# WEB SCRAPPING

**Mobil Bekas OLX Indonesia** 



#### **Table Of Content**

- Konfigurasi virtual env dan requirements
- Import library python
- Start Chromedriver + Define URL
- Main Code
- Part 1 -> membuka url + membuat file csv
- Part 2 -> Inspek elemen li
- Part 2 -> Mencari element yang ingin diambil
- Part 2 -> Mengekstrak teks dari element tag
- Part 2 -> Menyimpan semua isi variable untuk dimasukkan ke file csv
- Part 3 -> exception dan menutup chromedeiver
- File csv hasil crawling
- Hasil file csv yang dibaca dengan pandas
- Referensi



# Konfigurasi Virtual Env dan requirements

- Buat folder project
  - Big\_data\_crawl
- Install python 3.8
- Install virtualenv
  - pip install virtualenv
- Buat env
  - Python3.8 -m venv .env3.8
  - Source .env3.8\bin\activate
- Install library
  - pip install selenium
  - pip install Pandas
  - · pip install beautifulsoup4
  - Pip install ipykernel
- Install chromedriver (sesuaikan dengan versi chrome)
  - Chromedriver 119.0.6045.105

# Import Library Python

```
import time
import csv
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.support import expected_conditions as EC
```



# Start Chromedriver + Define URL

```
# Path to the Chrome driver
driver_path = 'chromedriver-win64/chromedriver.exe'
service = Service(driver_path)
service.start()
driver = webdriver.Chrome(service=service)

url = 'https://www.olx.co.id/mobil-bekas_c198'
```



## Main Code



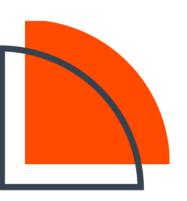
```
...
                            driver.get(url)
                             scroll_count = 0
Part 1
                             with open('crawl_langsung.csv', mode='w', newline='', encoding='utf-8') as file:
                                writer = csv.writer(file)
                                writer.writerow(['complete_link', 'merk', 'title_car', 'year', 'kilometers_range', 'price']) # Add the header
                                while True:
                     10
                     11
                                        wait = WebDriverWait(driver, 10) # Adjust the timeout value as needed
                     12
                                        time.sleep(15)
                     13
                                    except Exception as e:
                     14
                                        print(f"An error occurred: {e}")
                     15
                     16
                     17
                                    soup = BeautifulSoup(driver.page_source, 'html.parser')
                     18
                     19
                                    target_li_elements = soup.find_all('li', {'data-aut-category-id':'198'})
                     20
                     21
                                    for target_li in target_li_elements:
                     22
                                        target_a_tag = target_li.find('a')
                     23
                                        itemPrice = target_li.find('span', {'data-aut-id': 'itemPrice'})
                     24
                                        itemSubTitle = target_li.find('div', {'data-aut-id': 'itemSubTitle'})
                     25
                                        itemTitle = target_li.find('div', {'data-aut-id': 'itemTitle'})
                     26
                     27
                                        if target_a_tag:
                     28
                                            link = target_a_tag['href']
Part 2
                     29
                                            complete_link = 'https://www.olx.co.id' + link
                     30
                     31
                                        # take text for item
                     32
                                        span_item_price = itemPrice.text
                     33
                                        price = span_item_price.split()[-1]
                     34
                     35
                                        # take text for year and kilometers range
                     36
                                        div_item_sub_title = itemSubTitle.text
                     37
                                        parts = div_item_sub_title.split(' - ')
                     38
                                        year = parts[0]
                     39
                                        kilometers_range = parts[1].split(' ')[0]
                     40
                                        # taek text for merk and car title
                     41
                     42
                                        div_item_title = itemTitle.text
                    43
                                        merk = div_item_title.split(' ')[0]
                                        title_car = div_item_title.split(' ')[-1]
                     45
                     46
                                        writer.writerow([complete_link, merk, title_car, year, kilometers_range, price ])
                     47
                     48
                                    scroll_count += 1
                     49
                        except Exception as e:
                            print(f"An error occurred: {e}")
Part 3
                     53 finally:
                            driver.quit()
```

