爬虫快速上手

使用 aspider 写第一个爬虫

凤凰山 github.com/gxtrobot/aspider

2019-11

爬虫快速上手 - 使用 aspider 写第一个爬虫

课程内容

今天的课程将介绍爬虫的基本原理，并使用 aspider 库快速编写一个爬虫实战程序

aspider 库地址：https://github.com/gxtrobot/aspider

课程目标

- 了解爬虫基本原理
- 了解爬虫程序基本组成部分
- 使用传统 requests 库编写一个爬虫
- 使用 aspider 编写同样爬虫，并进行对比
- 了解 asipder 库完成的工作，以及特点

爬虫基本原理

所谓爬虫程序，基本就是利用程序获取远程服务器的页面或数据，进行分析提取有效部分，并存储以便后续处理的过程

基本步骤

1. 确定目标爬取根页面
2. 分析页面并提取有效信息，包括更多的目标爬取链接页面
3. 将更多链接加入处理队列
4. 从队列提取一个新页面链接，获取有效信息，并获取新链接，回到步骤 3
5. 直到队列所有页面都被处理，并且没有新页面被发现
6. 爬虫程序结束

代码实战演示

使用 requests 编写爬虫

代码见 github 仓库的 example/douban_requests.py

使用 aspider 编写爬虫

代码见 github 仓库的 example/douban_aspider.py

aspider 库总结

aspider 库完成的工作

- 处理链接自动发现
- 内部维护新链接的队列
- 高并发
- 提供爬虫处理报告

使用 aspider 库写爬虫步骤

1. 定义链接发现的正则表达式
2. 编写页面内容提取函数

爬虫其他相关知识及库

- 匹配 url, 提取内容 (正则表达式, python re 库, Beautfaul Soup, request_html)
- 定位 html 元素 (Css selector , Xpath)
- 模拟用户真实操作 (Selenium)
- 突破反爬机制 (代理池)

课后练习

- 写一个 daouban top 250 的爬虫，这次需要提取电影的标题，以及评分

(代码见 github 仓库的 example/douban_250_scores.py)

- 基于以上爬虫程序，提取其他电影信息

问题