需求规格说明书

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

一、引言

1.1 编写目的

编写本说明书的目的是让软件开发小组的成员能更好的了解市场的需求，了解该大数据产生的背景，以市场为导向。该文档对大数据、数据可视化所需要达到功能、界面及运行环境等作出了详细的说明。他作为对该系统概要设计的依据，帮助开发人员了解本系统的框架思想及实现功能，并验证核实该产品能否满足用户要求的标准，便于技术文档和需求变化的管理。同时也是用户与开发人员双方对软件需求取得共同理解的基础。

软件开发小组的每一位成员都应详细阅读此说明书，明确开发目的，按要求完成软件的开发，经使用方认可的需求说明将作为产品特征评价、仲裁的重要考。

预期读者：大赛评委、产品用户、开发人员，指导老师。

1.2 项目背景

项目名称：基于互联网数据的大数据分析

功能实现：数据抓取、数据存储、数据整理、数据可视化

项目提供者：齐鲁软件设计大赛组委会

开发团队：GOOKLUCK 团队

所属学校：青岛科技大学

项目组成员：张莹、刘桐源、邵华、乔善宝、寻宇星

1.3 定义

网络爬虫：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。

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

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。

Matplotlib: Matplotlib 是一个 Python 的 2D绘图库，它以各种硬拷贝格式和跨平台的交互式环境生成出版质量级别的图形[1] 。

通过 Matplotlib，开发者可以仅需要几行代码，便可以生成绘图，直方图，功率谱，条形图，错误图，散点图等。

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。

1.4 参考资料

1..参考书籍：

《Python 学习手册》 --Mark Lutz 著

http://matplotlib.org/

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

二、项目概述

2.1 产品描述

我们首先写出网络爬虫，使用网络爬虫爬取招聘网站信息，将爬取到的信息存放到数据库中，将原始信息进行清洗得到所需要的数据，对清洗得到的数据进行可视化，使用柱状图、饼图等进行展示。

2.2 用户特点

操作人员会浏览网页即可。

2.3 假设和依据

开发日期为为2017年7月至9月

需要的语言为Python, html,css,javascript,c++

可以根据网络爬虫爬取到的各类信息进行分类整理，然后将整理后的数据进行可视化。

三、具体需求

3.1 功能需求

功能可划分成如下几个方面：

抓取网络数据：使用Python语句爬取著名招聘网站上的招聘信息得

到原始数据

数据存储：对爬取的数据使用MySQL进行存储，使数据与数据之间具有一定的网格结构。

数据整理：通过对原始半结构化数据清洗、转换和汇总形成结构化的数据后，实现（1）计算机专业薪水最高的前十名招聘职位；（2）大数据职位需求量最高的前十名城市；（3）大数据职位需求量最高的前十大行业；（4）计算机专业工作经验要求分布情况；（5）企业对哪类大数据人才需求量最为迫切

并格外增加：

1. 公布有效数据：

爬取数据的来源分布 --> 极坐标系下的堆叠柱状图

不同网站不同领域比例(大数据，开发，测试，运维) -->嵌套环形图

2. 全国范围内的数据分布：

2.1 全国（平均）薪资分布情况 --> 地图散点图，薪资高的颜色深

2.2 不同领域的平均薪资所占比例 --> 【饼图】或者南丁格尔图

2.4 全国公司的城市分布 --> 地图散点图

2.3 在全国岗位需求量分布 --> 饼图或者地图散点图，岗位多的颜色深，

2.5 大数据在全国的岗位需求量分布 --> 柱状图或者地图散点图

2.6 计算机专业工作经验要求分布情况（应届、1～3 年、3～5 年）

--> 嵌套环形图

2.7 全国各行业占比

3. 排行表：

3.1 计算机专业薪水最高的前 10 名招聘职位+岗位需求量 --> 折线图

3.2 大数据职位需求量最高的前 10 名城市+大数据前六个岗位需求量排行 -->极坐标系下的堆叠柱状图-polar-stack-radial

3.3 大数据职位需求量最高的前 10 名行业（如互联网、金融、电子商务等） -->南丁格尔玫瑰图

3.4 企业对哪类大数据人才需求最为迫切（大数据分析师、大数据架构师等）

3.5 计算机专业编程语言职位需求量前十名饼状图 --> 矩形树图

3.6 计算机专业编程语言平均薪资 –>柱状图

3.6 计算机专业不同编程语言平均薪资柱状图

4. 不同要素的关系图表：

4.1 学历和薪资(min,max)关系 --> 折线图

4.2 工作经验和薪资(min,max)的关系 --> 折柱图

数据可视化：使用 Echarts 进行可视化，以饼图、柱状图、折线图、地图、雷达图、玫瑰图等进行数据的展示。