# Part 1 –> membuka url + membuat file csv

```
try:
    driver.get(url)
    scroll_count = 0

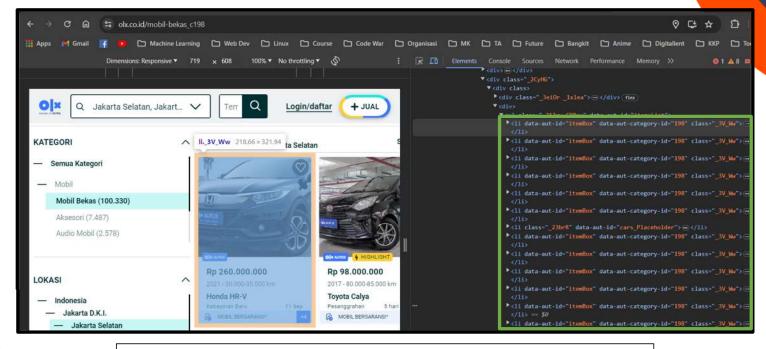
with open('crawl_langsung.csv', mode='w', newline='', encoding='utf-8') as file:
    writer = csv.writer(file)
    # Add the header
    writer.writerow(['complete_link','merk', 'title_car', 'year', 'kilometers_range', 'price'])
```

#### Part 2



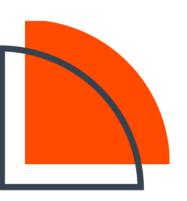
```
. . .
while True:
               wait = WebDriverWait(driver, 10) # Adjust the timeout value as needed
           except Exception as e:
               print(f"An error occurred: {e}")
        soup = BeautifulSoup(driver.page_source, 'html.parser')
        target_li_elements = soup.find_all('li', {'data-aut-category-id':'198'})
           for target_li in target_li_elements:
           target_a_tag = target_li.find('a')
               itemPrice = target_li.find('span', {'data-aut-id': 'itemPrice'})
               itemSubTitle = target_li.find('div', {'data-aut-id': 'itemSubTitle'})
             itemTitle = target_li.find('div', {'data-aut-id': 'itemTitle'})
             → if target_a_tag:
                   link = target_a_tag['href']
                   complete link = 'https://www.olx.co.id' + link
               # take text for item
               span item price = itemPrice.text
               price = span_item_price.split()[-1]
               # take text for year and kilometers range
               div item sub title = itemSubTitle.text
               parts = div_item_sub_title.split(' - ')
               year = parts[0]
               kilometers_range = parts[1].split(' ')[0]
               # taek text for merk and car title
               div item title = itemTitle.text
               merk = div_item_title.split(' ')[0]
             title_car = div_item_title.split(' ')[-1]
               writer.writerow([complete_link, merk, title_car, year, kilometers_range, price ])
           scroll_count += 1
```

# Part 2 -> Inspek elemen li



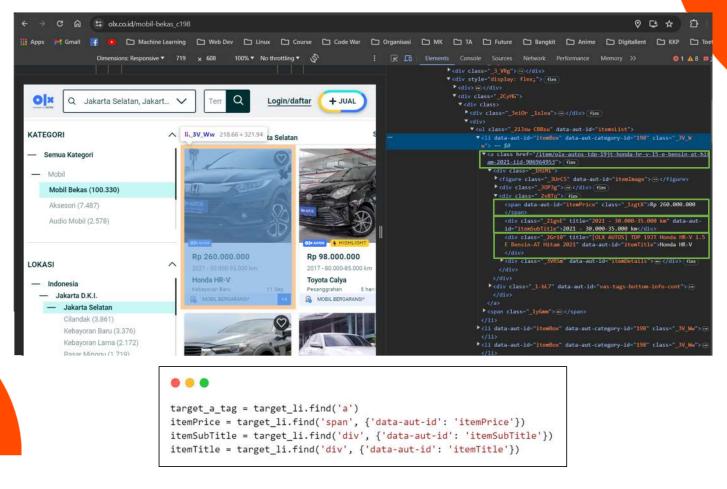
```
soup = BeautifulSoup(driver.page_source, 'html.parser')
target_li_elements = soup.find_all('li', {'data-aut-category-id':'198'})
```

