

实验三 网络信息抓取

【实验目的】

1. 了解 scraper 的使用与设置。
2. 掌握网络信息抓取的基本原理和实现方法。

【实验学时】

建议 2 学时

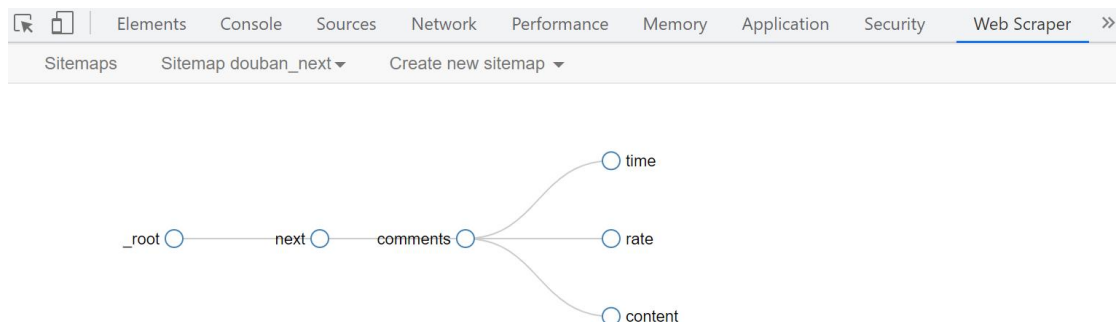
【实验环境配置】

- 1、Windows 环境
- 2、Chrome 浏览器
- 3、Scraper 插件

【实验原理】

基于 Chrome 浏览器插件 Scraper，自动抓取网络数据：

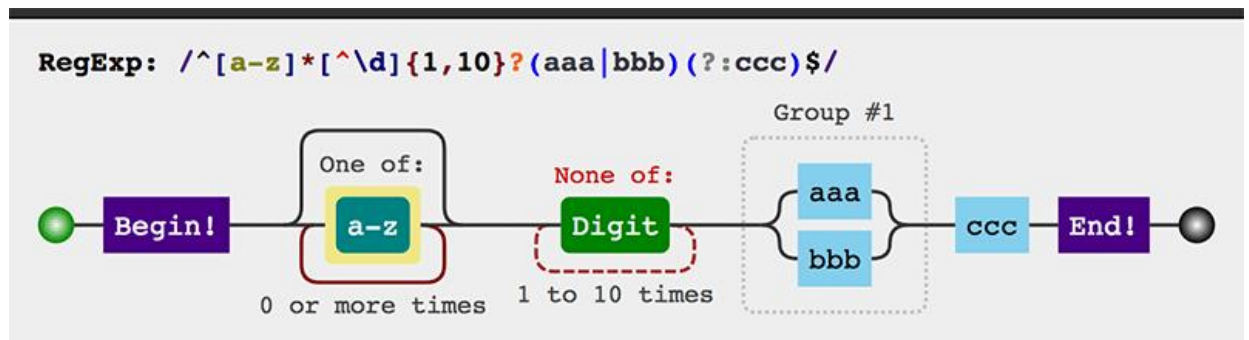
- 1、安装 scraper 插件，打开相关数据网页
- 2、使用 Scraper 设置相关抓取流程，并抓取数据



参考链接：

- 1、<https://www.webscraper.io/documentation>
- 2、<https://www.jianshu.com/p/1a6affc36d55>

3、正则关系式的使用



参考链接:

- 1、<https://www.runoob.com/regexp/regexp-tutorial.html>
- 2、<https://tool.oschina.net/uploads/apidocs/jquery/regexp.html>

【实验步骤】

1. 打开数据网页，指定相关网页数据

选取豆瓣电影页面抓取电影



豆瓣电影 Top 250

排名

1



电影名

肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]

导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins / ...

