Python爬虫入门（1）：综述

* [Python爬虫入门（1）：综述](http://python.jobbole.com/81332/)
* [Python爬虫入门（2）：爬虫基础了解](http://python.jobbole.com/81334/)
* [Python爬虫入门（3）：Urllib库的基本使用](http://python.jobbole.com/81336/)
* [Python爬虫入门（4）：Urllib库的高级用法](http://python.jobbole.com/81339/)
* [Python爬虫入门（5）：URLError异常处理](http://python.jobbole.com/81341/)
* [Python爬虫入门（6）：Cookie的使用](http://python.jobbole.com/81344/)
* [Python爬虫入门（7）：正则表达式](http://python.jobbole.com/81346/)
* [Python爬虫入门（8）：Beautiful Soup的用法](http://python.jobbole.com/81349/)

大家好哈，最近博主在学习Python，学习期间也遇到一些问题，获得了一些经验，在此将自己的学习系统地整理下来，如果大家有兴趣学习爬虫的话，可以将这些文章作为参考，也欢迎大家一共分享学习经验。

Python版本:2.7，Python 3请另寻其他博文。

首先爬虫是什么？

网络爬虫（又被称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动的抓取[万维网](http://cpro.baidu.com/cpro/ui/uijs.php?c=news&cf=1001&ch=0&di=8&fv=17&jk=bc21e4bc63a1cc37&k=%CD%F2%CE%AC%CD%F8&k0=%CD%F2%CE%AC%CD%F8&kdi0=0&luki=4&n=10&p=baidu&q=35091180_cpr&rb=0&rs=1&seller_id=1&sid=37cca163bce421bc&ssp2=1&stid=0&t=tpclicked3_hc&tu=u1959022&u=http%3A%2F%2Fcuiqingcai%2Ecom%2F927%2Ehtml&urlid=0)信息的程序或者脚本。

根据我的经验，要学习Python爬虫，我们要学习的共有以下几点：

* Python基础知识
* Python中urllib和urllib2库的用法
* Python正则表达式
* Python爬虫框架Scrapy
* Python爬虫更高级的功能

## 1.Python基础学习

首先，我们要用Python写爬虫，肯定要了解Python的基础吧，万丈高楼平地起，不能忘啦那地基，哈哈，那么我就分享一下自己曾经看过的一些Python教程，小伙伴们可以作为参考。

### 1) 慕课网Python教程

曾经有一些基础的语法是在慕课网上看的，上面附有一些练习，学习完之后可以作为练习，感觉效果还是蛮不错的，不过稍微遗憾的是内容基本上都是最基础的，入门开始的话，就这个吧

学习网址：慕课网Python教程

### 2) 廖雪峰Python教程

后来，我发现了廖老师的Python教程，讲的那是非常通俗易懂哪，感觉也是非常不错，大家如果想进一步了解Python就看一下这个吧。

学习网址：[廖雪峰Python教程](http://www.liaoxuefeng.com/wiki/001374738125095c955c1e6d8bb493182103fac9270762a000)

### 3) 简明Python教程

还有一个我看过的，简明Python教程，感觉讲的也不错

学习网址：[简明Python教程](http://woodpecker.org.cn/abyteofpython_cn/chinese/pr01.html#s01)

## 2.Python urllib和urllib2 库的用法

urllib和urllib2库是学习Python爬虫最基本的库，利用这个库我们可以得到网页的内容，并对内容用正则表达式提取分析，得到我们想要的结果。这个在学习过程中我会和大家分享的。

## 3.Python 正则表达式

Python正则表达式是一种用来匹配字符串的强有力的武器。它的设计思想是用一种描述性的语言来给字符串定义一个规则，凡是符合规则的字符串，我们就认为它“匹配”了，否则，该字符串就是不合法的。这个在后面的博文会分享的。

## 4.爬虫框架Scrapy

如果你是一个Python高手，基本的爬虫知识都已经掌握了，那么就寻觅一下Python框架吧，我选择的框架是Scrapy框架。这个框架有什么强大的功能呢？下面是它的官方介绍：

