**懂球帝网站的足球新闻数据爬取问题记录**

1. 安装scrapy就遇到很多兼容关联包兼容方面的问题，现将目前安装的所有包以及版本做 一下记录：

PYTHON环境：3.8

执行：pip list 后的结果如下：

attrs==23.2.0

Automat==22.10.0

beautifulsoup4==4.7.1

certifi==2022.6.15

cffi==1.16.0

charset-normalizer==2.1.0

constantly==23.10.4

cryptography==36.0.2

cssselect==1.2.0

distlib==0.3.8

filelock==3.15.1

grpcio==1.54.0

grpcio-tools==1.54.0

h2==3.2.0

hpack==3.0.0

hyperframe==5.2.0

hyperlink==21.0.0

idna==3.3

incremental==22.10.0

itemadapter==0.9.0

itemloaders==1.3.1

jmespath==1.0.1

lxml==5.2.1

packaging==24.0

parsel==1.9.1

pillow==10.3.0

pipenv==2024.0.1

platformdirs==4.2.2

priority==1.3.0

Protego==0.3.1

protobuf==4.22.3

pyasn1==0.6.0

pyasn1\_modules==0.4.0

pycparser==2.22

PyDispatcher==2.0.7

PyExecJS==1.5.1

pymongo==3.7.2

pyOpenSSL==22.0.0

pypiwin32==223

pywin32==306

queuelib==1.7.0

requests==2.28.1

Scrapy==2.5.0

selenium==3.141.0

service-identity==24.1.0

six==1.16.0

soupsieve==2.5

Twisted==22.10.0

twisted-iocpsupport==1.0.4

typing\_extensions==4.12.1

urllib3==1.26.11

virtualenv==20.26.2

w3lib==2.2.0

zope.interface==6.4.post2

备注：

记得先更新一下pip的版本。

这里需要特别关注的包是：

Scrapy==2.5.0

Twisted==22.10.0

如果Scrapy与Twisted包的版本不匹配，会导致Scrapy项目，在创建与启动的过程中，遇到很多报错的问题现象。

比如：

ImportError: cannot import name ‘HTTPClientFactory‘ from ‘twisted.web.client‘ (unknown location)

pillow==10.3.0

这个包的作用会体现在通过Scrapy将图片资源下载到本地的功能点中

如果不进行正常的安装，则会得到如下的报错提示：

2024-06-14 18:37:32 [scrapy.middleware] WARNING: Disabled DongqiudiImagePipeline: ImagesPipeline requires installing Pillow 4.0.0 or later

1. 默认的user-agent: ‘dongqiudi (+http://www.yourdomain.com)’，需要转换为随机的user-agent，来对抗反爬机制。

此处使用中间件：

'dongqiudi.middlewares.MyRequest'来进行具体实现

1. 爬取新闻数据的过程中，需要处理数据的注意点：

3-1、需要获取当前时间作为crawl\_time，抓取时间：

代码片段：

import time;

crawl\_time = time.strftime(“%Y-%m-%d %H:%M:%S”)

3-2、获取新闻详情的正文内容时，需要去除内容中的HTML标签

代码片段：

from bs4 import BeautifulSoup

# 获取新闻资讯文本内容  
html\_content = response.xpath("//div[@class='con']").extract\_first()  
# 需要去除HTML标签  
soup = BeautifulSoup(html\_content, 'html.parser')  
content = soup.get\_text()

3-3、获取新闻图片时，需要注意XPATH表达式中元素节点的位置：

# 获取新闻图片URL  
image\_urls = response.xpath("//div[@class='con']//img/@data-src").extract()  
dongqiudi\_info['image\_urls'] = image\_urls

问题点思考：XPATH表达式中”/”与”//”的区别。

1. DongqiudiItem的导入问题，在IDE提示中，出现有报错信息，但是，在实际的运行过程中，却是正常的。

# from dongqiudi.dongqiudi.items import DongqiudiItem  
from dongqiudi.items import DongqiudiItem

可以参考这个问题页面链接：

<https://class.imooc.com/course/qadetail/222689>

补充：

from dongqiudi import db\_config

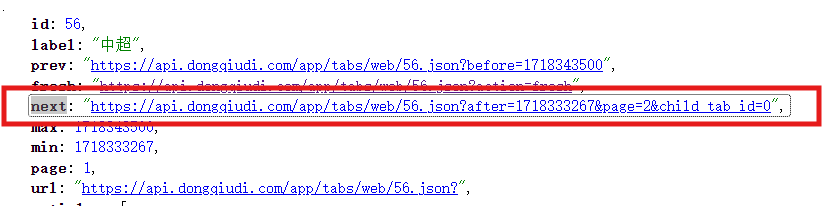
类似的问题还有这个地方。

1. yield语法的使用问题
2. 关于下一页的数据的获取，好像懂球帝站点在各个主题新闻列表页下的”下一页”新闻数据好像没有最后一页，就是没有底页的感觉。

例如：“中超”资讯列表数据页的接口：

https://api.dongqiudi.com/app/tabs/web/56.json

返回结果中的”next”字段好像一直都有值，这个就引发思考了，如果一直按照流程一页页的爬取足球新闻下去，是不是无底洞，然后会被发现么？



对应以下这段代码：

# 获取下一页的数据  
next\_page\_url = page\_response\_data['next']  
if next\_page\_url is not None and next\_page\_url.strip() != '':  
 # print('下一页的page\_url为：' + next\_page\_url)  
 yield scrapy.Request(url=next\_page\_url, callback=self.parse)

1. IP代理中间件，由于未购买任何供应商提供的IP代理服务，因此暂未真正实现。
2. crawl\_dongqiudi.py中的方法名称问题

def parse(self, response):

这个方法名是否是固定定的呢？如果改成其他方法名称的话，可能造成运行结果和实际需求有差异么？

1. 作业注意点与说明：

9-1、MONGODB数据源配置

请对db\_config.py中的下述配置信息做正确设置

# 数据源配置信息

MONGODB\_ADDR = '192.168.31.11'

MONGODB\_PORT = '27017'

9-2、存在问题下拉列表数据”下一页”字段是个无底洞的顾虑，在爬虫程序通过执行main.py，被成功启动后，可能需要做一下人为手工的关停操作来终止程序的不断执行。

9-3、由于网站更新迭代的原因，

参考了目前版本的懂球帝PC网站如下页面，进行数据的抓取。

日期时间：20240614

页面链接：<https://www.dongqiudi.com/articles>

主要新闻数据来源为：头条、中超、英超、意甲、西甲、德甲

以上这6个模块下的新闻数据。