1994 / 美国 / 犯罪 剧情

评分

★★★★★ 9.7 2480139人评价

“希望让人自由。”

概要


☐ 我没看过的

2. 设置 Scraper，编写抓取流程

新建某个网页的抓取 Sitemap

豆瓣电影排行榜

豆瓣新片榜



沙丘 / 沙丘瀚战(港) / Dune: Part One

2021-09-03(威尼斯电影节) / 2021-09-11(多伦多电影节) / 2021-10-22(美国/中国大陆) / 蒂莫西·柴勒梅德 / 丽贝卡·弗格森 / 奥斯卡·伊萨克 / 戴夫·巴蒂斯塔 / 杰森·莫玛 / 乔什·布洛林 / 哈维尔·巴登 / 斯特兰·斯卡斯加德 / 夏洛特·莱斯特 / 三浦春马

★★★★★ 7.8 (316169人评价)



摩加迪沙 / 绝路狂逃(港) / 逃出摩加迪沙(台)

2021-07-28(韩国) / 金允石 / 赵寅成 / 许峻豪 / 具教焕 / 金素真 / 郑满植 / 金在华 / 朴庆惠 / 尹敬浩 / 韩国 / 柳昇完 / 121分钟 / 摩加迪沙 / 剧情 / 动作 / 李基贤 Giecheol Lee / 柳昇完 Seung-wan Ryoo /

分类排行榜

剧情 喜剧 动作 悬疑 惊悚 恐怖 同性 音乐 歌舞 历史 战争 犯罪 灾难 武侠 古装

一周口碑榜

1 美好的世界

创建新的Sitemap

Sitemaps Sitemap Create new sitemap

Search Sitemaps

Create Sitemap Import Sitemap

ID Domain

Web Scraper

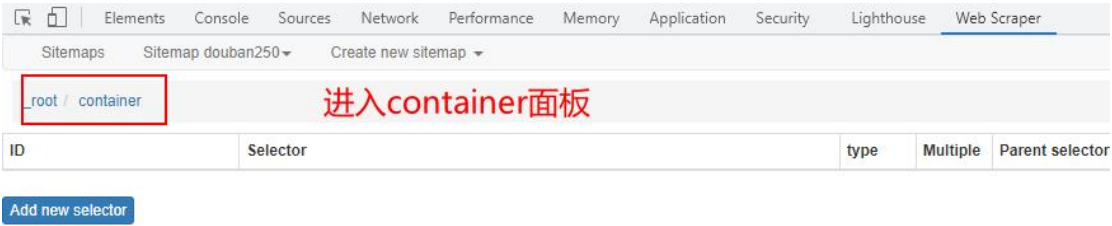
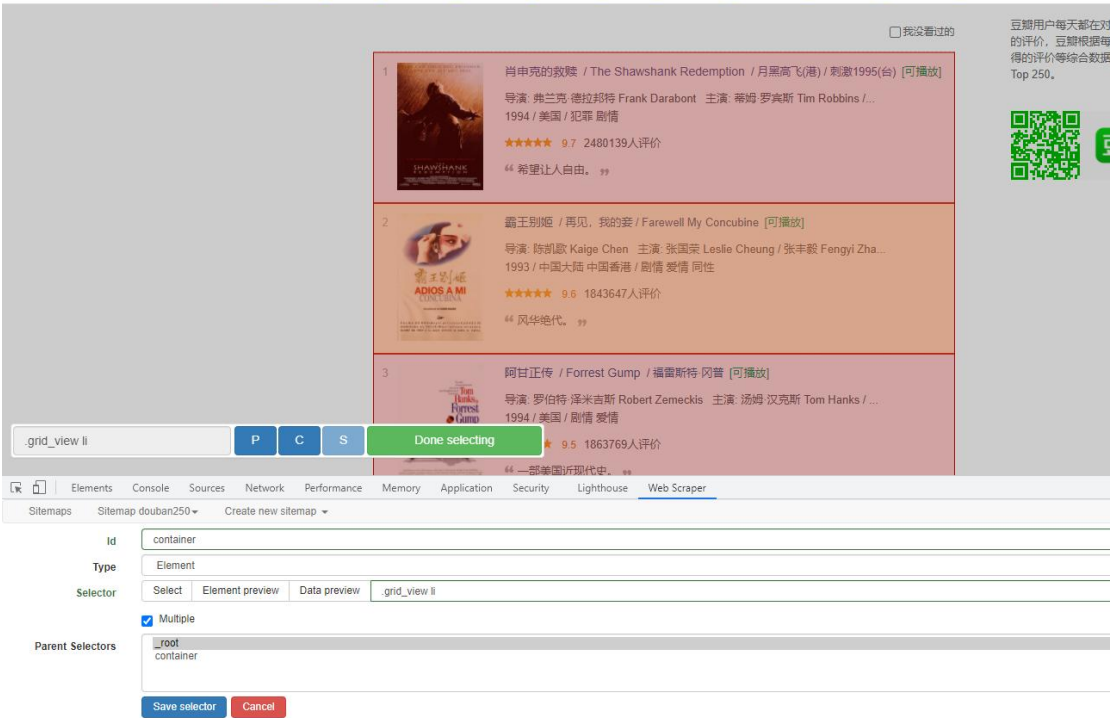
Create new sitemap

