Curso de Scrapy

1 - Intro Scrapy

Criar um arquivo e faça algo nessa estruta:

```
class BlogSpider(scrapy.Spider):
   name = 'blogspider'
   start_urls = ['https://blog.scrapinghub.com']

   def parse(self, response):
       for title in response.css('.post-header>h2'):
           yield {'title': title.css('a ::text').get()}

       for next_page in response.css('a.next-posts-link'):
           yield response.follow(next_page, self.parse
```

para executar o código

```
scrapy runspider my_project.py
```

2- Iniciar um projeto

```
scrapy startproject nome_projeto
```

3 - Criar um Spider

```
scrapy genspider nome_class_spider url_project
```

4 - Executar o Scrapy modo projeto

```
scrapy crawl nome_class_spider
```

5 - Executar e salvar o arquivo

```
scrapy crawl nome_class_spider -o arquivo.formato
```

6 - Comandos úteis

verificar o conteúdo html do elemento seletor

```
response.get()
response.extract()
response.extract_first()
```

Verificar o texto de uma pesquisa xpath

```
response.xpath('tag/.text()').extract() ou .extract_first()
```

Xpath contains

```
response.xpath('//tag[contains(@attr, '')]')
response.xpath('//tag[contains(text(), 'ipsum')]')
```

Scrapy shell - aceitar a página em pt-br

```
from scrapy import Request

req = Request('url', headers={'Accept-Language':'pt-br'})
fetch(req)
#  response ativado a partir do fetch
<p class="mume-header " id="response-ativado-a-partir-do-fetch"></p>
```

Scrapy FormResquest

```
from scrapy.http import FormRequest
def start_requests(self):
    url='http://imobiliariabelamorada.com.br/filtro/locacao/\
        apartamentos/pato-branco-pr/?busca=1'
    formdata={
        'cat1': '2.locacao',
        'cat3': '4.apartamentos',
        'cidade': '5362.pato-branco-pr',
        'valor':'',
        'cod':''
    }
    yield FormRequest(url, callback=self.parse, formdata=formdata, method='POST')
```

Alterar o robots.txt na class Spider:

```
    def start_requests(self):

        url = 'https://www.almeidaw.com.br'
        form_data = {
            'cat1': '2.locacao',
        }
        # comando que subscreve ROBOTSTXT_OBEY do settings
        meta = {'dont_obey_robotstxt':'False'}
        yield scrapy.FormRequest(url=url, callback=self.parse, formdata=form_data ,method=
```

FormResquest Shell

exemplo básico de funcionamento:

```
scrapy shell
```

```
url='http://imobiliariabelamorada.com.br/filtro/locacao/ \
    apartamentos/pato-branco-pr/?busca=1'
```

```
formdata={
    'cat1': '2.locacao',
    'cat3': '4.apartamentos',
    'cidade': '5362.pato-branco-pr',
    'valor':'',
    'cod':''
    }
```

```
fetch(scrapy.FormRequest(url, formdata=formdata, method='POST'))
```

Passar argumento category

```
scrapy crawl projeto -a category=nome_elemento
```

Ajustar o start_request

```
Nesse método é onde faz o callback para o parse, caso seja passado algum argumento,
é necessário fazer a verificação e/ou alteração do resquet.
```

Scrapy command line - List spiders

```
scrapy list
```