《Python网络爬虫与信息提取》大作业说明

icourse163账号：[jinruiyx@163.com](mailto:jinruiyx@163.com)

昵称：RichardJin

项目概要：获取中国主要省份/城市的空气质量数据并写入csv文件。

技术路线：requests库获取页面，BeautifulSoup库解析页面，正则表达式匹配<span>标签内string内容，并写入列表。最后将列表按行写入csv文件。

具体的：

1. 网址：http://aqicn.org/map/cn/
2. robots\_test(url)函数测试是否有robots协议；
3. get\_HTMLText(url)函数获取页面；
4. getStationAQI(city\_name, city\_url)函数获取数据并return一个列表；
5. 传入一个列表，用内建的csv库写入csv文件。

数据源网站AQI值每小时更新一次，设想将本脚本布置在服务器上每小时运行一次，积累AQI数据。

后续工作：将站点AQI值与地址位置绑定，并进行数据可视化分析。

个人Github地址：<https://github.com/ds17>

本课程的练习/作业以及自己平时写的一些脚本都上传至github中。请老师多多指教。

非常感谢老师带我走入爬虫的大门。谢谢！

以上！