传智播客

《解析 Python 网络爬虫：核心技术、Scrapy 框架、分布式爬虫》教学设计

课程名称：解析 Python 网络爬虫

授课年级：2018 年级

授课学期：2018 学年第一学期

教师姓名：某某老师

2018 年 09 月 09 日

课题名称	第 4 章 抓取网页数据	计划学时	6 学时
内容分析	基于爬虫的实现原理，我们进入爬虫的第一个阶段：抓取网页数据，即下载包含目标数据的网页。抓取网页需要通过爬虫向服务器发送一个 HTTP 请求，然后接收服务器返回的响应内容中的整个网页源代码。 Python 要想完成这个过程，既可以使用内置的 urllib 库，也可以使用第三方库 requests。使用这两个库，我们在抓取网页数据的时候，就只需要关心请求的 URL 格式，以及要传递什么参数，要设置什么样的请求头，而不需要关心它们的底层是怎样实现的。		
教学目标及基本要求	1、 了解什么是 urllib 库 2、 会使用 urllib 库爬取网页 3、 掌握如何转换 URL 编码，可以使用 GET 和 POST 两种方式实现数据传输 4、 知道伪装浏览器的用途 5、 掌握如何自定义 opener 6、 了解服务器的超时 7、 熟悉一些常见的网络异常 8、 掌握 requests 库的使用		
教学重点	1、 会使用 urllib 库爬取网页 2、 掌握如何转换 URL 编码，可以使用 GET 和 POST 两种方式实现数据传输 3、 知道伪装浏览器的用途 4、 掌握 requests 库的使用		
教学难点	1、 掌握如何自定义 opener		
教学方式	教学采用教师课堂讲授为主，使用教学 PPT 讲解		
教学过程	**第一课时** **（什么是 urllib 库，快速使用 urllib 爬取网页）** **一、创设情境，导入 urllib 库** （1） 教师通过爬虫的原理，引出抓取网页数据。 基于前面介绍的爬虫的原理，我们需要根据一个初始的 URL，发送网络请求到指定的服务器，去获取其对应的网络数据，这就是网络爬虫的第一个阶段：抓取网页数据。 （2） 什么是抓取网页数据？ 抓取网页需要通过爬虫向服务器发送一个 HTTP 请求，然后接收服务器返回		

的响应内容中的整个网页源代码。

（3） 教师通过课件，讲述 Python 提供的用于抓取网页数据的库。

Python 要想完成抓取网页数据的过程，既可以使用内置的 urllib 库，也可以使用第三方库 requests。

（4） 明确学习目标

- 要求学生掌握 urllib 库的基本使用

二、进行重点知识的讲解

（1） 什么是 urllib 库？

urllib 库是 Python 内置的 HTTP 请求库，它包含了 4 个模块：urllib.request、urllib.error、urllib.parse、urllib.robotparser。

（2） 教师根据课件，讲述使用 urllib 库快速爬取一个网页，并通过代码进行演示。

（3） 教师根据课件的多学一招，讲述 Python2 与 Python3 中 urllib 的不同之处。

Python2 中使用的是 urllib2 库来下载网页，Python3 出现后，之前 Python2 中的 urllib2 库被移到了 urllib.request 模块中，之前 urllib2 中很多函数的路径也发生了变化，希望大家在使用的时候多加注意。

（4） 教师根据课件，讲述 urlopen 方法的使用，并通过代码进行演示。

（5） 教师根据课件，讲述 HTTPResponse 对象的使用，并通过代码进行演示。

使用 urllib.request 模块中的 urlopen 方法发送 HTTP 请求后，服务器返回的响应内容封装在一个 HTTPResponse 类型的对象中。

（6） 教师根据课件，讲述构建 Request 对象，并通过代码进行演示。

当我们使用 urlopen 方法发送一个请求时，如果希望执行更为复杂的操作，比如增加 HTTP 报头，则必须创建一个 Request 对象来作为 urlopen 方法的参数。

三、归纳总结，布置作业/随堂练习

（1） 回顾上课前的学习目标，并对本节课要掌握的内容进行总结。

教师总结本节课需要掌握的知识点，包括 urlopen 方法的使用、使用 HTTPResponse 对象和构建 Request 对象。

（2） 布置随堂练习，检查学生的掌握情况。

根据博学谷和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。

（3） 使用博学谷系统下发课后作业。

第二课时

（URL 编码转换，处理 GET 请求，处理 POST 请求）

一、回顾上节课的内容，继续讲解本课时的知识

（1） 教师对学生们的疑问进行统一答疑。

（2） 回顾总结上节课内容，继续介绍本课时的内容。

上节课我们介绍了 urllib 库的基本使用，接下来，本节课将继续深入地探讨 urllib 库，学习如何使用 urllib 库发送 GET 或 POST 请求来传输数据。

