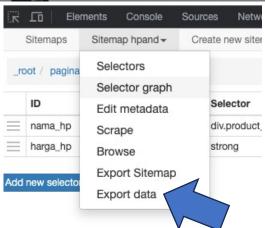












web-scraper-order	web-scraper-start-url	nama hp	harga hp	
1692063197-1	https://www.blibli.com/cari/handphone%20android	Xiaomi Redmi 9C Smartphone [4GB/ 64GB]	Rp1.589.000	
1692063197-2	https://www.blibli.com/cari/handphone%20android	Asus Zenfone 3 ZE520KL Smartphone [32GB/ 4GB]	Rp1.299.000	
1692063197-3	https://www.blibli.com/cari/handphone%20android	Xiaomi Poco M3 Pro 5G Smartphone [4GB/64GB]	Rp2.250.000	
1692063197-4	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy S21 FE 8/256GB RAM 8 ROM 256 Handphone Hp Smartphone Android Garansi Resmi	Rp7.099.000	
1692063197-5	https://www.blibli.com/cari/handphone%20android	Vivo Y15S 3/64 GB RAM 3 ROM 64GB Handphone Hp Smartphone Android Original Murah Garansi Resmi	Rp1.480.000	
1692063197-6	https://www.blibli.com/cari/handphone%20android		Rp2.547.902	
1692063197-7	https://www.blibli.com/cari/handphone%20android	Advan NasaPlus 2GB 16GB Smartphone Android10 Garansi Resmi	Rp845.000	
1692063197-8	https://www.blibli.com/cari/handphone%20android	OPPO A57 4/64 GB Smartphone Android Original Garansi Resmi	Mulai	Rp2.399.0
1692063197-9	https://www.blibli.com/cari/handphone%20android	Advan NasaPlus 2GB 16GB Smartphone Android10 Garansi Resmi Deep Blue	Rp845.000	
1692063197-10	https://www.blibli.com/cari/handphone%20android	Hp xiomi not 9	Rp2.100.000	
1692063197-11	https://www.blibli.com/cari/handphone%20android		Rp1.137.000	
1692063197-12	https://www.blibli.com/cari/handphone%20android	Infinix Hot 30i 8/128 RAM 8GB ROM 128GB 8 128 GB HP Android	Rp1.475.000	
1692063197-13	https://www.blibli.com/cari/handphone%20android	POCO C40 3/32 GB - Resmi	Rp1.258.000	
1692063197-14	https://www.blibli.com/cari/handphone%20android	Handphone OPPO A37	Rp700.000	
1692063197-15	https://www.blibli.com/cari/handphone%20android	XIAOMI REDMI 10A RAM 3GB - 64GB GARANSI RESMI	Mulai	Rp1.149.0
1692063197-16	https://www.blibli.com/cari/handphone%20android	Xiaomi MI 11T Pro 12/256 GB RAM 12 ROM 256GB Handphone Smartphone Android Original Garansi Resmi	Rp7.899.000	11,51.113.0
1692063197-17	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy A13 4/128 6/128 4GB 6GB 128GB RAM 4 6 ROM 128 GB HP Smartphone Android	Mulai	Rp2.500.0
1692063197-18	https://www.blibli.com/cari/handphone%20android	handphone samsung A12 6/128 resmi sein	Rp2.750.000	11,000.0
1692063197-19	https://www.blibli.com/cari/handphone%20android	Advan Nasa Pro 2 GB 32 GB Smartphone Android 11	Rp799.000	
1692063197-20	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy A73 5G 8/256 8GB 256GB RAM 8 ROM 256 GB Smartphone Android	Rp8.518.000	
1692063197-21	https://www.blibli.com/cari/handphone%20android	ALDO A9 SMARTPHONE RAM 3GB INTERNAL 16GB ANDROID	Rp680.000	
1692063197-22	https://www.blibli.com/cari/handphone%20android	HP Oppo A53S murah terbaru 2022 OPPO A53 S 5G RAM 8 + 256 GB Original New Baru Smartphone Android 6.52in HP OPPO Murah garansi 1 tahun	Rp2.459.000	
1692063197-23	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy S21 FE 8/256 GB RAM 8 ROM 256 Handphone Hp Smartphone Android Garansi Resmi	Rp7.149.000	
