实验三 下载网页并用正则表达式获取数据

3.1 实验介绍

此实验完成单网页数据爬取，并用正则表达式分解页面元素。

3. 2 实验目标

1. 掌握单页面数据采集方法；
2. 掌握 requests 库用法
3. 掌握正则表达式用法

3. 3 实验原理与方法

requests 库介绍

requests 库是 Python 中最简单易用的 HTTP 库。

HTTP 协议的主要请求

GET	获取 URL 位置的资源
HEAD	获取响应报告中的头部信息
POST	向 URL 位置的资源后附加数据
PUT	向 URL 位置存储资源
PATCH	局部更新 URL 位置的资源
DELETE	删除 URL 位置的资源

HTTP 协议的这些请求，在 requests 库中都有对应的函数实现。

requests 中主要函数说明

requests.request()	构造请求
requests.get()	获取网页（HTTP 的 GET）
requests.head()	获取网页的头部信息（HTTP 的 HEAD）
requests.post()	提交 post 请求（HTTP 的 POST）
requests.put()	提交 put 请求（HTTP 的 PUT）
requests.patch()	提交局部修改请求（HTTP 的 PATCH）
requests.delete()	提交删除请求（HTTP 的 DELETE）

requests 库中包含了两个主要的对象——Request 对象和 Response 对象，这两个对象贯穿的了 requests 主要函数的使用，其中常使用 Response 对象的属性获取网页的重要信息。

requests 库中的异常及相关函数

request.ConnectionError	网络连接错误异常
request.HTTPError	HTTP 错误异常
request.URLRequired	URL 缺失异常
requestTooManyRedirects	重定向异常（超过最大重定向次数）
request.ConnectTimeout	连接超时异常
request.Timeout	请求 URL 超时异常
request.Response.raise_for_status()	Response 对象判断状态码方法，非 200 则产生异常

更多 requests 库相关内容请参考官方文档。

正则表达式介绍

正则表达式用于可以用简洁的形式表达一组有规律的字符串。

操作符	说明
.	表示任意字符
[ ]	字符集，表示某一位置上的字符的取值集合。 [ab]表式某位置可以取 a、b，[a-z]表示某位置可以取 a 至中的字符
[^]	非字符集，表示某一位置上不能取的字符集合
*	前一字符可以扩展任意次数，包括 0 次
+	前一字符可以扩展任意次数，至少 1 次
?	前一字符取或不取，即前一字符扩展 0 次或 1 次
|	或操作，左右两边的字符串取其中一个
{n}	对前一字符扩展 n 次
{m, n}	对前一字符扩展 m 至 n 次，包括 n 次
^	匹配字符串起始位置
$	匹配字符串结束位置

( )	分组标记
\d	表示 0-9 的数字字符，等价[0-9]
\w	表示数字、大小写字母和下划线的字符，等价[A-Za-z_]

正则表达式常用操作符如下。

re 库介绍

命令行中使用如下命令安装 re 库。

pip install re

代码中使用如下语句导入 re 库

import re

Python 中使用在字符串前加 r 来表达正则表达式，例如 r'[0-9]'、r'[A-Za-z]'，事实上，字符串前加 r 表示字符串不解释转义字符。

re 库中常用函数的说明

函数	说明
re.search()	在字符串中找到和正则表达式匹配的第一个位置
re.match()	在字符串起始位置开始匹配正则表达式
re.split()	将字符串按照正则表达式分割为一组子串
re.sub()	替换所有匹配正则表达式的子串
re.findall()	找到字符串中所有匹配正则表达式的子串
re.finditer()	返回所有匹配正则表达式的子串的迭代器

更多函数的使用参考官方文档。

3.4 实验步骤

1. 简单学习 requests 库的相关函数及其参数意义、两个主要对象及其属性、异常信息。
2. 学习正则表达式的使用方法和正则表达式库的使用方法。
3. 在本地启动 flask 网站，使用 requests 库爬取网页，使用 re 库提取关键信息。

3.5 实验要求

1. 按照实验步骤编写 Python 脚本爬取网页，代码尽可能完备（考虑异常情况）、尽可能具有灵活性。

2. 将提取的关键信息保存到文件.
3. 撰写实验报告。