Sitemap name: douban250

Start URL: https://movie.douban.com/top250?start=[0-250,25]&filter=...


Create Sitemap

然后点击 Add new selector 选项



☐ 我没看过的

- 1




肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]

导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins / ...

1994 / 美国 / 犯罪 剧情

★★★★★ 9.7 2480139人评价

“希望让人自由。”
- 2




霸王别姬 / 再见，我的妾 / Farewell My Concubine [可播放]

导演: 陈凯歌 Kaige Chen 主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha...

1993 / 中国大陆 中国香港 / 剧情 爱情 同性

★★★★★ 9.6 1843647人评价

“风华绝代。”
- 3



阿甘正传 / Forrest Gump / 福雷斯特·冈普 [可播放]


导演: 罗伯特·泽米吉斯 Robert Zemeckis 主演: 汤姆·汉克斯 Tom Hanks / ...

1994 / 美国 / 剧情 爱情

★ 9.5 1863769人评价


“一部美国近现代史。”

1




肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995
导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins / ...
1994 / 美国 / 犯罪 剧情
★★★★★ 9.7 2480139人评价
“希望让人自由。”

2



霸王别姬 / 再见，我的妾 / Farewell My Concubine [可播放]
导演: 陈凯歌 Kaige Chen 主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha
1993 / 中国大陆 中国香港 / 剧情 爱情 同性
★★★★★ 9.6 1843647人评价
“风华绝代。”

3



阿甘正传 / Forrest Gump / 福雷斯特·冈普 [可播放]
导演: 罗伯特·泽米吉斯 Robert Zemeckis 主演: 汤姆·汉克斯 Tom Hanks / ...
1994 / 美国 / 剧情 爱情
★★★★★ 9.5 1863769人评价
“一部美国近现代史。”

P

C

S

Done selecting

Elements Console Sources Network Performance Memory Application Security Lighthouse Web Scraper

Sitemaps Sitemap douban250 Create new sitemap

Id

num

Type

Text

Selector

Select Element preview Data preview

☐ Multiple

Regex

Parent Selectors

_root

container

Save selector

Cancel

The screenshot shows the Web Scraper interface with a list of movies and a configuration panel for a selector.

Movie List:

1. 肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]
导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins / ...
1994 / 美国 / 犯罪 剧情
★★★★★ 9.7 2480139人评价
64 希望让人自由。 99
2. 霸王别姬 / 再见，我的妾 / Farewell My Concubine [可播放]
导演: 陈凯歌 Kaige Chen 主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha...
1993 / 中国大陆 中国香港 / 剧情 爱情 同性
★★★★★ 9.6 1843647人评价
64 风华绝代。 99
3. 阿甘正传 / Forrest Gump / 福雷斯特 冈普 [可播放]
导演: 罗伯特·泽米吉斯 Robert Zemeckis 主演: 汤姆·汉克斯 Tom Hanks / ...
1994 / 美国 / 剧情 爱情
★★★★★ 9.5 1863769人评价
64 一部美国近现代史。 99

