杭州電子科技大学

《可视计算基础》

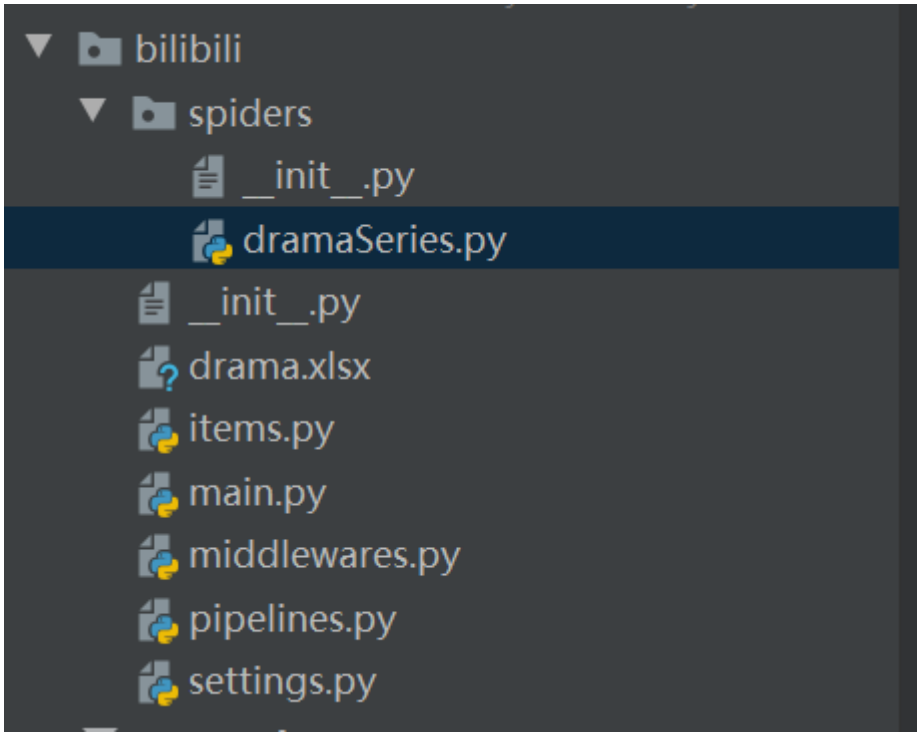
代码说明文档

Bilibili 番剧爬取

实验人员：

18051624 沈建鑫

首先用 scarpy projectstart name 创建一个 scarpy 工程，工程目录如下：



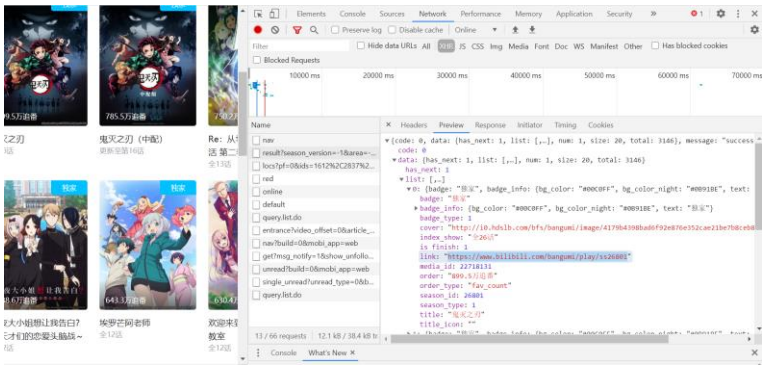
其中，spiders 里面放爬虫代码，items 中定义了爬取的字段，main 是用来代替命令行输入 crawl，middlewares 是中间层处理函数，pipelines 是数据处理相关文件，settings 存放基本配置。

第一步，我们确定要爬取的网址：

https://www.bilibili.com/anime/index/#season_version=-1&area=-1&is_finish=-1&copyright=-1&season_status=-1&season_month=-1&year=-1&style_id=-1&order=3&st=1&sort=0&page=1

剔除一些没必要的信息后：

https://api.bilibili.com/pgc/season/index//result?page=1&season_type=1&pagesize=20&type=1



通过观察可以得到，要想爬取多页的数据只要改变 page 的信息即可，因此可以确定我们的 url，这里我们爬取 100 页。

```
class DramaSeries(scrapy.Spider):
    name = 'drama'
    allowed_domains = ['https://api.bilibili.com/']
    i = 1
    start_urls = ['https://api.bilibili.com/pgc/season/index//result?page=%s&season_type=1&pagesize=20&type=1' % s for s
                  in range(1, 101)]
```

接下来在 items 中定义我们需要爬取的字段，通过观察该文档的第一幅图即可得到哪些字段是需要爬取的。

```
class BilibiliItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    number = scrapy.Field()
    badge = scrapy.Field()
    cover_img = scrapy.Field()
    index_show = scrapy.Field()
    link = scrapy.Field()
    media_id = scrapy.Field()
    order_type = scrapy.Field()
    season_id = scrapy.Field()
    title = scrapy.Field()
    pass
```

