Day06回顾

多线程爬虫

- **使用流程**

```
# 1、URL队列
q.put(url)
# 2、线程事件函数
while True:
    if not url_queue.empty():
        ...get()、请求、解析
    else:
        break
# 创建并启动线程
t_list = []
for i in range(5):
    t = Thread(target=parse_page)
    t_list.append(t)
    t.start()
# 阻塞等待回收线程
for i in t_list:
    i.join()
```

json模块

- **json转python**

```
变量名 = json.loads(res.text))
```

- **python转json（保存为json文件）**

```
# 保存所抓取数据为json数据
with open(filename,'a') as f:
    json.dump(字典/列表/元组,f,ensure_ascii=False)
```

selenium+phantomjs/chrome/firefox

- **特点**

```
1、简单，无需去详细抓取分析网络数据包，使用真实浏览器
2、需要等待页面元素加载，需要时间，效率低
```

- **使用流程**

```
from selenium import webdriver

# 创建浏览器对象
browser = webdriver.Firefox()
browser.get('https://www.jd.com/')

# 查找节点
node = browser.find_element_by_xpath('')
node.send_keys('')
node.click()

# 获取节点文本内容
content = node.text

# 关闭浏览器
browser.quit()
```

- **设置无界面模式（chromedriver | firefox）**

```
options = webdriver.ChromeOptions()
options.add_argument('--headless')

browser = webdriver.Chrome(options=options)
browser.get(url)
```

京东爬虫

- **执行JS脚本,把进度条拉到最下面**

```
1、js脚本
browser.execute_script(
'window.scrollTo(0,document.body.scrollHeight)'
)
2、利用节点对象的text属性获取当前节点及后代节点的文本内容,想办法处理数据
```

scrapy框架

- 五大组件

```
引擎（Engine）
爬虫程序（Spider）
调度器（Scheduler）
下载器（Downloader）
管道文件（Pipeline）
# 两个中间件
下载器中间件（Downloader Middlewares）
蜘蛛中间件（Spider Middlewares）
```

- 工作流程

```
1、Engine向Spider索要URL,交给Scheduler入队列
2、Scheduler处理后出队列,通过Downloader Middlewares交给Downloader去下载
3、Downloader得到响应后,通过Spider Middlewares交给Spider
4、Spider数据提取：
   1、数据交给Pipeline处理
   2、需要跟进URL,继续交给Scheduler入队列，依次循环
```

- 常用命令

```
# 创建爬虫项目
scrapy startproject 项目名

# 创建爬虫文件
cd 项目文件夹
scrapy genspider 爬虫名 域名

# 运行爬虫
scrapy crawl 爬虫名
```

- scrapy项目目录结构

```
Baidu
├── Baidu               # 项目目录
│   ├── items.py        # 定义数据结构
│   ├── middlewares.py  # 中间件
│   ├── pipelines.py    # 数据处理
│   ├── settings.py     # 全局配置
│   └── spiders
│       ├── baidu.py    # 爬虫文件
└── scrapy.cfg          # 项目基本配置文件
```

- settings.py全局配置

```
1、USER_AGENT = 'Mozilla/5.0'
2、ROBOTSTXT_OBEY = False
3、CONCURRENT_REQUESTS = 32
4、DOWNLOAD_DELAY = 1
5、DEFAULT_REQUEST_HEADERS={}
6、ITEM_PIPELINES={'项目目录名.pipelines.类名':300}
```

Day07笔记

selenium补充

切换页面

1、适用网站

```
页面中点开链接出现新的页面，但是浏览器对象browser还是之前页面的对象
```

2、应对方案

```
# 获取当前所有句柄（窗口）
all_handles = browser.window_handles
# 切换到新的窗口
browser.switch_to_window(all_handles[1])
```

3、民政部网站案例

3.1 目标: 将民政区划代码爬取到数据库中，按照层级关系（分表 -- 省表、市表、县表）

3.2 数据库中建表

```
# 建库
create database govdb charset utf8;
use govdb;
# 建表
create table province(
p_name varchar(20),
p_code varchar(20)
)charset=utf8;
create table city(
c_name varchar(20),
c_code varchar(20),
c_father_code varchar(20)
)charset=utf8;
create table county(
x_name varchar(20),
x_code varchar(20),
x_father_code varchar(20)
)charset=utf8;
```

