详细设计说明书

所属学校：青岛科技大学

参赛项目：基于互联网数据的大数据分析

团队名称：GOODLUCK 队

团队成员：张莹、刘桐源、邵华、乔善宝、寻宇星

指导老师: 周艳平

2017 年 9 月

目　　录

第一部分 前言

1.1 编写目的

本文档说明程序设计的关键部分,即详细设计说明书，另其他详细代码实现请查看代码源。且本文档是描述网络爬虫、数据库技术、可视化技术设计文档。

1.2 应用背景

1.2.1 发展背景

进入2012年，大数据(big data)一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然现在企业可能并没有意识到数据爆炸性增长带来问题的隐患，但是随着时间的推移，人们将越来越多的意识到数据对企业的重要性。大数据时代对人类的数据驾驭能力提出了新的挑战，也为人们获得更为深刻、全面的洞察能力提供了前所未有的空间与潜力。

近几年来，随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生的数据呈爆炸性增长。动辄达到数百TB甚至数十至数百PB规模的行业/企业大数据已远远超出了现有传统的计算技术和信息系统的处理能力，因此，寻求有效

的大数据处理技术、方法和手段已经成为现实世界的迫切需求。百度目前的总数据量已超过1000PB，每天需要处理的网页数据达到10PB～100PB;淘宝累计的交易数据量高达 100PB;Twitter每天发布超过2亿条消息，新浪微博每天发帖量达到8000万条;中国移动一个省的电话通联记录数据每月可达 0.5PB～1PB;一个省会城市公安局道路车辆监控数据三年可达200亿条、总量120TB。据世界权威IT信息咨询分析公司IDC研究报告预测：全世界数据量未来10年将从2009年的0.8ZB增长到2020年的35ZB(1ZB=1000EB=1000000PB)，10年将增长44倍，年均增长 40%。

早几年人们把大规模数据称为“海量数据”，但实际上，大数据(Big Data)这个概念早在2008年就已被提出。2008年，在Google成立10周年之际，著名的《自然》杂志出版了一期专刊，专门讨论未来的大数据处理相关的一系列技术问题和挑战，其中就提出了“Big Data”的概念。

随着大数据概念的普及，人们常常会问，多大的数据才叫大数据?其实，关于大数据，难以有一个非常定量的定义。维基百科给出了一个定性的描述：大数据是指无法使用传统和常用的软件技术和工具在一定时间内完成获取、管理和处理的数据集。进一步，当今“大数据”一词的重点其实已经不仅在于数据规模的定义，它更代表着信息技术发展进入了一个新的时代，代表着爆炸性的数据信息给传统的计算技术和信息技术带来的技术挑战和困难，代表着大数据处理所需的新的技术和方法，也代表着大数据分析和应用所带来的新

发明、新服务和新的发展机遇。

1.2.2 技术背景 网络爬虫技术

网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列,直到满足系统的一定停止条件。聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

1.3 参考资料

《Python 学习手册》 --Mark Lutz 著

PHP 与 MySQL 程序设计(第 4 版)》 --W. JasonGilmore 著

HTML5+CSS3 从入门到精通 --李东博 著

响应式 Web 设计:HTML5 和 CSS3 实战 --BenFrain (作者), 王永强 (译者)

JavaScript DOM 编程艺术 --**基思 (Jeremy Keith)** (作者), **桑布尔斯 (Jeffrey Sambells)** (作者), **魏忠** (合著者)

JavaScript 高级程序设计 --泽卡斯 (Zakas. Nicholas C.) (作者), 李松峰 (译者), 曹力 (译者)

jQuery API 中文文档

Bootstrap3.0 官方文档

ECharts3 官方配置项手册

第二部分　程序设计说明

2.1 程序描述

在这样的大数据大背景下，我们使用 Python 语言开发网络爬虫用于爬取互联网中 IT 公司提供的计算机岗位，并将结果进行了可视化，方便用户直观的看出个计算机岗位的就业情况。通过比较不同地区、不同岗位

