

本章总结

本章介绍了爬取静态网站常用的两个库：requests 库和BeautifulSoup4 库，以及介绍了HTML 基础知识。在爬虫中，这一些都是基础且必备的技能，学习过程中需要勤加练习，熟练使用。

下面总结一下：

- requests库中主要学习get()方法，通过它向网站发起请求，获取网址数据；get()方法返回的Response对象拥有4个属性，可用于查看网页请求的状态、对返回的数据进行编码和将返回的数据转化为二进制或字符串形式。
- BeautifulSoup4库可用来解析网页数据和提取网页数据。
- BeautifulSoup对象和Tag对象都有两个重要的方法find()和find_all()可以用于提取数据；Tag对象还有两个很重要的属性text和['属性名']，分别用于提取标签中的字符串和属性对应的属性值。
- HTML基础知识包括HTML层级结构、常见的标签、属性以及标签和属性二者的呈现形式；在爬虫实战中，能够看懂HTML 的内容，能够定位到需要爬取的数据是在哪个标签和属性中即可。