3.3 思路

```
1、selenium+Chrome打开一级页面，并提取二级页面最新链接
2、增量爬取：和数据库version表中进行比对，确定之前是否爬过（是否有更新）
3、如果没有更新，直接提示用户，无须继续爬取
4、如果有更新，则删除之前表中数据，重新爬取并插入数据库表
5、最终完成后：断开数据库连接，关闭浏览器
```

3.4 代码实现

```

```

3.5 SQL命令练习

```
# 1. 查询所有省市县信息（多表查询实现）

# 2. 查询所有省市县信息（连接查询实现）

```

Web客户端验证

弹窗中的用户名和密码如何输入?

```
不用输入，在URL地址中填入就可以
```

示例: 爬取某一天笔记

```
from selenium import webdriver

url = 'http://tarenacode:code_2013@code.tarena.com.cn/AIDCode/aid1903/12-spider/spider_day07_note.zip'
browser = webdriver.Chrome()
browser.get(url)
```

scrapy框架

- 创建爬虫项目步骤

```
1、新建项目 ： scrapy startproject 项目名
2、cd 项目文件夹
3、新建爬虫文件 ： scrapy genspider 文件名 域名
4、明确目标(items.py)
5、写爬虫程序(文件名.py)
6、管道文件(pipelines.py)
7、全局配置(settings.py)
8、运行爬虫 ： scrapy crawl 爬虫名
```

- pycharm运行爬虫项目

```
1、创建begin.py(和scrapy.cfg文件同目录)
2、begin.py中内容:
    from scrapy import cmdline
    cmdline.execute('scrapy crawl maoyan'.split())
```

小试牛刀

- 目标

```
打开百度首页，把 '百度一下，你就知道' 抓取下来，从终端输出
```

- 实现步骤

1. **创建项目Baidu 和 爬虫文件baidu**

```
1、scrapy startproject Baidu
2、cd Baidu
3、scrapy genspider baidu www.baidu.com
```

2. **编写爬虫文件baidu.py，xpath提取数据**

```python
# -*- coding: utf-8 -*-
import scrapy

class BaiduSpider(scrapy.Spider):
    name = 'baidu'
    allowed_domains = ['www.baidu.com']
    start_urls = ['http://www.baidu.com/']

    def parse(self, response):
        result = response.xpath('/html/head/title/text()').extract_first()
        print('*'*50)
        print(result)
        print('*'*50)

```

3. **全局配置settings.py**

```
USER_AGENT = 'Mozilla/5.0'
ROBOTSTXT_OBEY = False
DEFAULT_REQUEST_HEADERS = {
    'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
    'Accept-Language': 'en',
}
```

4. **创建run.py（和scrapy.cfg同目录）**

```
from scrapy import cmdline

cmdline.execute('scrapy crawl baidu'.split())
```

5. **启动爬虫**

```
直接运行 run.py 文件即可
```

思考运行过程

猫眼电影案例

- 目标

```
URL: 百度搜索 -> 猫眼电影 -> 榜单 -> top100榜
内容:电影名称、电影主演、上映时间
```

- 实现步骤

1. **创建项目和爬虫文件**

```
# 创建爬虫项目
scrapy startproject Maoyan
# 创建爬虫文件
cd Maoyan
scrapy genspider maoyan maoyan.com
```

2. **定义要爬取的数据结构（items.py）**

```
name = scrapy.Field()
star = scrapy.Field()
time = scrapy.Field()
```

3. **编写爬虫文件（maoyan.py）**

```
1、基准xpath,匹配每个电影信息节点对象列表
    dd_list = response.xpath('//dl[@class="board-wrapper"]/dd')
2、for dd in dd_list:
    电影名称 = dd.xpath('./a/@title')
    电影主演 = dd.xpath('.//p[@class="star"]/text()')
    上映时间 = dd.xpath('.//p[@class="releasetime"]/text()')
```

代码实现一

```

```

代码实现二

```
1
```

代码实现三

```
1
```

3. **定义管道文件（pipelines.py）**

```
1
```

5. **全局配置文件（settings.py）**

```
USER_AGENT = 'Mozilla/5.0'
ROBOTSTXT_OBEY = False
DEFAULT_REQUEST_HEADERS = {
    'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
    'Accept-Language': 'en',
}
ITEM_PIPELINES = {
   'Maoyan.pipelines.MaoyanPipeline': 300,
}
```

