**爬虫学习**

1. **爬虫原理**
2. **什么是爬虫**

请求网站，并提取数据的自动化程序

1. **爬虫工作流程**

* 第一步：发起请求。一般是通过HTTP库，对目标站点进行请求。等同于自己打开浏览器，输入网址。
* 第二步： 获取响应内容（response）。如果请求的内容存在于服务器上，那么服务器会返回请求的内容，一般为：HTML，二进制文件（视频，音频），文档，Json字符串等。
* 第三步：解析内容。对于用户而言，就是寻找自己需要的信息。对于Python爬虫而言，就是利用正则表达式或者其他库提取目标信息。
* 第四步：持久化数据。解析得到的数据可以多种形式，如文本，音频，视频保存在本地。

1. **请求与响应**
2. **爬虫与反爬虫**

* headers
  + Referer
  + UserAgent
  + cookie
* IP被封
* 验证码
* JS渲染
* 接口加密与JS混淆
* 行为分析

1. **开发工具**

* Charles
* Chrome、firefox等浏览器
* Pycharm

1. **Python爬虫基础篇**
2. Urllib库使用详解与项目实战

* Urllib库的基本使用
* urllib.request 请求模块
* urllib.error 异常处理模块
* urllib.parse url 解析模块
* urllib.robotparser 解析robots.txt协议模块

1. Urllib3库的使用

* 第三方库，最终也是使用python自带的http模块发送数据包给服务器
* requests使用了urllib3
* pip也使用了urllib3
* 多次请求中可重复利用同一socket连接,应用了keepalive特性，减少TCP握手次数和慢启动次数
* 支持File传输
* 内置重定向和重试
* 支持gzip和deflate解码
* 线程安全
* 支持代理

1. Requests库安装使用与项目实战
2. Pyquery解析库详解与基本使用
3. BeautifulSoup4解析库安装使用

xpath、BS4

1. **Python爬虫进阶篇**

破解反爬技术

1. JS渲染与参数加密
2. 模拟浏览器

* Selenium（webdriver)
* Phantomjs

1. Pyexecjs引擎解析JS
2. 验证码识别

* 图形验证码识别
* 极验验证码识别
* 点触验证码识别
* 微博宫格验证码识别
* 打码平台对接

1. IP代理池实现
2. pyspider框架
3. Scrapy框架
4. 分布式爬虫
5. 爬虫部署