项目开发总结

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 团队

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 引言

1.1 编写目的

1、可以极大的提高学生的大数据挖掘、分析实践经验，有助于大学生对大数据挖掘、分析工具、语言的熟练应用，有助于学校学习的理论知识和实践相结合。

2、可以让学生在校学习时了解哪些职业比较热门，收入较高，自己应该学习哪些课程和技术，有助于学生自我的学习计划。

3、可以对各个高校在设置专业的时候了解企业对人才的要求，对学生人才培养方向上有比较实际的参考价值。

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：对各职业的分布城市、薪资、需求量、工作经验要求、学历要求及公司的分布城市进行分析，并通过图表载入 html 网页展示

项目提供者：齐鲁软件设计大赛组委会

开发团队：goodluck 团队

所属学校：青岛科技大学

项目组成员：张莹、刘桐源、邵华、寻宇星、乔善宝

1.3 定义

Python Scrpy：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。

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

《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

第二部分 自我评价

3.1 生产率评价

程序的平均生产效率：700 行\每人天

3.2 技术方案评价

我们使用 python 的 scrapy 去抓取招聘网站中的数据，事实证明它还是比较高效的，在抓取数量和抓取质量方面，我们都非常满意。在数据可视化方面，我们最开始使用了 python 的 matplotlib 库，可视化的结果将会以静态图片的方式展示，我们都觉得不太美观，经过一番搜索资料，我们发现了 ECharts,从而开启了动态网页展示数据的大门，它的交互性特别强，用户只需轻轻移动鼠标就能看到不同的视觉效果。

3.3 产品质量评价

我们的数据有近一百五十万，涵盖一百多个城市，我们采用多种图表可视化我们的数据，其中包括漏斗图，玫瑰图，雷达图，中国地图，折线图，柱状图以及各种复合图表，做到了非常直观地展示数据，不仅如此，我们还把所有图表嵌入到了 html 中，从而使浏览更加方便。由于招聘网站发布信息的不全面问题，有些城市可能招聘信息较少甚至没有招聘信息，所以这些城市将不会展现在我们的地图上，因为这个原因，有些柱状图可能很短，我们还会不断地丰富我们的数据，从而达到很好的可视化效果。

第三部分 自我总结

通过本次项目开发我们组总结出以下经验和教训：

1.每位成员之间交流不够，项目计划不够详细使得项目进展慢了很多。

2.团队有五个成员，每个人负责不同的部分，所以有的时候没办法交流问题。

3.尽管有着重重的障碍与压力，但是我们积极查找资料，动手实践，最后我们成功地抓取到了近一百五十万的数据，并进行了可视化。

4.我们因为这个项目而收获了很多，在做这个项目之前，我们只会 C++， C 以及 java,现在我们能用 python 写分布式网络爬虫，进行数据分析，可视化，也能用 html,css,javascript 写网页，我们实践了很多，不再像以前一样只会理论知识了。

5.所有的代码都是由我们自己编写的，也许没有可视化工具看起来美观，但是看着自己的代码能展示出可视化结果，我们都很有成就感。