6. **创建并运行文件（run.py）**

```
from scrapy import cmdline
cmdline.execute('scrapy crawl maoyan'.split())
```

知识点汇总

- **节点对象.xpath('')**

```
1、列表,元素为选择器 ['<selector data='A'>]
2、列表.extract() ：序列化列表中所有选择器为Unicode字符串 ['A','B','C']
3、列表.extract_first() 或者 get() :获取列表中第1个序列化的元素(字符串)
```

- **pipelines.py中必须由1个函数叫process_item**

```
def process_item(self,item,spider):
    return item ( * 此处必须返回 item )
```

- **日志变量及日志级别(settings.py)**

```
# 日志相关变量
LOG_LEVEL = ''
LOG_FILE = '文件名.log'

# 日志级别
5 CRITICAL : 严重错误
4 ERROR    : 普通错误
3 WARNING  : 警告
2 INFO     : 一般信息
1 DEBUG    : 调试信息
# 注意：只显示当前级别的日志和比当前级别日志更严重的
```

- **管道文件使用**

```
1、在爬虫文件中为items.py中类做实例化，用爬下来的数据给对象赋值
    from ..items import MaoyanItem
    item = MaoyanItem()
2、管道文件（pipelines.py）
3、开启管道（settings.py）
    ITEM_PIPELINES = { '项目目录名.pipelines.类名':优先级 }
```

数据持久化存储(MySQL)

实现步骤

```
1、在setting.py中定义相关变量
2、pipelines.py中导入settings模块
    def open_spider(self,spider):
        # 爬虫开始执行1次,用于数据库连接
    def close_spider(self,spider):
        # 爬虫结束时执行1次,用于断开数据库连接
3、settings.py中添加此管道
    ITEM_PIPELINES = {'':200}

# 注意 ： process_item() 函数中一定要 return item ***
```

练习

把猫眼电影数据存储到MySQL数据库中

保存为csv、json文件

- 命令格式

```
scrapy crawl maoyan -o maoyan.csv
scrapy crawl maoyan -o maoyan.json
```

盗墓笔记小说抓取案例（三级页面）

- 目标

```
# 抓取目标网站中盗墓笔记1-8中所有章节的所有小说的具体内容，保存到本地文件
1、网址 ： http://www.daomubiji.com/
```

- 准备工作xpath

```
1、一级页面xpath（此处响应做了处理）： //ul[@class="sub-menu"]/li/a/@href
2、二级页面xpath: /html/body/section/div[2]/div/article
  基准xpath ： //article
3、三级页面xpath: response.xpath('//article[@class="article-content"]//p/text()').extract()
```

- 项目实现

1. **创建项目及爬虫文件**

```
创建项目 ： Daomu
创建爬虫 ： daomu  www.daomubiji.com
```

2. 定义要爬取的数据结构（把数据交给管道）

```
import scrapy

class DaomuItem(scrapy.Item):
    # 卷名
    juan_name = scrapy.Field()
    # 章节数
    zh_num = scrapy.Field()
    # 章节名
    zh_name = scrapy.Field()
    # 章节链接
    zh_link = scrapy.Field()
    # 小说内容
    zh_content = scrapy.Field()
```

3. **爬虫文件实现数据抓取**

```

```

4. **管道文件实现数据处理**

```
1
```

图片管道(360图片抓取案例)

- 目标

```
1  www.so.com -> 图片 -> 美女
```

- 抓取网络数据包

```
1  2、F12抓包,抓取到json地址 和 查询参数(QueryString)
2        url = 'http://image.so.com/zj?ch=beauty&sn={}&listtype=new&temp=1'.format(str(sn))
3        ch: beauty
4        sn: 90
5        listtype: new
6        temp: 1
```

- 项目实现

1. 创建爬虫项目和爬虫文件

```
1
```

2. 定义要爬取的数据结构

```
1
```

3. 爬虫文件实现图片链接抓取

```
1
```

4. **管道文件（pipelines.py）**

```
1
```

5. **设置settings.py**

```
1
```

6. **创建bigin.py运行爬虫**

```
1
```

今日作业

```
1、把今天内容过一遍
2、Daomu错误调一下(看规律,做条件判断)
3、腾讯招聘尝试改写为scrapy
   response.text ： 获取页面响应内容
```