第2天两行代码下载网站数据

一、 需求

现在我们的需求是把下面网站上的小说下载下来，并把处理后的内容，写的本地的一个文件中

网址是: http://www.pingfandeshijie.net/di-yi-bu-01.html

二、分析技术点

1. 模拟浏览器的行为

我们可以使用第三方模块来完成这件事

其实这几就是最底层的爬虫原理

1.1 requests 获取数据

安装模块

```
[root@qfedu.com ~]# pip3 install requests
```

使用模块

```
import requests
url = 'http://www.pingfandeshijie.net/di-yi-bu-01.html'

# 模拟浏览器发送请求并得到回应，返回的是一个 response 对象
r = requests.get(url)

# 从返回的对象中得到字符串类型的数据 r.content 是二进制的数据
html = str(r.content, encoding='utf-8')

# 这里为了测试的目的，可以把爬取的网页内容写的一个文件中
# 因为频繁的测试，会导频繁的请求，有可能被网站禁掉我们的 IP
# 写好后就可以把这段代码先注释掉了
with open('平凡的世界.txt', 'w', encoding='utf-8') as f:
    f.write(html)
```

2. python 操作文件

2.1 语法

```
with open("文件路径", "文件的打开模式",  encoding="字符编码")  as 变量(文件对象):
    执行文件对象对应的方法，这里的代码必须缩进 4 个空格
    多行缩进要一致
```

- 文件路径支持绝对路径和相对路径
- 文件的打开模式有：
    - r 只读，文件对象可以被 `for` 循环，每次循环一行
    - w 只写，`f.write(接收的是字符串)` `f.writeline(接收一个列表)` 每次都会把之前的内容先清空，之后才写入新的内容

2.2 处理数据

我们处理数据可以把刚才写入的文件再读到内存中，进行进一步的处理

```
with open('平凡的世界.html', 'r', encoding='utf-8') as f:
    for line in f:
        if line.startswith('<p>') and '=' not in line:
            print(line)
```

输出：

```
<p>1 9 7 5年二、三月间，一个平平常常的日子，细蒙蒙的雨丝夹着一星半点的雪花，正纷纷淋淋地向大地飘洒着。时令已快到惊蛰，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。黄土高原严寒而漫长的冬天看来就要过去，但那真正温暖的春天还远远地没有到来。</p>
.....篇幅原因，后面略了.....
```

三、作业

1. 说明

> 继续爬取 到 第五章，并把数据处理后,写如到一个文件。 处理过的数据，不能有 html 的标签。

示例：

１９７５年二、三月间，一个平平常常的日子，细蒙蒙的雨丝夹着一星半点的雪花，正纷纷淋淋地向大地飘洒着。时令已快到惊蛰，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。黄土高原严寒而漫长的冬天看来就要过去，但那真正温暖的春天还远远地没有到来

2. 解题代码


```python
# 第三方模块，需要自己安装
# 如何安装?
# 在 shell 中，执行如下命令
# pip3 install  requests
import requests

def query_html(url):
    """
    请求url，并返回 html 页面内容
    :return: str html 页面内容
    """
    r = requests.get(url)

    # 把 二进制的内容转换成字符串，  就是页面的内容
    html = str(r.content, encoding='utf-8')
    return html


def parse_data(data):
```

```
    """
    清洗数据
    :param data:  str html page
    :return: list  处理之后数据列表
    """
    content = []
    url = ''

    for line in data.splitlines():
        if '<p>下一章: ' in line:
            url = line.split()[1]
            url = url.split('"')[1]
            break
        elif '<p>' in line:
            # 去除 每行两端的 p 标签
            line = line[3:-4]
            content.append(line + '\n')
    return content, url


def write_file(conent='', file_name=''):
    """写文件"""
    if not file_name:
        file_name = '平方的世界-路遥.txt'

    with open(file_name, 'a', encoding='utf-8') as
f:
        f.writelines(conent)



def main(url_path):

    while url_path:
        url_path = main(url_path)
```

```
        html = query_html(url_path)
        content_list, url_path = parse_data(html)
        write_file(content_list)

        if '06.html' in url_path:
            break


if __name__ == '__main__':
    url_path = 'http://www.pingfandeshijie.net/di-yi-bu-01.html'
    main(url_path)

```