HTML, XML源数据 选择及提取 的内置支持  
提供了一系列在spider之间共享的可复用的过滤器(即 Item Loaders)，对智能处理爬取数据提供了内置支持。  
通过 feed导出 提供了多格式(JSON、CSV、XML)，多[存储](http://cpro.baidu.com/cpro/ui/uijs.php?c=news&cf=1001&ch=0&di=8&fv=17&jk=bc21e4bc63a1cc37&k=%B4%E6%B4%A2&k0=%B4%E6%B4%A2&kdi0=0&luki=2&n=10&p=baidu&q=35091180_cpr&rb=0&rs=1&seller_id=1&sid=37cca163bce421bc&ssp2=1&stid=0&t=tpclicked3_hc&tu=u1959022&u=http%3A%2F%2Fcuiqingcai%2Ecom%2F927%2Ehtml&urlid=0)后端(FTP、S3、本地[文件系统](http://cpro.baidu.com/cpro/ui/uijs.php?c=news&cf=1001&ch=0&di=8&fv=17&jk=bc21e4bc63a1cc37&k=%CE%C4%BC%FE%CF%B5%CD%B3&k0=%CE%C4%BC%FE%CF%B5%CD%B3&kdi0=0&luki=1&n=10&p=baidu&q=35091180_cpr&rb=0&rs=1&seller_id=1&sid=37cca163bce421bc&ssp2=1&stid=0&t=tpclicked3_hc&tu=u1959022&u=http%3A%2F%2Fcuiqingcai%2Ecom%2F927%2Ehtml&urlid=0))的内置支持  
提供了media pipeline，可以 自动下载 爬取到的数据中的图片(或者其他资源)。  
高扩展性。您可以通过使用 signals ，设计好的API(中间件, extensions, pipelines)来定制实现您的功能。  
内置的[中间件](http://cpro.baidu.com/cpro/ui/uijs.php?c=news&cf=1001&ch=0&di=8&fv=17&jk=bc21e4bc63a1cc37&k=%D6%D0%BC%E4%BC%FE&k0=%D6%D0%BC%E4%BC%FE&kdi0=0&luki=7&n=10&p=baidu&q=35091180_cpr&rb=0&rs=1&seller_id=1&sid=37cca163bce421bc&ssp2=1&stid=0&t=tpclicked3_hc&tu=u1959022&u=http%3A%2F%2Fcuiqingcai%2Ecom%2F927%2Ehtml&urlid=0)及扩展为下列功能提供了支持:  
cookies and session 处理  
HTTP 压缩  
HTTP 认证  
HTTP 缓存  
user-agent模拟  
robots.txt  
爬取深度限制  
针对非英语语系中不标准或者错误的编码声明, 提供了自动检测以及健壮的编码支持。  
支持根据模板生成爬虫。在加速爬虫创建的同时，保持在大型项目中的代码更为一致。详细内容请参阅 genspider 命令。  
针对多爬虫下性能评估、失败[检测](http://cpro.baidu.com/cpro/ui/uijs.php?c=news&cf=1001&ch=0&di=8&fv=17&jk=bc21e4bc63a1cc37&k=%BC%EC%B2%E2&k0=%BC%EC%B2%E2&kdi0=0&luki=5&n=10&p=baidu&q=35091180_cpr&rb=0&rs=1&seller_id=1&sid=37cca163bce421bc&ssp2=1&stid=0&t=tpclicked3_hc&tu=u1959022&u=http%3A%2F%2Fcuiqingcai%2Ecom%2F927%2Ehtml&urlid=0)，提供了可扩展的 状态收集工具 。  
提供 [交互式](http://cpro.baidu.com/cpro/ui/uijs.php?c=news&cf=1001&ch=0&di=8&fv=17&jk=bc21e4bc63a1cc37&k=%BD%BB%BB%A5%CA%BD&k0=%BD%BB%BB%A5%CA%BD&kdi0=0&luki=3&n=10&p=baidu&q=35091180_cpr&rb=0&rs=1&seller_id=1&sid=37cca163bce421bc&ssp2=1&stid=0&t=tpclicked3_hc&tu=u1959022&u=http%3A%2F%2Fcuiqingcai%2Ecom%2F927%2Ehtml&urlid=0)shell终端 , 为您测试XPath表达式，编写和调试爬虫提供了极大的方便  
提供 System service, 简化在生产环境的部署及运行  
内置 Web service, 使您可以监视及控制您的机器  
内置 Telnet终端 ，通过在Scrapy进程中钩入Python终端，使您可以查看并且调试爬虫  
Logging 为您在爬取过程中捕捉错误提供了方便  
支持 Sitemaps 爬取  
具有缓存的DNS解析器

官方文档：<http://doc.scrapy.org/en/latest/>

等我们掌握了基础的知识，再用这个 Scrapy 框架吧！

扯了这么多，好像没多少有用的东西额，那就不扯啦！

下面开始我们正式进入爬虫之旅吧！