

第三周

学习 `urllib`、`urllib3` 的使用方法。

`urllib` 是 Python 标准库中用于网络请求的库。在学习过程中，我使用了 `urllib.request` 模块中的 `urlopen` 方法请求百度首页，获得了它页面的源代码。同时该函数还可以设置请求超时，以及使用 `data` 参数提交数据。不同于 `urlopen` 方法，`Request` 方法能够加入请求头、代理、Cookie 等信息获取指定 URL 页面源代码。

在练习中容易感觉到 `urllib` 是最常用、也是最容易用的 Python 网络请求库。从理论上讲，`urllib` 能做到的，其他网络库通常也能做到，只是使用方法和性能不同而已。

`urllib` 中的 API 大多与 URL 有关，所以可以看出 `urllib` 是侧重于 URL 的请求构造，而 `urllib3` 则是服务与升级的 HTTP1.1 标准，且拥有高校 HTTP 连接池管理及 HTTP 代理服务的功能库。它和 `urllib` 的功能类似，只是使用方法和原理不同，包括发送 GET 请求、发送 POST 请求、设置请求头、超时设置等，初次之外它对网站中的文件上传提供了很大的支持。

第一周

了解网络爬虫，熟悉 Python 语言。

网络爬虫是一种自动化数据采集程序，其中，搜索引擎就是个很好的例子，搜索引擎技术中大量使用爬虫，它爬取整个互联网的内容，存储在数据库中做索引。例如，百度搜索、谷歌搜索就是一只大爬虫。

第二周

学习有关爬虫和网页的基本知识。

了解爬虫的常见类型、基本结构和工作流程。包括通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深度网络爬虫。爬虫的工作流程可概括为：解析 URL 得到网页源码，按照一定需求和规则抓取信息。

选择一个常用的网站 <https://www.csdn.net>。通过一系列操作学习有关网页的知识，包括 HTTP 的基本原理、网页的组成、网页的结构，以及 Session 和 Cookie 的相关内容。