#### Part 2



```
. . .
while True:
               wait = WebDriverWait(driver, 10) # Adjust the timeout value as needed
           except Exception as e:
               print(f"An error occurred: {e}")
        soup = BeautifulSoup(driver.page_source, 'html.parser')
        target_li_elements = soup.find_all('li', {'data-aut-category-id':'198'})
           for target_li in target_li_elements:
           target_a_tag = target_li.find('a')
               itemPrice = target_li.find('span', {'data-aut-id': 'itemPrice'})
               itemSubTitle = target_li.find('div', {'data-aut-id': 'itemSubTitle'})
             itemTitle = target_li.find('div', {'data-aut-id': 'itemTitle'})
             → if target_a_tag:
                   link = target_a_tag['href']
                   complete link = 'https://www.olx.co.id' + link
               # take text for item
               span item price = itemPrice.text
               price = span_item_price.split()[-1]
               # take text for year and kilometers range
               div item sub title = itemSubTitle.text
               parts = div_item_sub_title.split(' - ')
               year = parts[0]
               kilometers_range = parts[1].split(' ')[0]
               # taek text for merk and car title
               div item title = itemTitle.text
               merk = div_item_title.split(' ')[0]
             title_car = div_item_title.split(' ')[-1]
               writer.writerow([complete_link, merk, title_car, year, kilometers_range, price ])
           scroll_count += 1
```

# Part 2 -> Mencari element yang ingin diambil



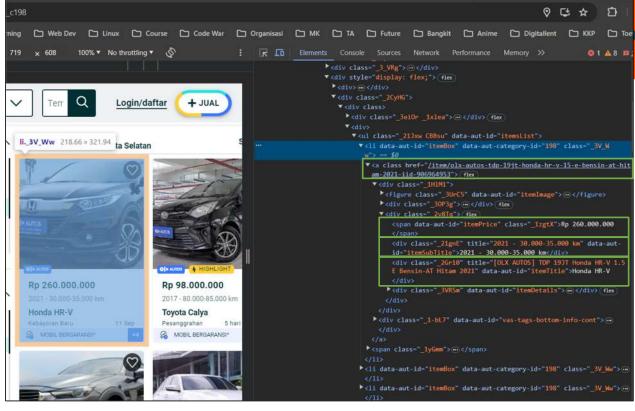
#### Part 2



```
...
while True:
               wait = WebDriverWait(driver, 10) # Adjust the timeout value as needed
           except Exception as e:
               print(f"An error occurred: {e}")
        soup = BeautifulSoup(driver.page_source, 'html.parser')
        target_li_elements = soup.find_all('li', {'data-aut-category-id':'198'})
           for target_li in target_li_elements:
           target_a_tag = target_li.find('a')
               itemPrice = target_li.find('span', {'data-aut-id': 'itemPrice'})
               itemSubTitle = target_li.find('div', {'data-aut-id': 'itemSubTitle'})
             itemTitle = target_li.find('div', {'data-aut-id': 'itemTitle'})
             → if target_a_tag:
                   link = target_a_tag['href']
                   complete link = 'https://www.olx.co.id' + link
               # take text for item
               span item price = itemPrice.text
               price = span_item_price.split()[-1]
               # take text for year and kilometers range
               div item sub title = itemSubTitle.text
               parts = div_item_sub_title.split(' - ')
               year = parts[0]
               kilometers_range = parts[1].split(' ')[0]
               # taek text for merk and car title
               div item title = itemTitle.text
               merk = div_item_title.split(' ')[0]
             title_car = div_item_title.split(' ')[-1]
               writer.writerow([complete_link, merk, title_car, year, kilometers_range, price ])
           scroll_count += 1
```