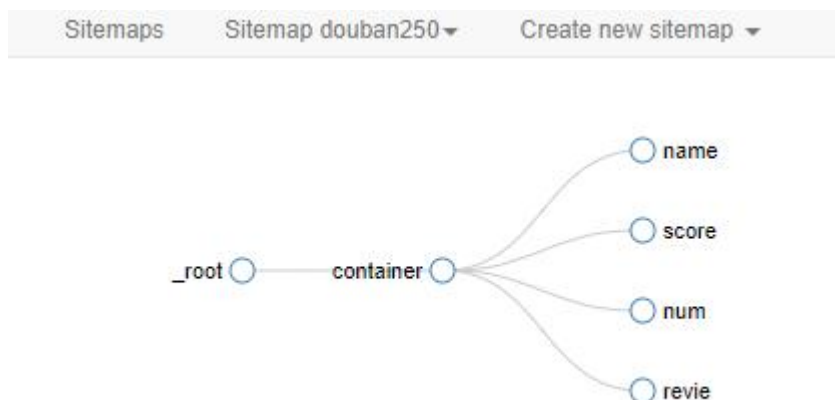
Selector Configuration:

- Id: revie
- Type: Text
- Selector: Select
- Multiple: ☐
- Regex:
- Parent Selectors: _root, container
- Buttons: Save selector, Cancel

每一个电影对应一个 li 标签，然后再在 container 选择器中定义四个元素 element 的选择器，每一个电影中分别只有一个 name、score、num、revie，所以不需要选择 multiple。

ID	Selector	type	Multiple	Parent selectors	Actions
name	span.title:nth-of-type(1)	SelectorText	no	container	Element preview Data preview Edit
score	span.rating_num	SelectorText	no	container	Element preview Data preview Edit
num	em	SelectorText	no	container	Element preview Data preview Edit
revie	span.inq	SelectorText	no	container	Element preview Data preview Edit

如下是 select graph 图标



点击 scrapre, 开启抓取 z



参考链接：

- Scrapy 框架分为几个部分，分别为 Scrapy Engine(引擎)、Scheduler(调度器)、Downloader(下载器)、Spiders(爬虫)、Item Pipeline(管道)、Downloader Middlewares(下载中间件)、Spider Middlewares(爬虫中间件)。

先进行 scrapy 的安装 pip install scrapy, 创建项目 “Douban”, 之后进入 Douban 目录下输入 scrapy genspider 爬虫名 域名, 就会生成

```
Windows PowerShell
PS F:\PythonFile\爬虫\Scrapy> scrapy --help
Scrapy 2.5.1 - no active project

Usage:
scrapy <command> [options] [args]

Available commands:
bench          Run quick benchmark test
commands
fetch          Fetch a URL using the Scrapy downloader
genspider      Generate new spider using pre-defined templates
runspider      Run a self-contained spider (without creating a project)
settings       Get settings values
shell          Interactive scraping console
startproject   Create new project
version        Print Scrapy version
view           Open URL in browser, as seen by Scrapy

[ more ]      More commands available when run from project directory

Use "scrapy <command> -h" to see more info about a command
PS F:\PythonFile\爬虫\Scrapy> scrapy startproject Douban
Scrapy 2.5.1 - no active project




Unknown command: startproject

Use "scrapy" to see available commands
PS F:\PythonFile\爬虫\Scrapy> scrapy startproject Douban
New Scrapy project 'Douban', using template directory 'd:\python\lib\site-packages\scrapy\templates\project', created in:
F:\PythonFile\爬虫\Scrapy\Douban

You can start your first spider with:
cd Douban
scrapy genspider example example.com
PS F:\PythonFile\爬虫\Scrapy>
```



```
Windows PowerShell
PS F:\PythonFile\爬虫\Scrapy\Douban> scrapy genspider douban movie.douban.com
Created spider 'douban' using template 'basic' in module:
Douban.spiders.douban
PS F:\PythonFile\爬虫\Scrapy\Douban>
```

