SCRAPING OLX

"Mobil Bekas"

Habil Nasution

2108107010089

Tahapan Scraping OLX

01

Membuat environment

02

Menginstal Scrupy

03

Melakukan Scraping

04

Hasil Scraping

1. Membuat Environtment

```
noitusanx@noitusanx:~/studies$ python -m venv env-bigdata-course
noitusanx@noitusanx:~/studies$ source env-bigdata-course/bin/activate
(env-bigdata-course) noitusanx@noitusanx:~/studies$
```

Membuat environment dan mengaktifkan Python virtual environment

2. Menginstal Scrapy

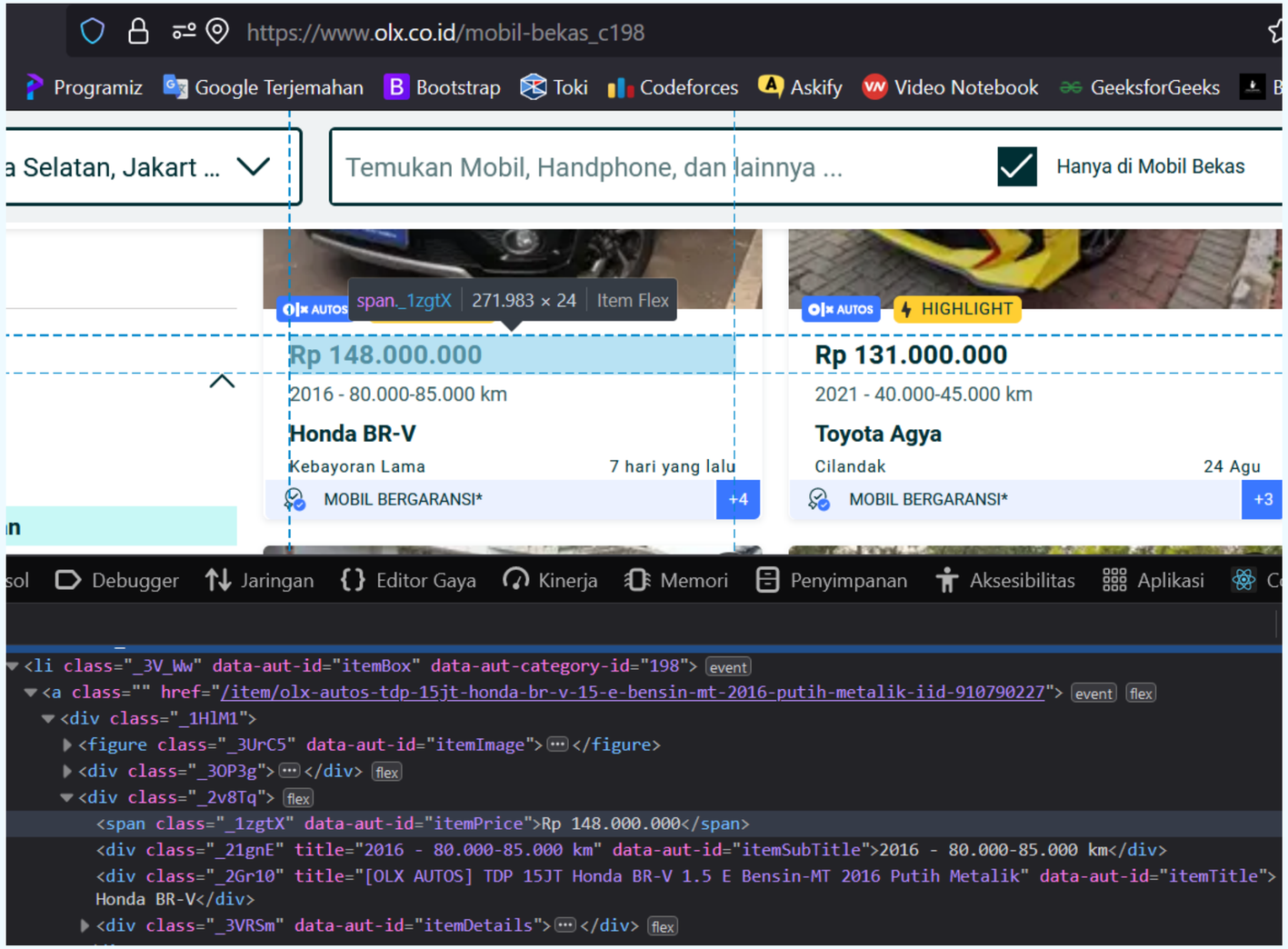
```
(env-bigdata-course) noitusanx@noitusanx:~/studies$ mkdir scrapping-olx
noitusanx@noitusanx:~/studies$ source env-bigdata-course/bin/activate
(env-bigdata-course) noitusanx@noitusanx:~/studies/scrapping-olx$ pip install scrapy
```