# Part 2 -> Mengekstrak teks dari element tag

```
. .
if target_a_tag:
   link = target a tag['href']
    complete_link = 'https://www.olx.co.id' + link
# take text for item
span_item_price = itemPrice.text
                                    Rp 260.000.000
price = span_item_price.split()[-1]
# take text for year and kilometers range
div_item_sub_title = itemSubTitle.text 2021 - 30.000-35.000 kms
parts = div_item_sub_title.split(' - ')
year = parts[0]
kilometers range = parts[1].split(' ')[0]
# taek text for merk and car title
div item title = itemTitle.text
merk = div_item_title.split(' ')[0]
title_car = div_item_title.split(' ')[-1]
```



Part 2->
Menyimpan semua
isi variable untuk
dimasukkan ke file
csv



```
. . .
while True:
               wait = WebDriverWait(driver, 10) # Adjust the timeout value as needed
           except Exception as e:
               print(f"An error occurred: {e}")
         soup = BeautifulSoup(driver.page_source, 'html.parser')
         target_li_elements = soup.find_all('li', {'data-aut-category-id':'198'})
           for target_li in target_li_elements:
            target a tag = target li.find('a')
               itemPrice = target_li.find('span', {'data-aut-id': 'itemPrice'})
               itemSubTitle = target_li.find('div', {'data-aut-id': 'itemSubTitle'})
              itemTitle = target_li.find('div', {'data-aut-id': 'itemTitle'})
             if target_a_tag:
                   link = target_a_tag['href']
                   complete link = 'https://www.olx.co.id' + link
               # take text for item
               span_item_price = itemPrice.text
               price = span_item_price.split()[-1]
               # take text for year and kilometers range
               div_item_sub_title = itemSubTitle.text
               parts = div_item_sub_title.split(' - ')
               year = parts[0]
               kilometers_range = parts[1].split(' ')[0]
               # taek text for merk and car title
               div item title = itemTitle.text
               merk = div_item_title.split(' ')[0]
              title car = div item title.split(' ')[-1]
               writer.writerow([complete_link, merk, title_car, year, kilometers_range, price ])
           scroll_count += 1
```

## Main Code



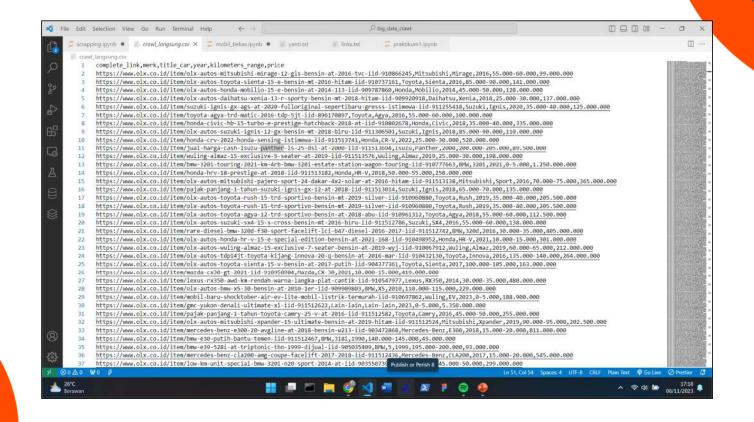
```
...
                            driver.get(url)
                             scroll_count = 0
Part 1
                             with open('crawl_langsung.csv', mode='w', newline='', encoding='utf-8') as file:
                                writer = csv.writer(file)
                                writer.writerow(['complete_link', 'merk', 'title_car', 'year', 'kilometers_range', 'price']) # Add the header
                                while True:
                     10
                     11
                                        wait = WebDriverWait(driver, 10) # Adjust the timeout value as needed
                     12
                                        time.sleep(15)
                     13
                                    except Exception as e:
                     14
                                        print(f"An error occurred: {e}")
                     15
                     16
                     17
                                    soup = BeautifulSoup(driver.page_source, 'html.parser')
                     18
                     19
                                    target_li_elements = soup.find_all('li', {'data-aut-category-id':'198'})
                     20
                     21
                                    for target_li in target_li_elements:
                     22
                                        target_a_tag = target_li.find('a')
                     23
                                        itemPrice = target_li.find('span', {'data-aut-id': 'itemPrice'})
                     24
                                        itemSubTitle = target_li.find('div', {'data-aut-id': 'itemSubTitle'})
                     25
                                        itemTitle = target_li.find('div', {'data-aut-id': 'itemTitle'})
                     26
                     27
                                        if target_a_tag:
                     28
                                            link = target_a_tag['href']
Part 2
                     29
                                            complete_link = 'https://www.olx.co.id' + link
                     30
                     31
                                        # take text for item
                     32
                                        span_item_price = itemPrice.text
                     33
                                        price = span_item_price.split()[-1]
                     34
                     35
                                        # take text for year and kilometers range
                     36
                                        div_item_sub_title = itemSubTitle.text
                     37
                                        parts = div_item_sub_title.split(' - ')
                     38
                                        year = parts[0]
                     39
                                        kilometers_range = parts[1].split(' ')[0]
                     40
                                        # taek text for merk and car title
                     41
                     42
                                        div_item_title = itemTitle.text
                    43
                                        merk = div_item_title.split(' ')[0]
                                        title_car = div_item_title.split(' ')[-1]
                     45
                     46
                                        writer.writerow([complete_link, merk, title_car, year, kilometers_range, price ])
                     47
                     48
                                    scroll_count += 1
                     49
                        except Exception as e:
                            print(f"An error occurred: {e}")
Part 3
                     53 finally:
                            driver.quit()
```

