本科生毕业设计（论文）开题报告

题　　目：基于 python 的数据信息爬取及相关特征分析

院　　系　光学与电子信息学院

专业班级　电子 1404

姓　　名　宁泽骥

学　　号　U201413720

指导教师　郑立新

2018　年 2　月

开题报告填写要求

一、 开题报告主要内容：

1.课题来源、目的、意义。

2.国内外研究现况及发展趋势。

3.预计达到的目标、关键理论和技术、主要研究内容、完成课题的方案及主要措施。

4.课题研究进度安排。

5.主要参考文献。

二、 报告内容用小四号宋体字编辑，采用 A4 号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。

三、 指导教师评语、教研室（系、所）或开题报告答辩小组审核意见用蓝、黑钢笔手写或小四号宋体字编辑，签名必须手写。

四、 理、工、医类要求字数在 3000 字左右，文、管类要求字数在 2000 字左右。

五、 开题报告应在第八学期第二周之前完成。

一、课题来源

随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。搜索引擎(Search Engine)，例如传统的通用搜索引擎 AltaVista，Yahoo!和 Google 等，作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。但是，这些通用性搜索引擎也存在着一定的局限性。

二、课题目的

为了解决这些问题，定向抓取相关网页资源的聚焦爬虫应运而生。聚焦爬虫是一个自动下载网页的程序，它根据既定的抓取目标，有选择的访问万维网上的网页与相关的链接，获取所需要的信息。与通用爬虫(general purpose web crawler)不同，聚焦爬虫并不追求大的覆盖，而将目标定为抓取与某一特定主题内容相关的网页，为面向主题的用户查询准备数据资源。

因为 python 的脚本特性，python 易于配置，对字符的处理也非常灵活，加上 python 有丰富的网络抓取模块，所以两者经常联系在一起，因此本课题的研究就十分有意义。

该课题着重于使用 python 爬虫获取到 12306 的车票信息，对获取到的信息处理后存储到相应数据库中，最后使用相应的数据分析第三方库进行相应的数据分析，可以分析出车票的出售等情况。

三、课题意义

网络爬虫已经发展了很多年，并且搜索引擎也是爬虫的一种应用，通过搜索引擎能够更快速的获得有用的数据和信息。但是，一些通用性的搜索引擎也存在着一定的局限性，通用搜索引擎返回的结果可能包含了大量用户不关心的网页内容，而且通用搜索引擎有限的服务器资源与无限的网络资源之间存在的矛盾进一步加深，还有，就是通用搜索引擎不能支持给据语义的信息提出的查询和搜索。所以学习网络爬虫有很大的意义。

作为毕业设计，这一课题极具挑战性和概括力，是对过去四年所学的巩固，是对新知识的一次探索，是对网络资源获取的一个很好的演练。

四、国内外研究现况及发展趋势

Python 为我们提供了非常完善的基础代码库，覆盖了网络、文件、GUI、数据库、文本等大量内容，被形象地称作“内置电池（batteries included）”。用 Python 开发，许多功能不必从零编写，直接使用现成的即可。

因为 Python 是跨平台的，它可以运行在 Windows、Mac 和各种 Linux/Unix 系统上。在 Windows 上写 Python 程序，放到 Linux 上也是能够运行的。

Python 爬虫架构主要由五个部分组成，分别是调度器、URL 管理器、网页下载器、网页解析器、应用程序（爬取的有价值数据）。

调度器：相当于一台电脑的 CPU，主要负责调度 URL 管理器、下载器、解析器之间的协调工作。

URL 管理器：包括待爬取的 URL 地址和已爬取的 URL 地址，防止重复抓取 URL 和循环抓取 URL，实现 URL 管理器主要用三种方式，通过内存、数据库、缓存数据库来实现。

网页下载器：通过传入一个 URL 地址来下载网页，将网页转换成一个字符串，网页下载器有 urllib2（Python 官方基础模块）包括需要登录、代理、和 cookie，requests(第三方包)