（3） 明确学习目标

- 要求学生会转换 URL 编码
- 要求学生会处理 GET 请求和 POST 请求

二、进行重点知识的讲解

（1） 教师通过举例，讲述浏览器如何通过 URL 传递数据。

例如，教师打开百度首页，在搜索框中输入新闻两个字，可以看到浏览器显示了带有搜索结果的网页。此时，再次查看这个网页的 URL，发现这个网页的 URL 发生了变化，多了一些查询参数，这些就是要传递的数据。

（2） 为什么要转换 URL 编码？

当我们传递的 URL 包含中文或者其它特殊字符（例如，空格或/等）时，需要使用 urllib.parse 库中的 urlencode 方法将 URL 进行编码。

（3） 教师通过 4.3.1 的示例，讲述如何将 URL 进行编码转换，并通过示例代码进行演示。

（4） 教师通过举例，讲述 GET 请求的特点。

例如，使用 Fiddler 抓包工具捕获搜索新闻的网页后进行查看，发现这个请求是 GET 请求。GET 请求可以直接使用 URL 访问，在 URL 中已经包含了所有的参数。

（5） 教师通过 4.3.2 的示例，讲述如何使用 urllib 库处理 GET 请求，并通过实践进行演示。

（6） 教师通过举例，讲述 POST 请求的特点。

例如，使用 Fiddler 抓包工具捕获有道翻译的网页后进行查看，发现这个请求是 POST 请求。POST 请求的参数都放到数据体中，只能通过抓包工具进行查

看。

（7） 教师通过 4.3.3 的示例，讲述如何使用 urllib 库处理 POST 请求，并通过实践进行演示。

三、归纳总结，布置作业

（1） 回顾学习目标，对本节课的内容进行总结。

教师带领学生总结本节课需要掌握的内容，包括 URL 编码转换、处理 GET 请求和 POST 请求。

（2） 布置随堂练习，检查学生掌握情况。

根据博学谷和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。

（3） 使用博学谷系统下发课后作业。

第三课时

（添加特定 Headers—请求伪装，代理服务器）

一、回顾上节课内容，继续讲解本节课的内容

（1） 教师对学生们的疑问进行统一答疑。

（2） 回顾总结上节课内容，继续介绍请求伪装的知识。

在上一节课中，我们使用 urllib 库实现了数据传递，接下来继续介绍一些该库的其它使用技巧，包括添加请求头和设置代理服务器。

（3） 明确学习目标

- 要求学生会添加特定的请求头
- 要求学生会设置代理服务器

二、进行重点知识的讲解

（1） 什么是请求伪装？

对于一些需要登录的网站，如果不是从浏览器发出的请求，则得不到任何响应。因此，我们需要将爬虫程序发出的请求伪装成浏览器发出的请求。

（2） 教师根据 4.4 的示例，讲述如何自定义请求报头，并通过代码进行演示。

添加特定 Headers 的方式很简单，只需要调用 Request.add_header()即可。如果想查看已有的 Headers，可以通过调用 Request.get_header()查看。

（3） 什么是代理服务器？

提供代理服务的电脑系统或其它类型的网络终端称为代理服务器，大多被用

来连接互联网和局域网。

（4） 教师根据课件，讲述如何自定义 opener，并通过代码进行演示。

opener 是 urllib.request.OpenerDirector 类的对象，我们之前一直都在使用的 urlopen，就是模块帮我们构建好的一个 opener，但是它不支持代理、Cookie 等其他的 HTTP/HTTPS 高级功能。所以如果要想设置代理，不能使用自带的 urlopen，而是要自定义 opener。

（5） 教师根据课件，讲述怎样设置代理服务器，并通过代码进行演示。

免费开放代理的获取基本没有成本，我们可以在一些代理网站上收集这些免费代理，测试后如果可以用，就把它收集起来用在爬虫上面。

三、归纳总结，布置作业

（1） 回顾学习目标，对本节课需要掌握的内容进行总结

教师带领学生总结本节课需要掌握的知识点，包括请求伪装和代理服务器。

（2） 布置随堂练习，检查学生掌握情况。

根据博学谷和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。

（3） 使用博学谷系统下发课后作业。

第四课时

（超时设置，常见的网络异常，什么是 requests 库，requests 库初体验）

一、回顾上节课内容，继续讲解本节课的内容

（1） 教师讲解上节课有难度的作业，并对学生的疑问进行统一答疑。

（2） 回顾上节课的内容，继续介绍本节课的知识。

在上节课中，我们知道了伪装请求和代理服务器，其中伪装请求是应对反爬虫最常用的一种策略。接下来，本节课将为大家介绍一些 urllib 库的其它知识，包括超时设置、网络异常，并带领大家简单地认识一下 requests 库。

