《Python程序设计》大作业

实验报告

实验题目： 高校就业信息网爬取

姓名： 林亦旸

学号： 2020212142

日期： 2022/1/07

1. **需求分析**

本次作业分为两个部分进行。第一部分，是爬取北邮就业信息网，电子科技大学研究生就业信息网和西电就业信息网的招聘信息。第二部分是对爬取下来的数据做数据处理，得到最受关注几条的招聘信息。

首先，对需要爬取的网站进行页面分析。北邮就业信息网是动态页面，需要用selenium模拟用户打开页面或用json文件爬取。发现北邮就业信息网的翻页是通过脚本控制的，不能靠修改url实现翻页。而我们需要进入招聘信息二级页面进行爬取，这样会导致翻页无法正常进行。除非用一个浏览器爬取目录，另外一个爬取二级页面。但这样做多占用系统资源。故在源代码中查找读取json文件的脚本，找到获取json的方法。用json文件爬取，不仅速度快，内容也好解析，对网站服务器造成的负载也小。因此对于成电就业网我也采取了这种爬取方式。西电就业网因为在之前作业中爬取过，直接对之前的代码做修改即可完成任务。

1. **模型设计**

本次作业分为两个部分进行

1. **系统设计**

分为两个脚本进行编写，一个是爬虫脚本，一个是数据处理模块脚本

1. 对爬虫的设计

采用Scrapy框架，结合selenium，json模块进行爬取。

item.py中设置四个item：

*class* Jobtop10Item(*scrapy*.*Item*):

    job\_title = scrapy.Field()  # 招聘帖子的主题

    job\_date  = scrapy.Field()  # 招聘帖子的发帖时间

    job\_views = scrapy.Field()  # 招聘帖子的浏览次数

    job\_nums  = scrapy.Field()  # 招聘帖子的职位数量

pipeline.py中写好了对item的处理，将其写入对应csv

setting.py中设置了多爬虫启动和中间件

middlewares.py中，设计下载中间件。这边和之前作业的设计不一样。之前的作业在启动多爬虫时会多开几个没有执行任务的浏览器。这是因为scrapy对每个spider都实例化了一个下载中间件类。按照之前的作业代码，n个爬虫会开启n\*n个浏览器，浪费系统资源。并且之前作业中爬虫get\_driver调用的是类方法，这样做最终调用的都是最后一个被实例化的下载中间件类。可不可以在传给spider的response里面找到scrapy为这个spider实例化的中间件呢？可以，但是需要运行时观察变量，找到自定义的中间件在中间件列表里的下标。这样写代码不利于维护，因为一旦改变中间件顺序，就得改spider。因此，我认为应该在spider初始化的时候进行driver的声明，变成spider的类对象，下载中间件使用spider.driver.get()方法进行页面打开，并返回response。具体实现如下：

spider.py中的实现

@*classmethod*

*def* from\_crawler(*cls*, *crawler*, \**args*, \*\**kwargs*):

# 在crawler初始化的时候被调用一次

        spider = *super*(XDUJOBSpider, *cls*).from\_crawler(

            crawler, \*args, \*\*kwargs)

        # 创建浏览器

        spider.driver = webdriver.Firefox()

        crawler.signals.connect(spider.spider\_closed,

*signal*=signals.spider\_closed)

        return spider

*def* spider\_closed(*self*, *spider*):

        spider.driver.close()  # 关闭浏览器

        print("==========爬虫结束！==========")

        spider.logger.info('Spider closed:%s', spider.name)

middlewares.py中的实现：（只需要实现process\_request函数）

*def* process\_request(*self*, *request*, *spider*):

        re\_method = request.meta.get('method', 'default')

        if re\_method=='default': #selenium爬取

            spider.driver.get(request.url)

            time.sleep(3)  #等待页面加载完成

            return scrapy.http.HtmlResponse(*url*=request.url,

*body*=spider.driver.page\_source.encode('utf-8'),

*encoding*='utf-8', *request*=request, *status*=200)

        elif re\_method == 'stastic':

            return None #静态爬取，交给其他中间件处理

这边为了配合静态爬取json文件，在爬虫发出request的时候会在meta里传一个参数method，代表静态爬取还是selenium爬取

爬虫的实现（代码中包括了注释，不赘述功能）：

XDUJOB\_spider.py

*def* parse(*self*, *response*): #解析爬取的内容

        item = Jobtop10Item()

        #下一页的url

        next\_page\_href=response.css('li[class="next"]>a::attr(href)').extract()

        # 末尾页的url

        last\_page\_href=response.css('li[class="last"]>a::attr(href)').extract()

#确认现在页面不是最后一页

#在尾页页的时候页面发生改变，尾页的链接在li[class="last "]下

        #按上面代码无法获取，由于只有一页，直接用len判断

        if response.url!=last\_page\_href and len(last\_page\_href) :

*self*.xidian\_next\_page = 'https://job.xidian.edu.cn' + next\_page\_href[0]

        else:

*self*.xidian\_next\_page = ''

        c\_page\_url\_list = response.css('ul[class="infoList"]>li:nth-child(1)>a')

        for job in c\_page\_url\_list:  # 循环打开和解析每个详情页

            driver=*self*.driver #用爬虫定义的driver

            driver.get('https://job.xidian.edu.cn' + job.css('a::attr(href)').extract()[0])

            time.sleep(2)#等待2秒钟

            item['job\_title'] = [driver.find\_element(

'css selector', 'div[class="title-message"]>h5').text][0]

            date\_text = driver.find\_element('css selector', 'div[class="share"]>ul>li:nth-child(1)').text

            date\_text=date\_text[date\_text.find('：') + 1:]

            if datetime.strptime(date\_text,'%Y-%m-%d %H:%M')

<datetime.strptime('2021-09-01 00:00','%Y-%m-%d %H:%M'):

*self*.xidian\_next\_page=''

 #到了设定时间，停止爬取。刚好可以利用下一页的url作为结束标志

                break

            item['job\_date'] = [date\_text][0]

            views\_text = driver.find\_element('css selector', 'div[class="share"]>ul>li:nth-child(2)').text

            item['job\_views'] = [views\_text[views\_text.find('：') + 1:]][0]

            item['job\_nums']=1

            yield item

        #处理完列表页的所有二级页面后，继续打开下一页进行抓取

        if *self*.xidian\_next\_page!='':

        #如果没有下一页了，停止爬取

            yield scrapy.Request(*self*.xidian\_next\_page, *callback*=*self*.parse)

BUPTJOB \_spider.py

*class* BUPTJOBSpider(*scrapy*.*spiders*.*Spider*):

    name="BUPTJOB"

    allowed\_domains = ["job.bupt.edu.cn"]

    # start\_urls = ['https://job.bupt.edu.cn/f/recruitmentinfo/ajax\_frontRecruitinfo?pageNo=']

    pagenum=1

*def* start\_requests(*self*):

        #由于start\_urls不能传静态爬取参数，用start\_requests

        yield scrapy.Request('https://job.bupt.edu.cn/f/recruitmentinfo/ajax\_frontRecruitinfo?pageNo=',

*callback*=*self*.parse, *meta*={'method': "stastic"})

*def* parse(*self*, *response*):

        data = json.loads(response.body)

        # 获取网页的json数据

        jobpost\_list = data['object']['list']

        print("[BUPT]now pagenum:"+*str*(data['object']['pageNo']))

        print("[BUPT]job len:"+*str*(len(jobpost\_list)))

        for jobpost in jobpost\_list:

            if datetime.strptime(jobpost["startTime"], '%Y-%m-%d %H:%M:%S') < datetime.strptime('2021-09-01', '%Y-%m-%d'):

                break

            id = jobpost['id']

            #获取下一页url，到二级页面爬取

            yield scrapy.Request(*url*=*f*'https://job.bupt.edu.cn/f/recruitmentinfo/ajax\_show?recruitmentId={id}',

*callback*=*self*.parse\_details, *meta*={'method': "stastic"})

#回调函数改变

*self*.n=data['object']['pageNo']+1

        if not data['object']['lastPage']:

            #判断尾页

            yield scrapy.Request(*url*=*f*'https://job.bupt.edu.cn/f/recruitmentinfo/ajax\_frontRecruitinfo?pageNo={*self*.n}',

*callback*=*self*.parse, *meta*={'method': "stastic"})

*def* parse\_details(*self*, *response*):

        #爬详情页，代码非常简单

        item = Jobtop10Item()

        data = json.loads(response.body)

        job = data['object']['recruitmentinfo']

        item['job\_title'] = job['title']

        item['job\_date'] = job['startTime']

        item['job\_views'] = job['browseNumber']

        item['job\_nums'] = job['positionNum']

        yield item

UESTCJOB\_spider.py

*def* parse(*self*, *response*):  # 解析爬取的内容，连二级页面都不用爬了

        data = json.loads(response.body)   # 获取网页的json数据

        jobpost\_list = data['list']

        print("[UESTC]now pagenum:"+*str*(data['page']))

        print("[UESTC]job len:"+*str*(len(jobpost\_list)))

        for jobpost in jobpost\_list:

            jobtime=jobpost["date"]

            jobtime = jobtime.replace("年", '-').replace("月", '-').replace("日", '')

            # print(jobtime)

            if datetime.strptime(jobtime, '%Y-%m-%d') < datetime.strptime('2021-09-01', '%Y-%m-%d'):

                return None

            item = Jobtop10Item()

            item['job\_title'] = jobpost['title']

            item['job\_date'] = jobtime

            item['job\_views'] = jobpost['viewcount']

            item['job\_nums'] = 1

            yield item

*self*.n = data['page']+1

        if *self*.n <= data['totalPage']:

            yield scrapy.Request(*url*='https://yjsjob.uestc.edu.cn/coread/listeminfo.action?page={}'.format(*self*.n),*callback*=*self*.parse, *meta*={'method': "stastic"})

最后运行出来的结果，北邮和成电在很短的时间内完成爬取（约五分钟），西电爬取了2个小时。如果动态页面能用json爬取，无疑是比较优秀的方案。

1. 对数据处理模块的设计
2. **数据分析**
3. **作业小结**