网页解析器：将一个网页字符串进行解析，可以按照我们的要求来提取出我们有用的信息，也可以根据 DOM 树的解析方式来解析。网页解析器有正则表达式（直观，将网页转成字符串通过模糊匹配的方式来提取有价值的信息，当文档比较复杂的时候，该方法提取数据的时候就会非常的困难）、html.parser（Python 自带的）、beautifulsoup（第三方插件，可以使用 Python 自带的 html.parser 进行解析，也可以使用 lxml 进行解析，相对于其他几种来说要强大一些）、lxml（第三方插件，可以解析 xml 和 HTML），html.parser 和 beautifulsoup 以及 lxml 都是以 DOM 树的方式进行解析的。

应用程序：就是从网页中提取的有用数据组成的一个应用。

综上，在浏览器网络爬虫领域，很多技术已经很充分了，并且解决了诸如，数据爬取，页面分析等问题。但在根据语义爬取数据和聚焦爬虫的方面，还并不是很完善，万维网上大量的有用的信息，或者语义相近的有用信息得不到利用，这是一种巨大的损失，所以对于爬虫的研究还是非常有必要的。

五、预计达到的目标与主要研究内容

通过对 Python 基础知识的学习和对其网络相关的理论的研究，以及对相关的一些涉及到的第三方库比如 Tkinter、mongoDB 数据库的了解，我预计能达到设计一个简单的 UI 层界面供调用，在调用层实现对 12306 车票的查询，并对接收到的数据进行处理、收集，并存放到数据库中。最后对数据库中的车票、车次数据进行有关的数据分析，通过该课题的学习与设计，为以后的更深的爬虫学习奠定基础。

六、关键理论和技术

本课题有关的理论与技术较多，可以归纳为以下：

数据调试:Chrome 浏览器的开发者工具，可以看到网络请求，记录当前页面的网络操作，包括详细的时间，http 请求和相应，cookies 等，并可以用来操作 DOM 和样式，可以直接在上面进行编辑。

发起请求: 使用 http 库向目标站点发起请求，即发送一个 Request，Request 包含：请求头、请求体等 ，Request 模块缺陷：不能执行 JS 和 CSS 代码。

获取响应内容：如果服务器能正常响应，则会得到一个 Response，Response 包含：html，json，图片，视频等。

解析内容：解析 html 数据：正则表达式（RE 模块），第三方解析库如 Beautifulsoup，pyquery 等，解析 json 数据：json 模块，解析二进制数据：以 wb 的方式写入文件。

保存数据：数据库（MySQL，Mongdb、Redis），文件。

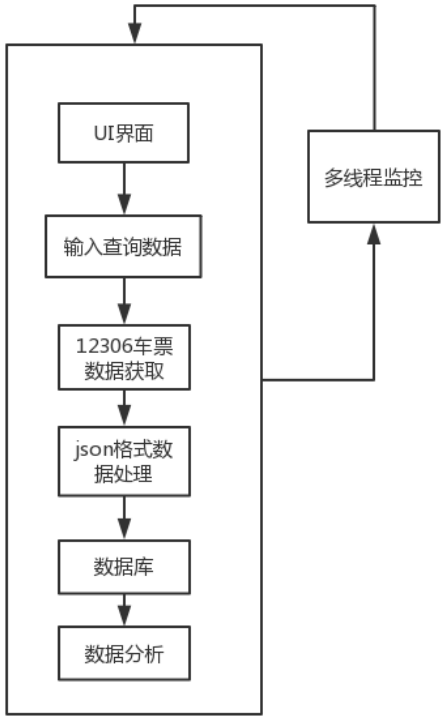
多线程的监控：therading.Thread 类中的并发执行任务。

爬虫框架：使用比较成熟的 Python 开源爬虫框架，来高效地爬虫内容。如 Scrapy 框架、Crawley 框架、Portia 框架、newspaper 框架、python-goose 框架。