Membuat direktori scrapping-olx dan melakukan instalasi scrapy

```
(env-bigdata-course) noitusanx@noitusanx:~/studies/scrapping-olx$ scrapy startproject scrapping
```

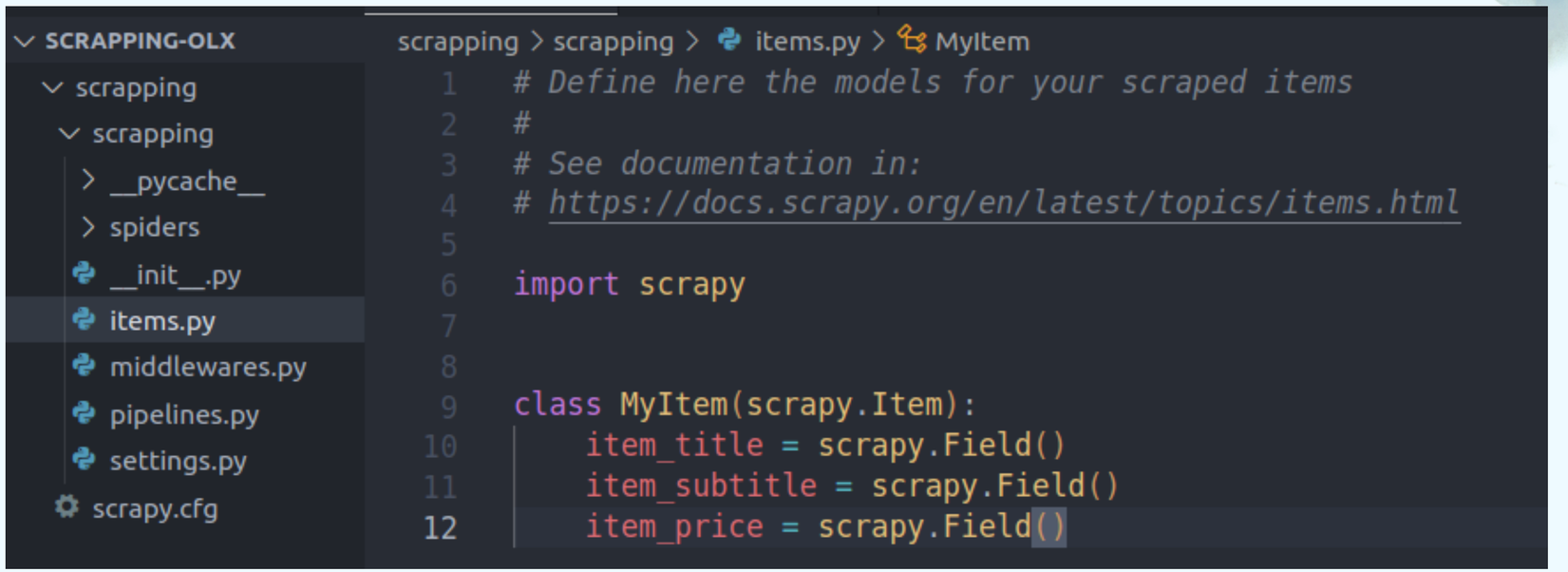
Membuat project Scrapy baru yang berisi kumpulan kode dan sumber daya yang digunakan Scrapy untuk menjelajahi situs web dan mengekstrak data

