天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

Python数据分析与挖掘实战

1、Urllib库实战

韦玮

自我介绍

- 天善商业智能和大数据社区 Python 讲师 – 韦玮
- 天善社区 ID - 韦玮
- https://www.hellobi.com – 学习过程中有任何相关的问题都可以提到技术社区 Python 版块。

本课概要

- 课前说明
- 作业讲解
- 关于学习方法
- urllib基础
- 超时设置
- 自动模拟HTTP请求

课前说明

1、作业提交方式与奖励：可以通过http://task.iqianyue.com提交，目前网站页面不太美观，但能实现相关功能，最终会求平均分，前十名的朋友至少可以获得老师亲笔签名的书籍一本（1月左右出版）。

2、课堂笔记可以咱们学员之间共享，**但不要发成博文或者公开**，如要公开，必须加上来源“韦玮老师课堂笔记”。因为：**目前国内很多网站版权意识太差**，咱们公布出去之后，很多网站直接采用爬虫自动采集，并且作者会成为别人，这些代码都是老师辛苦**原创**的，而书籍的出版时间会比课程延后，到时书籍中原创的代码却被当成盗版就不好了，并且到时实在要打官司，各位也比较麻烦，所以为了避免这种麻烦事，请各位配合。否则发现我课程就只能尽量少写代码了，那么这样课程质量必然下降，不希望这么做。

作业讲解

上节课的作业是如何爬取豆瓣出版社列表并写入文件中，目前（下午4：00）有110位同学提交了作业，辛苦。

接下来我们为大家讲解一下如何实现。

关于学习方法

各位同学的基础参差不齐，有些同学可能觉得跟不上，在此给出以下学习方法指导：

1、如果觉得跟不上，每次看完直播后，课后再看我们的录播，把咱们课程中讲到的知识点弄熟，因为各位听课的时候会发现能听懂的，但是可能课上不能完全消化，故而课后一定要花时间。

2、如果觉得很多基础知识不懂，就不要再去网上找各种资料了，因为你的精力有限。课程中涉及的内容一定是核心知识，所以，关键需要把课程弄熟悉，之后再去考虑扩展，否则，把课程弄熟即可。这几次课下来，一定会发现一个现象，就是上节课的难点知识，在下一节课中很多都会解决，因为我们课程大纲的承接性是非常好的，总之，做到“讲过必熟，未讲不急”就行。

3、一定要多敲代码、多敲代码、多敲代码。自己写的才是自己的。

4、遇到问题尝试独立解决，比如通过一些搜索引擎等工具，因为工作时更多需要有独立解决问题的能力，当然，难点问题希望群里踊跃讨论，在这个过程中，我们发现群里确实有高手。

urllib基础

要系统学习urllib模块，我们从urllib基础开始。这个知识点中，我们会为大家实战讲解urlretrieve()、urlcleanup()、info()、getcode()、geturl()等。

超时设置

由于网络速度或对方服务器的问题，我们爬取一个网页的时候，都需要时间。我们访问一个网页，如果该网页长时间未响应，那么我们的系统就会判断该网页超时了，即无法打开该网页。

有的时候，我们需要根据自己的需要，来设置超时的时间值，比如，有些网站反应快，我们希望2秒钟没有反应，则判断为超时，那么此时，timeout的值就是2，再比如，有些网站服务器反应慢，那么此时，我们希望100秒没有反应，才判断为超时，那么此时timeout的值就是100。接下来为大家实战讲解爬取时的超时设置。

自动模拟HTTP请求

客户端如果要与服务器端进行通信，需要通过http请求进行，http请求有很多种，我们在此会讲post与get两种请求方式。比如登陆、搜索某些信息的时候会用到。

接下来我们通过实战讲解。

天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

Python数据分析与挖掘实战

2、爬虫的异常处理

韦玮

自我介绍

- 天善商业智能和大数据社区 Python 讲师 – 韦玮
- 天善社区 ID - 韦玮
- https://www.hellobi.com – 学习过程中有任何相关的问题都可以提到技术社区 Python 版块。

本课概要

- 异常处理概述
- 常见状态码及含义
- URLError与HTTPError
- 异常处理实战

异常处理概述

爬虫在运行的过程中，很多时候都会遇到这样或那样的异常。如果没有异常处理，爬虫遇到异常时就会直接崩溃停止运行，下次再次运行时，又会重头开始，所以，要开发一个具有顽强生命力的爬虫，必须要进行异常处理。

常见状态码及含义

301 Moved Permanently：重定向到新的URL，永久性

302 Found：重定向到临时的URL，非永久性

304 Not Modified：请求的资源未更新

400 Bad Request：非法请求

401 Unauthorized：请求未经授权

403 Forbidden：禁止访问

404 Not Found：没有找到对应页面

500 Internal Server Error：服务器内部出现错误

501 Not Implemented：服务器不支持实现请求所需要的功能

URLError与HTTPError

两者都是异常处理的类，HTTPError是URLError的子类，HTTPError有异常状态码与异常原因，URLError没有异常状态码，所以，在处理的时候，不能使用URLError直接代替HTTPError。如果要代替，必须要判断是否有状态码属性。

接下来我们通过实战讲解。

异常处理实战

客户端如果要与服务器端进行通信，需要通过http请求进行，http请求有很多种，我们在此会讲post与get两种请求方式。比如登陆、搜索某些信息的时候会用到。

接下来我们通过实战讲解。

天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

Python数据分析与挖掘实战

3、爬虫的浏览器伪装技术

韦玮

自我介绍

- 天善商业智能和大数据社区 Python 讲师 – 韦玮
- 天善社区 ID - 韦玮
- https://www.hellobi.com – 学习过程中有任何相关的问题都可以提到技术社区 Python 版块。

本课概要

- 浏览器伪装技术原理
- 浏览器伪装技术实战

浏览器伪装技术原理

我们可以试试爬取csdn博客，我们发现会返回403，因为对方服务器会对爬虫进行屏蔽。此时，我们需要伪装成浏览器才能爬取。

浏览器伪装我们一般通过报头进行，接下来我们通过实战分析一下。

浏览器伪装技术实战

由于urlopen()对于一些HTTP的高级功能不支持，所以，我们如果要修改报头，可以使用urllib.request.build_opener()进行，当然，也可以使用urllib.request.Request()下的add_header()实现浏览器的模拟。

我们重点讲前者方法，后者方法是否掌握无所谓，有兴趣并有时间的同学可以自行研究第2种方法，接下来通过实战讲解。

天善智能
TIANSHAN SOFT

专注于商业智能BI和大数据的垂直社区平台

Python数据分析与挖掘实战

4、Python新闻爬虫实战

韦玮

自我介绍

- 天善商业智能和大数据社区 Python 讲师 – 韦玮
- 天善社区 ID - 韦玮
- https://www.hellobi.com – 学习过程中有任何相关的问题都可以提到技术社区 Python 版块。

本课概要

- 新闻爬虫需求及实现思路
- 新闻爬虫编写实战
- 作业

新闻爬虫需求及实现思路

需求：将新浪新闻首页（http://news.sina.com.cn/）所有新闻都爬到本地。

思路：先爬首页，通过正则获取所有新闻链接，然后依次爬各新闻，并存储到本地。

新闻爬虫编写实战

接下来为大家通过实战讲解如何编写新闻爬虫。

作业

爬取CSDN博客http://blog.csdn.net/首页显示的所有文章，每个文章内容单独生成一个本地网页存到本地中。

难点：浏览器伪装、循环爬各文章

思路：先爬首页，然后通过正则筛选出所有文章url，然后通过循环分别爬取这些url到本地。