1692063197-24	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy A04e Smartphone [3/32GB]	Rp1.500.000	
1692063197-25	https://www.blibli.com/cari/handphone%20android	Samsung A14 4/128	Rp2.399.000	
1692063197-26	https://www.blibli.com/cari/handphone%20android	Oppo Reno 8T 5G 8/256GB	Rp6.199.000	
1692063197-27	https://www.blibli.com/cari/handphone%20android	OPPO A96 8GB/256GB Garansi Resmi 1 Tahun	Rp3.231.000	
1692063197-27		Vivo V25E 8/128 Garansi resmi vivo	•	
	https://www.blibli.com/cari/handphone%20android	·	Rp3.699.000	
1692063197-29	https://www.blibli.com/cari/handphone%20android	Huawei Y6 II LTE Smartphone - Black [16GB/ 2GB]	Rp3.849.000	
1692063197-30	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy A54 8/128GB	Rp5.599.000	
1692063197-31	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy A54 8/256GB	Rp5.999.000	
1692063197-32	https://www.blibli.com/cari/handphone%20android	SAMSUNG A03 CORE 2/32 GB [GARANSI RESMI]	Rp1.199.000	
1692063197-33	https://www.blibli.com/cari/handphone%20android	Xiaomi Redmi 9C Smartphone [4GB/ 64GB]	Rp1.625.000	
1692063197-34	https://www.blibli.com/cari/handphone%20android	Vivo Y16 - 3GB/32GB	Rp1.799.000	D - 4 700 0
1692063197-35	https://www.blibli.com/cari/handphone%20android		Mulai	Rp1.799.0
1692063197-36	https://www.blibli.com/cari/handphone%20android	Tecno Pova 4 (8GB + 128GB) Garansi Resmi	Rp2.269.000	
1692063197-37	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy A13 6/128GB	Rp2.487.500	2 700 0
1692063197-38	https://www.blibli.com/cari/handphone%20android	Samsung Galaxy A23 5G 6/128GB	Mulai	Rp3.799.0
1692063197-39	https://www.blibli.com/cari/handphone%20android	Xiaomi Redmi 9A Smartphone [32 GB/ 2 GB]	Rp1.110.000	
1692063197-40	https://www.blibli.com/cari/handphone%20android	Vivo Y16 Smartphone [4GB/128GB]	Rp1.899.000	
1692063197-41	https://www.blibli.com/cari/handphone%20android		Rp1.470.000	
1692063197-42	https://www.blibli.com/cari/handphone%20android	Redmi 9A 2/32GB	Rp1.500.500	
1692063197-43	https://www.blibli.com/cari/handphone%20android	Infinix Note 7 Lite Smartphone [4GB- 64GB]	Rp2.100.000	
1692063197-44	https://www.blibli.com/cari/handphone%20android	Xiaomi redmi 9T RAM 4/64 GB garansi resmi	Rp2.475.000	
1692063197-45	https://www.blibli.com/cari/handphone%20android	Samsung M23 5G 6/128 garansi resmi	Rp3.309.800	

Pembagian Topik Projek Kelompok

Projek

- Dikerjakan secara berkelompok
- Setiap kelompok akan mendapatkan topik projek tertentu secara acak, dengan pilihan topiknya
 - 1. Topik 1: Harga rumah dan spesifikasinya
 - 2. Topik 2: Harga leptop dan spesifikasinya
 - 3. Topik 3: Harga mobil bekas dan spesifikasinya
 - 4. Topik 4: Harga handphone dan spesifikasinya
- Spesifikasi yang diambil meliputi jenis objek, dan beberapa karakteristiknya (minimal 5 peubah)

• Lakukan web scraping sesuai dengan topik terpilih dengan minimal objek yang diambil adalah 500 buah objek

Topik	P1	P2
Topik 1: Harga rumah dan spesifikasinya	K4	K2
Topik 2: Harga leptop dan spesifikasinya	K2	К3
Topik 3: Harga mobil bekas dan spesifikasinya	K1	K1
Topik 4: Harga handphone dan spesifikasinya	K3	K4

Diskusikanlah peubah-peubah yang akan digunakan

Terima kasih ©