3. Melakukan Scraping



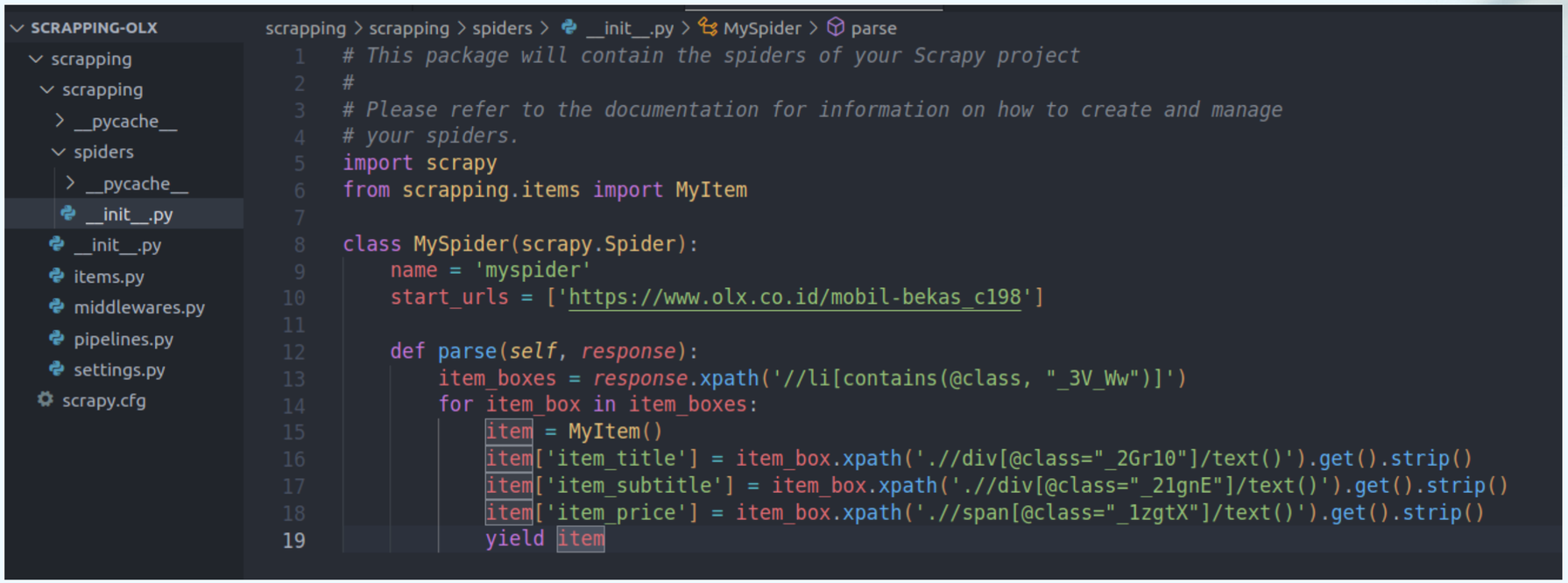
Melakukan inspect pada situs web olx mobil bekas untuk mendapatkan class dari data-data yang diperlukan seperti data harga, tahun, kilometer dan brand dari mobil bekas

3. Melakukan Scraping



Mendefinisikan field yang akan diambil yang berfungsi untuk menyimpan data yang diambil di situs web

3. Melakukan Scraping



Mengekstrak informasi daftar mobil bekas di situs olx yang terdiri dari brand, tahun, kilometer dan harga dari mobil bekas tersebut

3. Melakukan Scraping

```
(env-bigdata-course) noitusanx@noitusanx:~/studies/scrapping-olx/scrapping$ scrapy crawl myspider -o scrapping.json
2023-11-06 22:17:17 [scrapy.utils.log] INFO: Scrapy 2.11.0 started (bot: scrapping)
2023-11-06 22:17:17 [scrapy.utils.log] INFO: Versions: lxml 4.9.3.0, libxml2 2.10.3, cssselect 1.2.0, parsel 1.8.1, w3lib 2.1.2,
SSL 23.3.0 (OpenSSL 3.1.4 24 Oct 2023), cryptography 41.0.5, Platform Linux-6.2.0-36-generic-x86_64-with-glibc2.35
2023-11-06 22:17:17 [scrapy.addons] INFO: Enabled addons:
[]
2023-11-06 22:17:17 [asyncio] DEBUG: Using selector: EpollSelector
2023-11-06 22:17:17 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.asyncioreactor.AsyncioSelectorReactor
2023-11-06 22:17:17 [scrapy.utils.log] DEBUG: Using asyncio event loop: asyncio.unix_events._UnixSelectorEventLoop
2023-11-06 22:17:17 [scrapy.extensions.telnet] INFO: Telnet Password: b4b712a65f9bf8db
2023-11-06 22:17:17 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.memusage.MemoryUsage',
 'scrapy.extensions.feedexport.FeedExporter',
 'scrapy.extensions.logstats.LogStats']
2023-11-06 22:17:17 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'scrapping',
 'FEED_EXPORT_ENCODING': 'utf-8',
 'NEWSPIDER_MODULE': 'scrapping.spiders',
 'REQUEST_FINGERPRINTER_IMPLEMENTATION': '2.7',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['scrapping.spiders'],
 'TWISTED_REACTOR': 'twisted.internet.asyncioreactor.AsyncioSelectorReactor'}
2023-11-06 22:17:17 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
```

