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、网络爬虫初识

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
- 如何查看模块功能以及如何安装模块
- 网络爬虫是什么？
- 网络爬虫能做什么事情？

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将多个表中的内容合并到一起。作业比较难，接下来我们来讲解一下上节课的作业。

参考答案：

http://mp.qq.com/material_show/show/3235303732333834393 7-1477477765-1477322044394549-0.html?_wv=2281701505&v=3&sig=ff9e1e4ae141514b06cf6e586ddc52f0&_bid=2321

如何查看模块功能以及如何安装模块

很多朋友问到，当接触到一个新模块的时候，如何了解这个模块的功能。主要方法有：

1、help()--输入对应模块名

1、阅读该模块的文档，一些大型的模块都有，比如scrapy等。

2、查看模块的源代码，分析各方法的作用，当然也可以从名字进行相应的分析。

有些朋友在安装模块的时候，经常会出现超时然后自动断掉的问题，解决这种问题，我们可以这样做：

1、使用VPN（推荐）

2、多试几次

3、使用本地whl文件来安装（推荐）实战讲解。http://www.lfd.uci.edu/~gohlke/pythonlibs/

网络爬虫是什么？

简单来说，网络爬虫就是自动从互联网中定向或不定向地采集信息的一种程序。

网络爬虫有很多种类型，常用的有通用网络爬虫、聚焦网络爬虫等。

网络爬虫能做什么事情？

网络爬虫可以做很多事情，比如通用网络爬虫可以应用在搜索引擎中，聚焦网络爬虫可以从互联网中自动采集信息并代替我们筛选出相关的数据出来。具体来说，网络爬虫经常可以应用在以下方面：

1、搜索引擎

2、采集金融数据

3、采集商品数据

4、自动过滤广告

5、采集竞争对手的客户数据

6、采集行业相关数据，进行数据分析

……

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、网络爬虫原理

韦玮

自我介绍

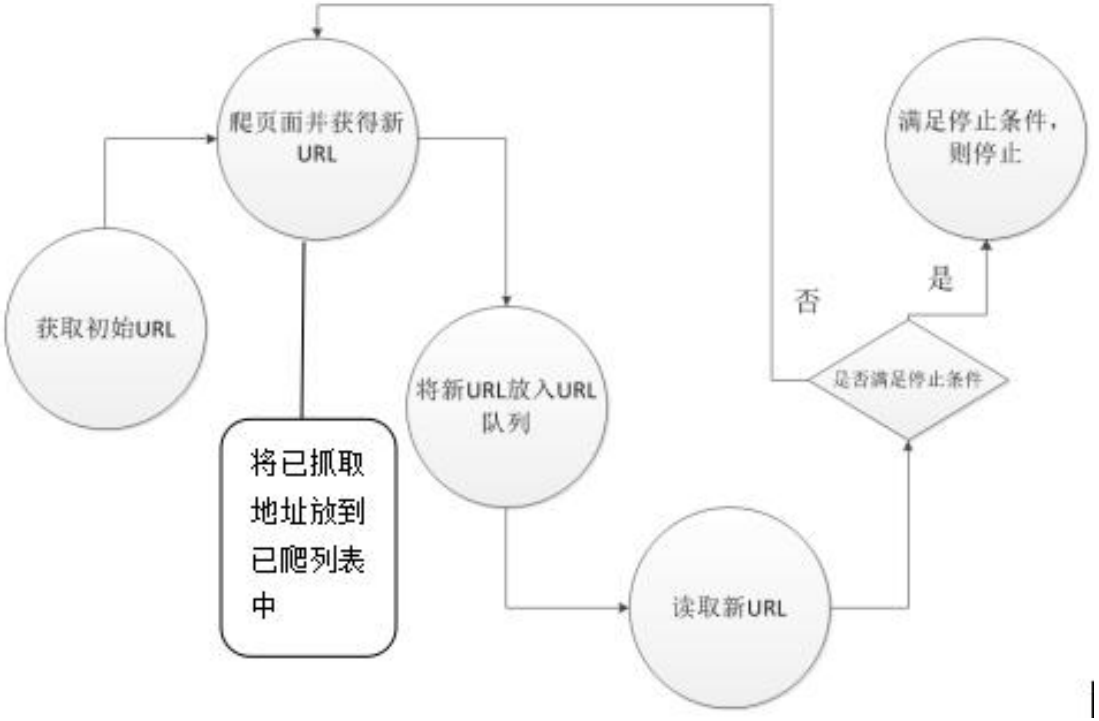
- 天善商业智能和大数据社区 Python 讲师 – 韦玮
- 天善社区 ID - 韦玮
- https://www.hellobi.com – 学习过程中有任何相关的问题都可以提到技术社区 Python 版块。