定义完后，回到爬虫代码，可以完成 parse 函数，实现爬虫解析相应的主要逻辑了。

```python
def parse(self, response):
    item = BilibiliItem()
    drama = json.loads(response.text)
    data = drama['data']
    data_list = data['list']
    #print(data_list)
    for filed in data_list:
        item['number'] = self.i
        item['badge'] = filed['badge']
        item['cover_img'] = filed['cover']
        item['index_show'] = filed['index_show']
        item['link'] = filed['link']
        item['media_id'] = filed['media_id']
        item['order_type'] = filed['order_type']
        item['season_id'] = filed['season_id']
        item['title'] = filed['title']
        #print(self.i, item)
        self.i += 1
        yield item
```

到这一步结束已经基本实现爬虫的功能了，接下去完善一下用 scrapy projectstart 创建出来的工程文件中的其他代码文件。

对于文件写入，根据刚刚解析出来的字段构建列表，然后用 openpyxl 的 Workbook 库来实现文件的直接写入。

```python
class BilibiliPipeline(object):
    excelBook = Workbook()
    activeSheet = excelBook.active
    file = ['number', 'title', 'link', 'media_id', 'season_id', 'index_show', 'cover_img', 'badge']
    activeSheet.append(file)
    def process_item(self, item, spider):
        files = [item['number'], item['title'], item['link'], item['media_id'], item['season_id'], item['index_show'],
                 item['cover_img'], item['badge']]
        self.activeSheet.append(files)
        self.excelBook.save('C:/Users/59723/PycharmProjects/bilibili/bilibili/drama.xlsx')
        return item
```

在 settings 中，我们添加 scrapy 发送 http 请求默认使用的请求头以及 pipelines 的设置。

```python
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 = {
   'pipelines.BilibiliPipeline': 100,
}
```

最后写一个简单的 main 函数来代替每次命令行运行输入 scrapy crawl name。

```
from scrapy import cmdline
import os
import sys

sys.path.append(os.path.dirname(os.path.abspath(__file__)))
cmdline.execute(['scrapy', 'crawl', 'drama'])
```

运行 main 就可以看到爬取出的结果了。

```
2020-11-16 14:45:11 [scrapy.core.scraper] DEBUG: Scraped from <200 https://api.bilibili.com/pgc/season/index//result?page=98&season_type=1&pagesize=20&type=1>
{'badge': '',
 'cover_img': 'http://i0.hdslb.com/bfs/bangumi/170a15388a44a826c0e0afb4644b81fba981590a.jpg',
 'index_show': '全1话',
 'link': 'https://www.bilibili.com/bangumi/play/ss3787',
 'media_id': 3787,
 'number': 1996,
 'order_type': '',
 'season_id': 3787,
 'title': '偶像大师 灰姑娘女孩 第一季 OVA'}
2020-11-16 14:45:12 [scrapy.core.scraper] DEBUG: Scraped from <200 https://api.bilibili.com/pgc/season/index//result?page=99&season_type=1&pagesize=20&type=1>
{'badge': '',
 'cover_img': 'http://i0.hdslb.com/bfs/bangumi/386bf66370e5ba4a56896cc44b6f8d017920ba97.jpg',
 'index_show': '全1话',
 'link': 'https://www.bilibili.com/bangumi/play/ss4723?theme=movie',
 'media_id': 4723,
 'number': 1997,
 'order_type': '',
 'season_id': 4723,
 'title': '乡村医生'}
2020-11-16 14:45:12 [scrapy.core.scraper] DEBUG: Scraped from <200 https://api.bilibili.com/pgc/season/index//result?page=100&season_type=1&pagesize=20&type=1>
{'badge': '',
 'cover_img': 'http://i0.hdslb.com/bfs/bangumi/39245a2e0bd128880b39bf9d31177f834bb53ce8.jpg',
 'index_show': '全13话',
```

Xlsx：

	A	B	C	D	E	F	G	H
1	number	title	link	media_id	season_i	index_sh	cover_im	badge
2	1	鬼灭之刃	https://v	22718131	26801	全26话	http://i	独家
3	2	鬼灭之刃	https://v	28229443	34004	更新至第	http://i	独家
4	3	Re：从零	https://v	28229233	33802	全13话	http://i	会员专享
5	4	工作细胞	https://v	102392	24588	全14话	http://i	独家
6	5	辉夜大小	https://v	28228367	32982	全12话	http://i	独家
7	6	在下坂本	https://v	3450	3450	全13话	http://i	独家
8	7	路人女主	https://v	28228738	5971	全12话	http://i	独家
9	8	命运之夜	https://v	28222114	28332	2019-07-	http://i	独家
10	9	炎炎消防	https://v	28221335	27959	全24话	http://i	出品
11	10	埃罗芒阿	https://v	5997	5997	全12话	http://i	独家
12	11	Re：从零	https://v	28226009	31151	全1话	http://i	会员专享
13	12	Fate/Zer	https://v	1650	1650	全13话	http://i	独家
14	13	动物狂想	https://v	28222618	28542	全12话	http://i	独家
15	14	JOJO的奇	https://v	135652	25681	全39话	http://i	独家
16	15	超能力女	https://v	78352	23850	全12话	http://i	独家
17	16	擅长捉弄	https://v	28221403	28006	全12话	http://i	独家
18	17	欢迎来到	https://v	6339	6339	全12话	http://i	独家
19	18	Re：从零	https://v	28226008	31150	全1话	http://i	会员专享
20	19	OVERLORD	https://v	102252	24596	全13话	http://i	独家
21	20	齐木楠雄	https://v	8812	21469	全24话	http://i	会员专享
22	21	笨女孩	https://v	6311	6311	全12话	http://i	独家