Python 网络爬虫与数据分析

摘　要 近年来，随着互联网的广泛使用、发展和深化，对于公司以及个人而言如何准确高效的获取想要的数据已经成为现代互联网发展的一项挑战。特别是在大数据时代，如何高效准确的获取更多即时数据也成为了一家公司以及个人的一项挑战。在这种情况下网络爬虫应运而生，成为企业以及个人在线提取数据最常用的方法之一。本文就是主要是基于网络爬虫即 web spider 来准确高效的获取互联网上的数据。本项目主要分三个步骤来实现，第一：网络数据爬取部分，采用 Scrapy 框架来爬取最近更新的全省各市的疫情情况和我国自疫情爆发至今每天的疫情情况，采用普通爬虫爬取最近人们最关心的问题即百度热搜。最后将爬取的数据保存至数据库。第二：数据读取及处理部分，将存储的数据从数据库读出，并在 Flask 框架中做数据的进一步处理。第三：数据可视化展示，使用 HTML 和 JavaScript、Echarts 编写网页框架及规划框架内容，最后完成 Flask 中数据与 Echarts 图表数据的对接。本项目采用的是 python 开发语言，python 是一个解释性、动态面向对象的高级语言程序，同时也是一个代码简便、易懂的语言。是网络爬虫的首选语言。

关键词 python; 爬虫; 数据; Web; HTML; JavaScript; Flask; Scrapy; 疫情; 热搜;

1 概述

1.1 课题背景

随着国与国之间的交往日益增长，经济交流也与日俱增。在经济全球化的大背景之下，爆发了新冠状肺炎疫情，就在最近短暂一个月的时间内新冠状肺炎病毒也已经实现了“全球化”。就在今年的 3 月 27 日，世界卫生组织向全球公布的关于新冠状肺炎疫情数据显示，自新冠肺炎疫情爆发至今疫情已经影响到了全球 200 多个国家和地区，全球各地新冠肺炎确诊病例人数已经超过 60 万例，而且全球的确诊人数还在增加。此次疫情在全球化不断深化的大背景下爆发，对各国的经济发展产生了不可估量的后果，这不仅是对世界经济全球化的产生了重大灾难也对世界各国人民产生了重大的灾难。此次新冠肺炎疫情的“全球化”过程表明，当一国爆发疫情时，其他国家也不可能独善其身置身事外。虽然我国的新冠肺炎疫情已经得到了有效控制，但是全球范围内的频繁人员往来使得我国外来病例逐渐增多。国内人民的安全还存在威胁，我们还不能放松警惕还应该时时刻刻关注我国的疫情情况，做好安全防护。根据疫情情况合理安排学习、工作和生活，保证自己安全就是对祖国人民负责。本课题就是在这样的背景下产生了，爬虫爬取全国疫情数据并实现可视化处理，更好的了解我国的最新疫情情况。

1.2 课题研究的目的

本课题的主要目的有两个：

第一：互联网是非结构化数据库。有效地搜索具有特别大的应用前景。现在人们都通过互联网来获取中国以及世界上的最新信息或者历史信息，通过互联网来获取世界各地的奇闻异事，浏览器成为人们获取信息的重要选择。然而，这些最常用浏览器也有存在一些限制。不同人们想要浏览的信息，想要获取的数据往往有所差别。一般搜索引擎返回的结果常常会包含一些人们不想得到的信息或数据。为了解决这一问题，爬虫逐渐走进了人们的视野，成为了一种重要的选择。通过本课题能更好的使我了解爬虫的基本原理、Scrapy 框架的架构和各组件的功能、深入了解 MySQL 数据库的基本语句的使用并实现使用 python 操作数据库、懂得如何将 Echarts 引入 HTML 代码并使用 JavaScript 实现商业级图表 Echarts 的使用。

第二：爬虫内容是爬取中国最近的疫情情况，中国自疫情爆发至今的时间内每天的确诊人数、治愈人数、死亡人数、疑似人数等相关数据以及最近人们关于疫情问题最关心的问题。通过爬虫能更好的了解我国最新的疫情情况，做好个人防护。

1.3 网络爬虫的定义

网络爬虫是一种能够模仿人们搜索信息浏览网页的过程，根据人们设定的搜索目标自动获取网页信息。从而得到人们想要的文字、图片或数据等信息，具有确定性不会包括一些人们不想要的数据。只需要在开始指定想要获取信息的网页地址，爬虫就可以运行获取包含需要的到信息在内的网页源代码，然后通过一些解析函数，就可以提取出想要的信息。对于获得的信息可以进行后期处理，如：存储到数据库、csv 文件等。

2 系统开发平台及运行环境

2.1 系统开发平台

Pycharm 是一个 pythonIDE，它有一套完善的工具可以使用户在使用 python 时可以提高效率。它还有很多高级功能，支持 Django, Flask 在内的专业网络开发。Pycharm 成为 python 初学者最受欢迎的强大工具。主要功能：

一、智能编码辅助：提供智能代码完成、错误代码显示、代码检查和导航功能。

二、内置开发人员工具：代码调试、远程开发、数据库工具等。

三、用于 web 开发：它提供特定的开发框架，支持 JavaScript、HTML、CSS 代码的编写。

四、科学分析工具：支持数据分析及数据展示的库函数，包括：Anaconda、Matplotliib、NumPy 等软件包。