2.1.1 功能主旨

基于互联网数据的大数据分析

2.2.2 函数库介绍

利用 Python 语言中丰富的可用于网络爬虫的库编写网络爬虫，简化了我们编写网络爬虫的难度，使用了 Python 的科学计算库 numpy，NumPy 系统是 Python 的一种开源的数值计算扩展。这种工具可用来

存储和处理大型矩阵，比 Python 自身的嵌套列表（nested list structure)结构要高效的多（该结构也可以用来表示矩阵（matrix））。

2. 2. 3 软件开发人员分派

1.任务提出者：齐鲁大学生软件设计大赛

2.开发者：青岛科技大学 GOODLUCK 团队

2.2.4 运用工具

Python Scripy：

Python 开发的一个快速、高层次的屏幕抓取和 web 抓取框架，用于抓取 web 站点并从页面中提取结构化的数据。Scrapy 用途广泛，可以用于数据挖掘、监测和自动化测试。

Scrapy 吸引人的地方在于它是一个框架，任何人都可以根据需求方便的修改。它也提供了多种类型爬虫的基类，如 BaseSpider、sitemap 爬虫等，最新版本又提供了 web2.0 爬虫的支持。

C++:

C++是 C 语言的继承，它既可以进行 C 语言的过程化程序设计，又可以进行以抽象数据类型为特点的基于对象的程序设计，还可以进行以继承和多态为特点的面向对象的程序设计。C++擅长面向对象程序设计的同时，还可以进行基于过程的程序设计，因而 C++就适应的问题规模而论，大小由之。

MySQL：

是一个关系型数据库管理系统，在 Web 应用方面 MySQL 是最好的 RDBMS(Relational Database Management System：关系数据库管理系

统)应用软件之一。

MySQL C Connector：

MySQL Wrapper（闭源）：

一款 MySQL C API 封装类库

HTML：

HTML用来定义了网页的内容。 HTML 是用来描述网页的一种超文本标记语言 (Hyper Text Markup Language) ，而不是一种编程语言。标记语言是一套标记标签 (markup tag)， HTML 使用标记标签来描述网页。Web 浏览器（如谷歌浏览器，Internet Explorer，Firefox，Safari）是用于读取 HTML 文件，并将其作为网页显示。浏览器并不是直接显示的 HTML 标签，但可以使用标签来决定如何展现 HTML 页面的内容给用户

CSS：

CSS 描述了网页的布局。CSS 指层叠样式表 (Cascading Style Sheets) ，样式将定义如何显示 HTML 元素 ，通常将样式存储在样式表中 ，而添加到 HTML 4.0 中，是为了解决内容与表现分离的问题 。外部样式表可以极大提高工作效率 ，通常把它存储在 CSS 文件中 ，多个样式定义可层叠为一 。

JavaScript：

JavaScript 定义了网页的行为。JavaScript 是脚本语言，是一种轻量级的编程语言，是可插入 HTML 页面的编程代码。JavaScript 插入 HTML 页面后，可由所有的现代浏览器执行。

jQuery：

jQuery 是一个 JavaScript 函数库，可以通过一行简单的标记被添加到网页中。Query 库包含 HTML 元素选取、HTML 元素操作、CSS 操作、HTML 事件函数、JavaScript 特效和动画、HTML DOM 遍历和修改、HTML DOM 遍历和修改 、AJAX、Utilities 等特性：

Bootstrap3.0：

Bootstrap 是一个用于快速开发 Web 应用程序和网站的前端框架，是基于 HTML、CSS、JAVASCRIPT 的。响应式设计：Bootstrap 的响应式 CSS 能够自适应于台式机、平板电脑和手机。它为开发人员创建接口提供了一个简洁统一的解决方案，包含了功能强大的内置组件，易于定制。 它还提供了基于 Web 的定制，是开源的。

ECharts 3：

ECharts，一个纯 Javascript 的图表库，可以流畅的运行在 PC 和移动设备上，兼容当前绝大部分浏览器（IE8/9/10/11，Chrome，Firefox，Safari 等），底层依赖轻量级的 Canvas 类库 ZRender，提供直观，生动，可交互，可高度个性化定制的数据可视化图表。而 ECharts 3 中更是加入了更多丰富的交互功能以及更多的可视化效果，并且对移动端做了深度的优化。