Menjelajahi situs web menggunakan Scrapy dan menyimpan hasilnya dalam bentuk JSON

4. Hasil Scraping

```
(env-bigdata-course) noitusanx@noitusanx:~/studies/scrapping-olx/scrapping$ cat scrapping.json
[
{"item_title": "Nissan X-Trail", "item_subtitle": "2017 - 105.000-110.000 km", "item_price": "Rp 225.000.000"},
{"item_title": "Wuling Almaz", "item_subtitle": "2019 - 25.000-30.000 km", "item_price": "Rp 204.000.000"},
{"item_title": "Nissan Livina", "item_subtitle": "2019 - 60.000-65.000 km", "item_price": "Rp 198.000.000"},
{"item_title": "Wuling Cortez", "item_subtitle": "2019 - 25.000-30.000 km", "item_price": "Rp 152.000.000"},
{"item_title": "Toyota Yaris", "item_subtitle": "2019 - 45.000-50.000 km", "item_price": "Rp 212.000.000"},
{"item_title": "Nissan X-Trail", "item_subtitle": "2015 - 135.000-140.000 km", "item_price": "Rp 177.000.000"},
{"item_title": "Nissan Serena", "item_subtitle": "2013 - 80.000-85.000 km", "item_price": "Rp 152.000.000"},
{"item_title": "Suzuki Ertiga", "item_subtitle": "2018 - 130.000-135.000 km", "item_price": "Rp 180.000.000"},
{"item_title": "Mercedes-Benz E300", "item_subtitle": "2011 - 80.000-85.000 km", "item_price": "Rp 295.000.000"},
{"item_title": "Daihatsu Rocky", "item_subtitle": "2021 - 30.000-35.000 km", "item_price": "Rp 203.000.000"},
{"item_title": "Toyota Kijang Innova", "item_subtitle": "2019 - 60.000-65.000 km", "item_price": "Rp 342.000.000"},
{"item_title": "Toyota Yaris", "item_subtitle": "2017 - 65.000-70.000 km", "item_price": "Rp 176.000.000"},
{"item_title": "Honda CR-V", "item_subtitle": "2019 - 40.000-45.000 km", "item_price": "Rp 375.000.000"},
{"item_title": "Toyota Yaris", "item_subtitle": "2021 - 95.000-100.000 km", "item_price": "Rp 247.000.000"},
{"item_title": "BMW 320i", "item_subtitle": "2016 - 40.000-45.000 km", "item_price": "Rp 317.000.000"},
{"item_title": "Honda HR-V", "item_subtitle": "2018 - 40.000-45.000 km", "item_price": "Rp 262.000.000"},
{"item_title": "Honda HR-V", "item_subtitle": "2018 - 40.000-45.000 km", "item_price": "Rp 262.000.000"},
{"item_title": "Toyota Yaris", "item_subtitle": "2015 - 100.000-105.000 km", "item_price": "Rp 160.000.000"},
{"item_title": "Mitsubishi Pajero", "item_subtitle": "2014 - 120.000-125.000 km", "item_price": "Rp 250.000.000"},
{"item_title": "Mitsubishi Xpander", "item_subtitle": "2019 - 5.000-10.000 km", "item_price": "Rp 196.000.000"},
{"item_title": "Daihatsu Xenia", "item_subtitle": "2015 - 155.000-160.000 km", "item_price": "Rp 106.500.000"},
{"item_title": "Daihatsu Ayla", "item_subtitle": "2017 - 50.000-55.000 km", "item_price": "Rp 102.000.000"},
{"item_title": "Honda Mobilio", "item_subtitle": "2014 - 120.000-125.000 km", "item_price": "Rp 124.000.000"},
```

```
(env-bigdata-course) noitusanx@noitusanx:~/studies/scrapping-olx/scrapping$ wc -l scrapping.json
41 scrapping.json
```

Menampilkan hasil dan jumlah data scraping