Scrapy CrawlSpider Notlar

Terminal komtuları ve anlamları:

#genspider içerisindeki şablonları görmek için terminal komutu.

scrapy genspider -l

Available templates:
basic
crawl
csvfeed
xmlfeed

#şablon oluşturmak:

scrapy genspider deneme x

```
importscrapy


classDenemeSpider(scrapy.Spider):
    name = 'deneme'
    allowed_domains = ['x']
    start_urls = ['http://x/']

defparse(self,response):
pass

#Bu ve aşağıdaki kodun base class farklarına dikkat edin lütfen. Yazdığınız kodu çalış
tırmak için iki ayrı komut kullanacaksınız.
```

#buradaki “deneme” spider adını, x ise başlangış url’ini temsil eder.

scrapy genspider -t crawl trendy http://www.trendyol.com/

#-t template(şablon) temsil eder, crawl seçeceğimiz şablonu, trendy ise spider adını. En son ise başlangıç Url’ini yazıyoruz.

```
importscrapy
fromscrapy.linkextractorsimportLinkExtractor
fromscrapy.spidersimportCrawlSpider, Rule


classTrendySpider(CrawlSpider):
    name = 'trendy'
    allowed_domains = ['www.trendyol.com']
    start_urls = ['http://www.trendyol.com/']

    rules = (
        Rule(LinkExtractor(allow=r'Items/'), callback='parse_item', follow=True),
    )

defparse_item(self, response):
        item = {}
        #item['domain_id'] = response.xpath('//input[@id="sid"]/@value').get()
        #item['name'] = response.xpath('//div[@id="name"]').get()
        #item['description'] = response.xpath('//div[@id="description"]').get()
returnitem
```

Allowed domain kısmını düzeltmemiz gerektiğinin farkına varmamız gerekiyor.

allowed_domains = ['www.trendyol.com'] =⇒ yanlış kullanım.

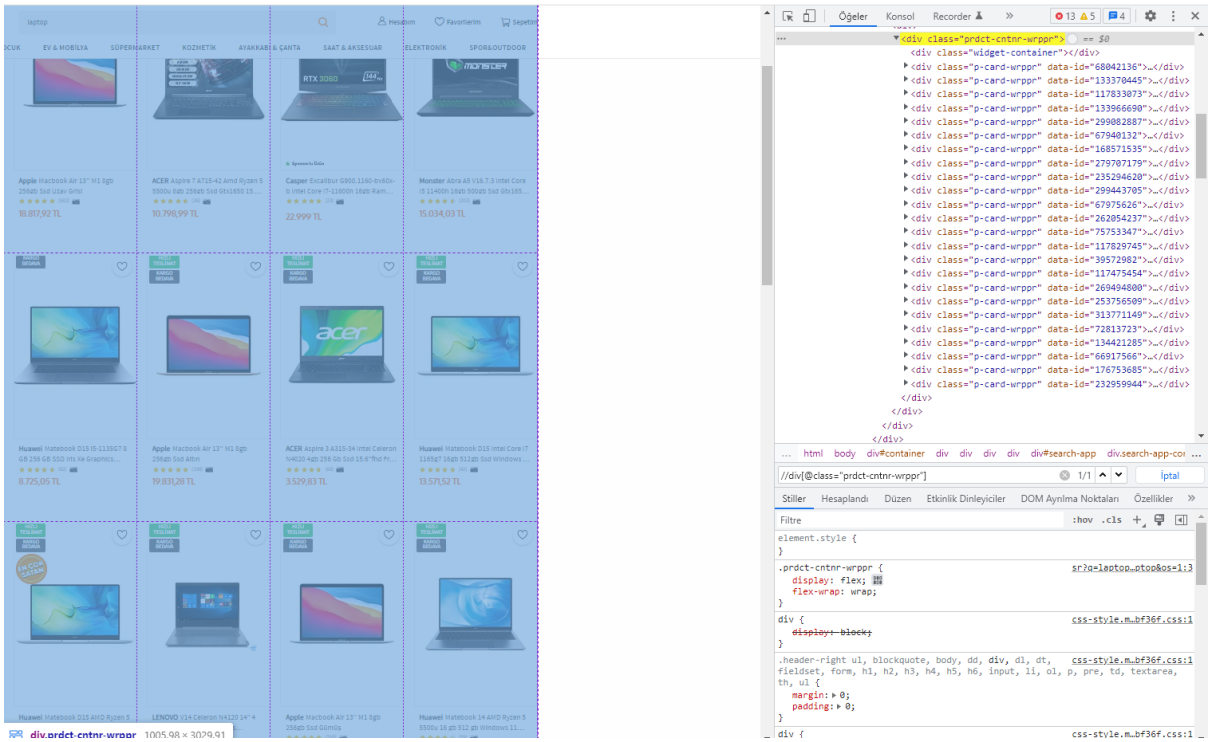
allowed_domains = ['trendyol.com'] =⇒ doğru kullanım.

Allowed domain kısmı bizim için çok önemli çünkü spider’ımızın domainin dışarısına çıkmasını istemiyoruz.

Rules

```
laptop_detail_link_rules = LinkExtractor(restrict_xpaths='//div[@class="prdct-cntnr-wr
ppr"]')
laptop_detail = Rule(laptop_detail_link_rules,
                     callback='parse_item',
                     follow=False)
rules = (
    laptop_detail,
)
```

laptop_detail adından bir local variable oluşturduktan sonra içerisine sayfa içerisinde takip etmesini istediğim linklerin xpathini verdim.



callback=’parse_item’

Sayfa içerisinden bir bilgi çıkartmak istiyorsak callback’i kullanmamız gerekiyor. Varsayılan olarak parse_item değerini alıyor.

follow=False

Eğer bu değer True olursa ziyaret edip bilgi çektiği sayfaların içerisinde LinkExtractor(restrict_xpaths='//div[@class="prdct-cntnr-wrppr"]') kuralına uyan linkler varsa onlarıda işleme dahil edecek.