

Python数据分析与挖掘实战

1、抓包分析实战

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

- 作业讲解
- 抓包分析概述
- 使用Fiddler进行抓包分析
- 抓取HTTPS数据包
- 爬取腾讯视频的评论

作业讲解

上节课的作业是如何爬取千图网的高清图片。

接下来我们为大家讲解一下如何实现。

抓包分析概述

所谓抓包分析，即将网络传输发送与接收的数据包进行抓取的操作，做爬虫时，数据并不一定就在HTML源码中，很可能隐藏在一些网址中，所以，我们要抓取某些数据，就需要进行抓包，分析出对应数据所隐藏的网址，然后分析规律并爬取。

使用Fiddler进行抓包分析

抓包工具有很多，Fiddler是其中一种，我们在爬虫项目中经常会使用Fiddler进行抓包。接下来通过实战为大家讲解如何使用Fiddler进行抓包分析。

抓取HTTPS数据包

Fiddler默认只能抓HTTP的数据，抓不到HTTPS的数据，如果想用Fiddler抓HTTPS的数据，需要进行相应设置。

可以参照韦玮老师的文章：

<https://ask.hellobi.com/blog/weiwei/5159>

爬取腾讯视频的评论

接下来我们为大家讲解如何抓取腾讯视频的评论数据，这些评论数据是需要抓包分析才能得到的，同时为大家讲解如何实现自动加载评论。

Python数据分析与挖掘实战

2、微信爬虫实战

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

- 什么是微信爬虫
- 微信爬虫编写实战

什么是微信爬虫

所谓微信爬虫，及自动获取微信的相关文章信息的一种爬虫。微信对我们的限制是很多的，所以，我们需要采取一些手段解决这些限制，主要包括伪装浏览器、使用代理IP等方式。

微信爬虫编写实战

接下来我们通过实战为大家讲解如何编写微信爬虫。

Python数据分析与挖掘实战

3、多线程爬虫实战

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

- 什么是多线程爬虫
- 糗事百科段子爬虫
- 将糗事百科段子爬虫改造为多线程爬虫

什么是多线程爬虫

所谓多线程，即程序中的某些程序段并行执行，合理地设置多线程，可以让爬虫效率更高。

糗事百科段子爬虫

接下来我们为大家先做一个普通的糗事百科段子爬虫。

将糗事百科段子爬虫改造为多线程爬虫

接下来我们为大家通过实战讲解如何将这个普通的糗事百科段子爬虫改变为多线程的爬虫。

Python数据分析与挖掘实战

4、Scrapy框架的安装

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

- 什么是Scrapy框架
- 安装Scrapy框架及各种常见错误解决技巧
- 少坑版安装方式
- 常见错误与解决
- 作业

什么是Scrapy框架

Scrapy是一个Python爬虫框架，非常适合做一些大型爬虫项目，并且开发者利用这个框架，可以不用过多关注细节，我们会重点说这个框架。

Scrapy的官网地址是：<http://scrapy.org/>

少坑版安装方式

由于Scrapy框架涉及太多依赖库，在此，如果想省事的朋友，可以按照这种方式去安装。如果想了解具体过程的朋友，可以先安装pip install scrapy，然后遇到错误参加后面的方式解决。

接下来讲解少坑安装方式。

0、开个VPN或者采用本地安装方式

1、首先，升级pip：

```
python -m pip install --upgrade pip
```

2、安装Visual Studio 2015专业版，去

http://download.microsoft.com/download/B/8/9/B898E46E-CBAE-4045-A8E2-2D33DD36F3C4/vs2015.pro_chs.iso下载

3、安装lxml---去<http://www.lfd.uci.edu/~gothlke/pythonlibs/>

找“lxml-”、pip install wheel、pip install lxml-3.6.4-cp35-cp35m-win_amd64.whl

4、pip install scrapy或pip install scrapy==1.1.0rc3

常见错误与解决1

```
You are using pip version 8.1.1, however version 8.1.2 is available.  
You should consider upgrading via the 'python -m pip install --upgrade pip' command.
```

```
running build_ext  
building 'lxml.etree' extension  
error: Unable to find vcvarsall.bat
```

```
-----  
Command "d:\python35\python.exe -u -c "import setuptools, tokenize;__file__='C:\Users\ADMINI~1\AppData\Local\Temp\pip-build-0fz6piyy\lxml\setup.py';exec(compile(getattr(tokenize, 'open', open)(__file__).read().replace('\r\n', '\n'), __file__, 'exec'))" install --record C:\Users\ADMINI~1\AppData\Local\Temp\pip-dxktqlun-record\install-record.txt --single-version-externally-managed --compile" failed with error code 1 in C:\Users\ADMINI~1\AppData\Local\Temp\pip-build-0fz6piyy\lxml\
```



常见错误与解决2

```
xmlXPathInitjlv76wvi.c
C:\Users\ADMINI~1\USE\AppData\Local\Temp\xmlXPathInitjlv76wvi.c(1): fatal er
ror C1083: 无法打开包括文件: "libxml\xpath.h": No such file or directory
*****
*****
Could not find function xmlCheckVersion in library libxml2. Is libxml2 insta
lled?
*****
*****
error: command 'E:\Program Files (x86)\Microsoft Visual Studio 14.0\VC\BIN\x86_amd64\cl.exe' failed with exit status 2

-----
Command "d:\python35\python.exe -u -c "import setuptools, tokenize;__file__='C:\
Users\ADMINI~1\USE\AppData\Local\Temp\pip-build-44llrit7\libxml\setup.py';
exec(compile(getattr(tokenize, 'open', open)(__file__).read().replace('\r\n', '\
n'), __file__, 'exec'))" install --record C:\Users\ADMINI~1\USE\AppData\Local\Te
mp\pip-jlrtdv6-record\install-record.txt --single-version-externally-managed --
compile" failed with error code 1 in C:\Users\ADMINI~1\USE\AppData\Local\Temp\pi
p-build-44llrit7\libxml\
```

```
C:\Users\Administrator.USER-20160828PN>pip install wheel
Requirement already satisfied (use --upgrade to upgrade): wheel in d:\python35\l
ib\site-packages

C:\Users\Administrator.USER-20160828PN>d:

D:\>cd vs2015pro

D:\vs2015pro>pip install lxml-3.6.4-cp35-cp35m-win_amd64.whl
Processing d:\vs2015pro\lxml-3.6.4-cp35-cp35m-win_amd64.whl
Installing collected packages: lxml
Successfully installed lxml-3.6.4
```

作业

- 本课作业就一个，把Scrapy框架安装好，然后将成功截图通过<http://task.iqianyue.com>提交给老师。