（3） 明确学习目标

- 要求学生会设置超时异常
- 要求学生了解一些常见的网络异常
- 要求学生能体会 requests 库的便捷之处

二、进行重点知识的讲解

(1) 教师通过举例，讲述为什么要设置超时。

例如，我们要爬取 1000 个网站，如果其中有 100 个网站需要等待 30s 才能返回数据，那如果要返回所有的数据，至少需要等待 3000 秒，如此长时间的等待显然是不可行的。

(2) 如何设置超时时间？

我们可以为 HTTP 请求设置超时时间，一旦超过这个时间，服务器还没有返回响应内容，那么就会抛出一个超时异常，这个异常需要使用 try 语句来捕获。

(3) 教师通过 4.6 节的示例，讲述如何给请求设置超时时间。

(4) 常见的网络异常有哪些？

URLError 和 HTTPError。

(5) 教师根据课件，讲述 URLError 异常的产生和捕获。

(6) 教师根据课件，讲述 HttpError 异常的产生和捕获。

(7) 什么是 requests？

requests 是基于 Python 开发的 HTTP 库，与 urllib 标准库相比，它不仅使用方便，而且能节约大量的工作。

(8) 教师根据课件，讲述 requests 库中常用的几个类。

requests.Request 表示请求对象，用于准备一个请求发送到服务器；requests.Response 表示响应对象，其中包含服务器对 HTTP 请求的响应；requests.Session 表示请求会话，提供 Cookie 持久性、连接池和配置。

(9) 教师根据课件，讲述分别使用 urllib 和 requests 抓取百度网站中“传智播客”关键字的搜索结果网页，并通过示例代码进行演示。

(10) 教师根据课件，总结与 urllib 相比 requests 的便捷之处。

三、归纳总结，布置作业

(1) 回顾学习目标，并对本节课需要掌握的内容进行总结。

教师带领学生总结本节课需要掌握的内容，包括超时设置、常见网络异常，以及什么是 requests 库。

(2) 布置随堂练习，检查学生的掌握情况。

根据博学谷和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。

(3) 使用博学谷系统下发随堂练习，检测学生对课堂内容的掌握情况。

第五课时

（发送请求，返回响应，案例—使用 urllib 库爬取百度贴吧）

一、回顾上节课内容，继续讲解本节课的内容

（1） 教师讲解上节课有难度的作业，并对学生的疑问进行统一答疑。

（2） 回顾上节课的内容，继续介绍本节课的知识。

上节课中，我们介绍了设置超时异常和常见的网络异常，并对这些网络异常产生的原因及处理进行了讲解。此外，我们初步体验了使用 requests 库抓取网页，更加简单。接下来，本节课将深入地介绍一些 requests 库的知识，使用 urllib 库爬取百度贴吧。

（3） 明确学习目标

- 要求学生掌握 requests 的基本使用
- 要求学生会使用 urllib 和 requests 抓取网页

二、进行重点知识的讲解

(1) 教师根据课件，讲述 requests 库提供的发送请求的函数。

教师可以从这些函数中随机选择两个，通过示例代码进行演示。

(2) 教师根据课件，讲述 requests 库提供的响应请求的 Response 类。

教师可以从这些函数中随机选择两个，通过示例代码进行演示。

(3) 教师通过 4.9 的案例，讲述如何使用 urllib 库爬取百度贴吧。

(4) 教师根据博学谷和随堂练习资源，给学生布置随堂练习，检测学生的掌握程度，并对学生出现的问题进行解决。

三、归纳总结，布置作业

（1）回顾本节课的知识点，要求学生掌握 requests 库的基本使用。

（2）使用博学谷系统下发随堂练习，检测学生对课堂内容的掌握情况。

第六课时

（上机练习）

上机练习主要针对本章中需要重点掌握的知识点，以及在程序中容易出错的内容进行练习，通过上机练习可以考察同学对知识点的掌握情况，对代码的熟练程度。

上机一：（考察知识点为使用 urllib 库发送 GET 请求）

形式：单独完成

题目：

	请按照以下要求完成。 要求如下： (1) 使用 urllib 库发送 GET 请求到百度搜索“有道翻译”结果的网页（https://www.baidu.com/s?wd=有道翻译）； (2) 把传递的数据进行 URL 编码。 (3) 使用 print 函数将最终爬到的数据打印出来。 **上机二：（考察知识点为使用 urllib 库发送 POST 请求）** **形式：单独完成** **题目：** **请按照以下要求操作。** **要求如下：** (1) 使用 Fiddler 工具抓取访问有道翻译网站的请求（http://fanyi.youdao.com）； (2) 记录访问网站时使用的 URL 地址和数据体； (3) 使用 print 函数将最终爬到的数据打印出来。
思考题和习题	见教材第 4 章配套的习题
教学后记	