基于Python语言的互联网数据收集软件的设计

林亦凡　刘孟雄(长安大学,陕西　西安　710054)

摘　要:随着数据的开放,互联网上公布的许多数据均可为科研所用。利用人工收集的方法获取这些数据费时费力,而采用网络爬虫技术采集这些数据可有效的提升我们的工作效率。本文利用Python语言及其丰富的第三方库,编写了对空气质量,气象等各类数据进行抓取的爬虫,并为其设计GUI,将其打包成可以在Windows下运行的软件。该软件操作简单,界面友好,对非计算机专业的人员运用网络爬虫收集科研数据提供了便利条件。

关键词:Python;网络爬虫;空气质量数据

网络爬虫是用户获取互联网数据的有效工具,可以用于编写网络爬虫的语言亦有多种,其中比较常用的语言有python,java,C++等。Python是一种面向对象的解释型计算机程序设计语言,可用于诸多方面,如:做Web前端与后端,设计软件的GUI,大数据分析等等。

1 基于Python语言的网络爬虫建立

本软件的编写环境与Python第三方库见表1。Requests库基于Python编写,它比urllib更加方便,可以节约我们大量的工作,完全满足HTTP测试需求[1]。BeautifulSoup是一个可以从HTML或XML文件中提取数据的Python库。PyQt5是QT专门为Python所写的第三方库,可以用于软件GUI的编写。Pyinstaller可以将Python的".py"文件封装成可以在Windows电脑上运行的".exe"可执行文件。

表1 软件建立所需的工具及其版本

编写环境与IDE	第三方库与版本号	库的作用
Python3.5.2 Windows10 PyCharm 2016.3 Sublime Text3	Requests 2.12.1	用于爬虫编写
	BeautifulSoup 4.5.1	用于爬虫编写
	PYQT5 5.8.1	用于软件GUI设计
	Pyinstaller	用于封装python程序

1.1 百度地图数据的抓取

一般我们研究区域污染状况成因时,需要找出污染源的坐标,就宏观研究而言,百度地图是一个很好的选择。经过分析,百度地图提供一个接口:http://map.baidu.com/?newmap=1&reqflag=pcmap&biz=1&from=webmap&da_par=direct&pcevaname=pc4.1&qt=con&from=webmap&c=233&wd={}&pn={}。其中wd代表的是所需的查找内容,pn用于翻页。需要指出的是,如果访问过于快速,会出现访问不成功现象,这时需要放慢抓取速度与设置断点续传的功能[3]。抓取数据坐标是百度墨卡托坐标,与经纬度坐标有很大出入,需要进行坐标转换。

1.2 空气质量数据的抓取设计

本软件的空气质量数据来源于网站:http://www.pm25.in/。在抓取中由于PC端的网页采取的Ajax所以学采用Selenium与Phantomjs抓取,但是会大量占用电脑内存与CPU。故笔者将Headers设置成手机以便正常抓取。由于代码较长不在此放出。

1.3 空间热度图与点密度图绘制的设计

可以运用Python的标准库matplotlib绘制空间的点密度,热密度图,及将上述地图中收集的数据反应的地图中。本软件只是应用hexbin函数做了简单的绘制。由于需要将matplotlib嵌入PyQt5开发的窗口内代码较长,不再给出。

2 基于PyQt5的GUI设计

2.1 功能的描述

运用PyQt5进行GUI设计,将上述6个爬虫进行封装[2]。其中空气质量按钮对应空气质量数据收集(自动每一小时收集一次)。其中,坐标转换按钮为百度地图收集的数据的坐标转换(需要key)。

2.2 界面的设计与运行情况

为了简化界面,将界面中的菜单栏与状态栏统统删除,只留下6个爬虫的相关按钮并尽可能放大,,如图2。为了保证界面的美观性,不再允许用户将界面最大化,即固定主界面的大小。其中空气质量数据,气象数据,百度地图,高德地图的二级界面基本一致,输入爬取内容与保存位置点击开始即可。

对于地图绘制功能,是本软件实现的难点内容。需要将matplotlib画布嵌到界面中,实现起来有些复杂[3]。该界面提供了西安市及其各个区的轮廓图、点密度与热密度图的绘制。

图2 软件主界面

3 讨论与展望

运用Python编写爬虫软件,以用于科研相关数据的抓取,可以更好的服务于科学研究。Python语言应用广泛,在统计分析方面也正在追赶R语言,在数学建模与数据分析方面亦正在追赶MATLIB,且其编写桌面程序的开发效率较高,可以较快的实现某些功能。

参考文献:

[1]狄博,王晓丹.基于Python语言的面向对象程序设计课程教学[J].计算机工程与科学,2014,36(S1):122-125.

[2]钱程,阳小兰,朱福喜.基于Python的网络爬虫技术[J].黑龙江科技信息,2016,(36):273.

[3]康计良.Python语言的可视化编程环境的设计与实现[D].西安电子科技大学,2012.