名称	修改日期	类型	大小
 _pycache_	2021/11/12 16:48	文件夹	
 _init_.py	2021/11/12 16:15	Python 源文件	1 KB
 douban.py	2021/11/12 16:48	Python 源文件	1 KB

The screenshot shows a code editor with a file explorer on the left and a code editor on the right. The file explorer shows a project structure with a 'venv' folder selected. The code editor shows the 'items.py' file with Python code for a Scrapy spider. The code includes a comment about documentation, an import statement for 'scrapy', and a class definition for 'DoubanItem' which inherits from 'scrapy.Item'. The class has fields for 'name', 'score', 'num', and 'review'.

```

3      # See documentation in:
4      # https://docs.scrapy.org/en/latest/topics/items.html
5
6      import scrapy
7
8
9      class DoubanItem(scrapy.Item):
10
11         # define the fields for your item here like:
12         # name = scrapy.Field()
13
14         name = scrapy.Field() # 电影名
15         score = scrapy.Field() # 分数
16         num = scrapy.Field() # 排名
17         review = scrapy.Field() # 概况
18
19     pass

```

The screenshot shows a code editor with a file explorer on the left and a code editor on the right. The file explorer shows a project structure with a 'Douban' folder containing 'spiders' and 'scrapy.cfg'. The 'spiders' folder is expanded, showing 'douban.py' selected. The code editor shows the following Python code:

```
1 import scrapy
2
3
4 class DoubanSpider(scrapy.Spider):
5     # 爬虫名
6     name = 'douban'
7     # 允许的域名
8     allowed_domains = ['movie.douban.com']
9     # 入口url
10    start_urls = ['https://movie.douban.com/top250']
11
12    def parse(self, response):
13        pass
14
```

在进行数据爬取前，首先设置一些网络代理，在 settings.py 文件中修改 USER_AGENT 变量

```
# Crawl responsibly by identifying yourself (and your website) on the User-agent
USER_AGENT = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:70.0) Gecko/20100101 Firefox/70.0'

# Obey robots.txt rules
```

Douban.py 中的内容，对爬取到的网页数据进行细分选取，通过 xpath 选择其中的电影名字、简介、评价、排名、分数。并保存为 DoubanItem 对象 Item，最后通过 yield 将 item 对象从 Spiders 返回到 Item 管道。并且通过拼接 url 形成下一页的 Request 请求。

```
from scrapy.Douban.Douban.items import DoubanItem
class DoubanSpider(scrapy.Spider):
    # 爬虫名
    name = 'douban'
    # 允许的域名
    allowed_domains = ['movie.douban.com']
    # 入口url
    start_urls = ['https://movie.douban.com/top250']

    def parse(self, response):
        # 首先抓取电影列表
        movie_list = response.xpath("//ol[@class='grid_view']/li")
        for selector in movie_list:
            # 遍历每个电影列表，从其中精准抓取所需要的信息并保存为item对象
            item = DoubanItem()
            item['num'] = selector.xpath("//div[@class='pic']/em/text()").extract_first()
            item['name'] = selector.xpath("//span[@class='title']/text()").extract_first()
            text = selector.xpath("//div[@class='bd']/p[1]/text()").extract()
            intro = ""
            for s in text: # 将简介放到一个字符串
                intro += "".join(s.split()) # 去掉空格
            item['review'] = intro
            item['score'] = selector.css('.rating_num::text').extract_first()
            # item['comments'] = selector.xpath("//div[@class='star']/span[4]/text()").extract_first()
            # item['describe'] = selector.xpath("//span[@class='inq']/text()").extract_first()
            # print(item)
            yield item # 将结果item对象返回给Item管道

        # 爬取网页中的下一个页面url信息
        next_link = response.xpath("//span[@class='next']/a[1]/@href").extract_first()
        if next_link:
            next_link = "https://movie.douban.com/top250" + next_link
            print(next_link)
            yield scrapy.Request(next_link, callback=self.parse) # 将Request请求提交给调度器
```