

# SCRAPY OLX



```
<p> Faiz Dzakirullail - 2008107010063</p>
```

# Preperation

First of all, we will create a `directory` and create a `virtual environment` for our `project`.

```
ozak@FaizLaptop:~$ mkdir olx
ozak@FaizLaptop:~$ cd olx
ozak@FaizLaptop:~/olx$ python3 -m venv myenv
ozak@FaizLaptop:~/olx$ ls
myenv
ozak@FaizLaptop:~/olx$ source myenv/bin/activate
(myenv) ozak@FaizLaptop:~/olx$ |
```

# Install SCRAPY

The next step is to install `scrapy` in that `directory`

```
(myenv) ozak@FaizLaptop:~/olx$ pip install scrapy
Collecting scrapy
  Using cached Scrapy-2.11.0-py2.py3-none-any.whl (286 kB)
Collecting protego>=0.1.15
  Using cached Protego-0.3.0-py2.py3-none-any.whl (8.5 kB)
```

# Create Project

using `command`:

`scrapy startproject [name your project]`

```
(myenv) ozak@FaizLaptop:~/olx$ scrapy startproject carscraper
New Scrapy project 'carscraper', using template directory '/home/ozak/olx/myenv/lib/python3.10/site-packages/scrapy/templates/project', created in:
  /home/ozak/olx/carscraper

You can start your first spider with:
  cd carscraper
  scrapy genspider example example.com
(myenv) ozak@FaizLaptop:~/olx$ ls
carscraper  myenv
(myenv) ozak@FaizLaptop:~/olx$ cd carscraper
(myenv) ozak@FaizLaptop:~/olx/carscraper$ ls
carscraper  scrapy.cfg
(myenv) ozak@FaizLaptop:~/olx/carscraper$ cd carscraper
(myenv) ozak@FaizLaptop:~/olx/carscraper/carscraper$ ls
__init__.py  items.py  middlewares.py  pipelines.py  settings.py  spiders
```

# Create Spider

If you want, you can create `spider` manually or using a basic template by running this `command`:

```
scrapy genspider [name] [link]
```

```
(myenv) ozak@FaizLaptop:~/olx/carscraper/carscraper/spiders$ scrapy genspider olxcarscraper https://www.olx.co.id/mobil-bekas_c198
Created spider 'olxcarscraper' using template 'basic' in module:
carscraper.spiders.olxcarscraper
```

# Customize Spider

In this stage we will code the **spider** according to the needs of this **project**

```
1 import scrapy
2
3
4 class CarscraperSpider(scrapy.Spider):
5     name = "carscraper"
6     allowed_domains = ["www.olx.co.id"]
7
8     def start_requests(self):
9         base_url = "https://www.olx.co.id/mobil-bekas_c198?page={}"
10        for page_number in range(51):
11            url = base_url.format(page_number)
12            yield scrapy.Request(url, self.parse)
13
14    def parse(self, response):
15        items = response.css('div._2v8Tq')
16
17        for item in items:
18            brand = item.css('._2Gr10::text').extract_first()
19
20            desc = item.css('._21gnE::text').extract_first()
21            split_desc = desc.split(' - ')
22            year = split_desc[0] if len(split_desc) > 0 else 'None'
23            km = split_desc[1] if len(split_desc) > 1 else 'None'
24
25            price = item.css('._1zgtX::text').extract_first()
26
```



# Save The Output

save the output that has been crawled using this  
command :

```
Scrapy crawl [spider name] -o output.json
```