七、完成课题的方案及主要措施

首先，我会大量阅读有关 python 的中英文教程资料，并安装 python 相关的开发环境，熟悉从理论上了解和认识相应的 python 常用库的基本原理和方法，熟悉有关 12306 车票数据的数据结构。然后查阅相关书籍与资料，熟悉如何用 Python 去获取相应的车票数据。在做了这些准备工作之后开始研究如果将数据存放到数据库中，并为课题加上简单的 UI 组件，并加以改进，可调参来实现。最后将爬取数据作具体分析，得出可视化的结论。

以下是该课题的主要实现思路如下图 1 所示：



图片 1

主要思路叙述：如图 1 所示，该爬虫可分为几个模块来开发，在爬取层，依情况使用 Scrapy 框架或者使用原生的 urllib 库来进行开发。在 UI 层使用相应的库来设置几个输入参数作为整个爬虫项目的参数，如“始发站”、“终点站”与“日期”等。在数据层，依据信息获取是否是 html 还是 json 格式的数据来决定是用正则表达式来从网页数据中获取，还是使用 json 解析库来对 json 数据进行解析。对于数据，学习使用 mangoDB 与 MySQL 后，将数据存放到数据库里面。而对于数据的分析，我们可以从车票数据中获取到相应的信息，比如可以知道某车票每天的销量等一般查询不到的信息。多线程模块的设置是为了让爬虫能在我们规定的时间内循环爬取车票数据。

可能会出现的问题及解决思路：

1. 频繁获取数据后 12306 网站可能会要求验证码识别，所以也有必要对验证码进行相应的处理。

2. 12306 网站只能在每天的 6:00-23:00 获取车票数据，所以完成的时间也需要限定在这个时间。

3. 12306 网站对于每个 IP 车票数据的获取会有所限制，可能需要使用其他代理的 IP。

八、课题研究进度安排

表 1□课题研究进度安排表

学期	周次	工作任务
2017-2018 第一学期	19 周——20 周	完成外文翻译
2017-2018 第二学期	1 周——2 周	完成开题报告、文献调研
	3 周——6 周	完成 python 的学习、课题提及的模块涉及库的学习
	7 周——12 周	深度学习模块的设计，完整的软件设计
	13 周——14 周	软件的测试及调优
	15 周——17 周	撰写论文、论文答辩

九、主要参考文献

[1]□韦玮.精通 Python 网络爬虫——核心技术、框架与项目实战.机械工业出版社,2017.

[2]□春. Python 核心编程 ： 第 3 版[M]. 人民邮电出版社, 2016.

[3]□翟自洋, 林昌东. 利用正则表达式进行查找/替换[J]. 中国科技期刊研究, 2009, 20(1):122-126.

[4]□.python 编程教程 https://www.liaoxuefeng.com/wiki/00143160895572 64a

[5]□.python 参考手册 https://docs.python.org/3/

[6]□.自己动手写网络爬虫 罗刚 清华大学出版社 2010.10

[7]□.如何轻松爬取网站：https://www.cnblogs.com/liuliliuli2017/p/6809 083.html

[8]□. HOWTO Fetch Internet Resources Using The urllib Package.[EB/OL] https://docs.python.org/3.1/howto/urllib2.html.1990-2012

[9]□.Basic Authentication - Authentication with Python.[EB/OL] http: //www.voidspace.org.uk/python/articles/authentication.shtml

华中科技大学本科生毕业设计（论文）开题报告评审表

<table>
<tr><td>姓名</td><td></td><td>学号</td><td></td><td>指导教师</td><td></td></tr>
<tr><td colspan="2">院（系）专业</td><td colspan="4"></td></tr>
<tr><td colspan="6">指导教师评语
1. 学生前期表现情况。
2. 是否具备开始设计（论文）条件？是否同意开始设计（论文）？
3. 不足及建议。</td></tr>
<tr><td colspan="6">

指导教师（签名）：

年　　月　　日</td></tr>
<tr><td colspan="6">教研室（系、所）或开题报告答辩小组审核意见</td></tr>
<tr><td colspan="6">

教研室（系、所）或开题报告答辩小组负责人（签名）：

年　　月　　日</td></tr>
</table>