# Part 3 -> exception dan menutup chromedriver

```
except Exception as e:
    print(f"An error occurred: {e}")

finally:
    # Quit the driver to close the browser session
    if 'driver' in locals():
        driver.quit()
```

# File csv hasil crawling



# Hasil file csv yang dibaca dengan pandas

	df = pd.read_csv('crawl_langsung.csv', deli	imiter=',')				
	df.head(10)					
1	0.0s					
	complete_link	merk	title_car	year	kilometers_range	price
)	https://www.olx.co.id/item/olx-autos-mitsubish	Mitsubishi	Mirage	2016	55.000-60.000	99.000.000
1	https://www.olx.co.id/item/olx-autos-toyota-si	Toyota	Sienta	2016	85.000-90.000	141.000.000
2	https://www.olx.co.id/item/olx-autos-honda-mob	Honda	Mobilio	2014	45.000-50.000	128.000.000
3	https://www.olx.co.id/item/olx-autos-daihatsu	Daihatsu	Xenia	2018	25.000-30.000	137.000.000
4	https://www.olx.co.id/item/suzuki-ignis-gx-ags	Suzuki	Ignis	2020	35.000-40.000	125.000.000
5	https://www.olx.co.id/item/toyota-agya-trd-mat	Toyota	Agya	2016	55.000-60.000	100.000.000
ŝ	https://www.olx.co.id/item/honda-civic-hb-15-t	Honda	Civic	2018	35.000-40.000	335.000.000
7	https://www.olx.co.id/item/olx-autos-suzuki-ig	Suzuki	Ignis	2018	85.000-90.000	110.000.000
3	https://www.olx.co.id/item/honda-crv-2022-hond	Honda	CR-V	2022	25.000-30.000	520.000.000
9	https://www.olx.co.id/item/jual-harga-cash-isu	Isuzu	Panther	2000	200.000-205.000	89.500.000

#### Referensi

https://www.olx.co.id/mobil-bekas c198

https://chromedriver.chromium.org/downloads/version-selection

https://chat.openai.com/share/758cc0d7-04f0-404e-b158-

49ede8aa7e9e

https://pypi.org/

https://www.freecodecamp.org/news/how-to-setup-virtual-

environments-in-python/



# Thank You