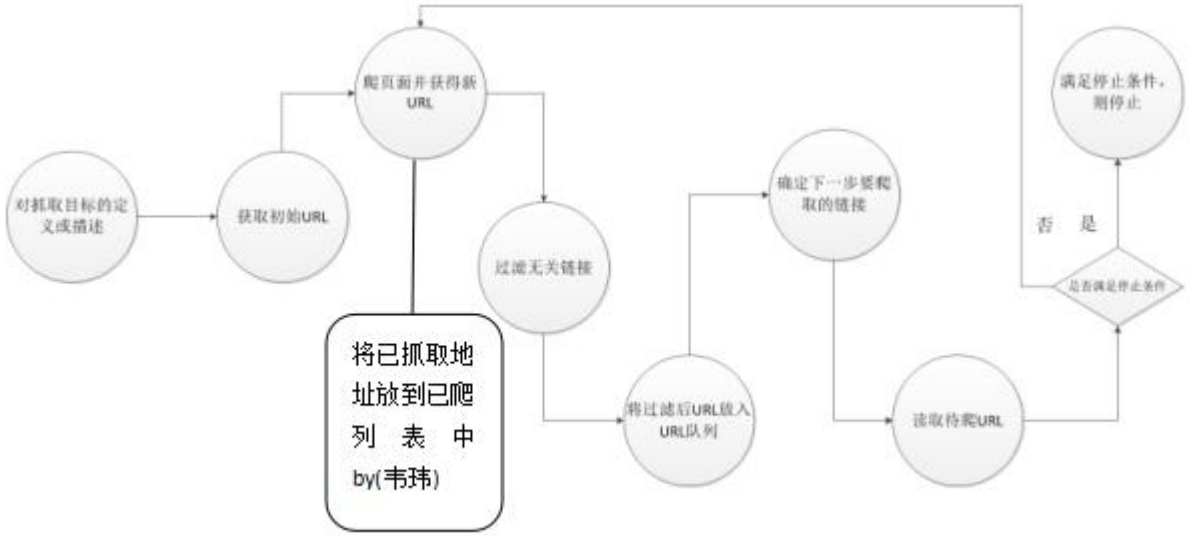
本课概要

- 网络爬虫的运行原理

网络爬虫的运行原理

通用与聚焦网络爬虫运行原理如下：





天善智能
TIANSHAN SOFT

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、正则表达式实战

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

- 什么是正则表达式
- 原子
- 元字符
- 模式修正符
- 贪婪模式与懒惰模式
- 正则表达式函数
- 常见正则实例
- 简单的爬虫
- 从网页中提取出qq群
- 作业：提取出版社信息并写入文件中

什么是正则表达式

世界上信息非常多，而我们关注的信息有限。假如我们希望只提取出关注的数据，此时可以通过一些表达式进行提取，正则表达式就是其中一种进行数据筛选的表达式。

原子

原子是正则表达式中最基本的组成单位，每个正则表达式中至少要包含一个原子。常见的原子类型有：

a普通字符作为原子

b非打印字符作为原子

c通用字符作为原子

d原子表

接下来进行实战讲解。

元字符

所谓的元字符，就是正则表达式中具有一些特殊含义的字符，比如重复N次前面的字符等。

接下来为大家通过实战讲解。

模式修正符

所谓的模式修正符，即可以在不改变正则表达式的情况下，通过模式修正符改变正则表达式的含义，从而实现一些匹配结果的调整等功能。

接下来为大家通过实战来讲。

贪婪模式与懒惰模式

贪婪模式的核心点就是尽可能多的匹配，而懒惰模式的核心点就是尽可能少的匹配。

接下来为大家通过实战来讲。

正则表达式函数

正则表达式函数有re.match()函数、 re.search()函数、全局匹配函数、re.sub()函数,接下来我们为大家进行分别讲解。

常见正则实例

接下来为大家讲解如何匹配.com或.cn网址，以及如何匹配电话号码。

简单的爬虫

简单的爬虫很好写，直接使用urllib即可编写，接下来我们为大家讲解如何爬取csdn。

从网页中提取出qq群

接下来为大家讲解如何爬取csdn的一个课程页，并自动提取出qq群。

作业：提取出版社信息并写入文件中

我们接下来为大家讲解，如何从https://read.douban.com/provider/all中将所有出版社提取出来，把无关信息过滤掉。

作业提交地址：http://task.iqianyue.com