DOI:10.16707/j.cnki.fjpc.2018.11.058

基于 Pyhton 的软件技术人才招聘信息分析与实现
——以前程无忧为例

王 涛
（长沙民政职业技术学院 软件学院 湖南 长沙 410004）

【摘 要】本文以前程无忧招聘网站上的涉及 Java、Python、大数据和安卓四个关键字的岗位在全国多个热门城市的相关招聘信息为数据来源，分析了当前我国这四个类型岗位的地域分布，热点分布和薪资水平的现状，呈现了研究结果的可视化关系，发现了目前我国软件技术专业的招聘需求。

【关键词】Python；数据分析；软件技术

一、软件行业发展与人才需求的背景

从信息技术时代走向数据技术时代，与之对应的是大数据技术的迅猛发展和行业应用需求的快速增长，为了应对大数据行业正面临着全球性的人才荒，加快实施国家大数据战略，推动大数据产业健康快速发展，2017 年 1 月，工业和信息化部印发《大数据产业发展规划（2016–2020 年）》[1]。

2018 年 1–4 月，我国软件和信息技术服务业继续稳中有升，收入增速提高，利润和出口增速保持增长，从业人数和工资总额稳步增加。中部地区软件业增势突出，中心城市软件业保持领先[2]。软件行业和技术不断发展的同时，软件行业对人才的需求也不断发生变化，使用的开发语言和技术随着应用的领域也在不断更迭，这个过程中软件企业对技术人才的需求是不是有了新的变化。针对于此本文开展了使用 Python 语言来实现对前程无忧招聘网站上的软件技术人才招聘信息分析和实现等工作。

二、Python 和数据分析

Python 语言上世纪 90 年代诞生，从 2004 年以后，使用率呈线性增长。2018 年 7 月，它在 Hacker News Hiring Trends 编程语言排行榜[3]中高居榜首。由于 Python 语言的简洁性、易读性以及可扩展性，众多开源的科学计算软件包都提供了 Python 的调用接口，而 Python 专用的科学计算扩展库就更多了，Python 语言及其众多的扩展库所构成的开发环境十分适合工程技术、科研人员处理实验数据、制作图表，甚至开发科学计算应用程序。

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程，还可以发现事物之间的规律。比如，发掘我国不同地区软件企业对人才的需求分布及岗位要求。

三、基于 Python 的数据分析与实现

数据挖掘需要经过数据采集、预处理、数据分析、结果表示等一系列过程，对数据信息进行有效的处理。

3.1 Python 编写爬虫采集数据

根据软件技术专业当前的人才培养方案的课程体系和面向的培养岗位，结合当前大数据产业和企业的发展情况，本文选取了 Java、Android、大数据、Python 这四个类型的岗位爬取了前程无忧招聘网上招聘企业信息。在前程无忧招聘网上分别对这四个岗位搜索，查看了一下 url、页码和需要爬取的数据，求出 xpath，使用 scrapy 框架进行爬取。

1. 引入 scrapy 框架 import scrapy
2. 选择爬取对象 allowed_domains = [´51job.com´]
3. 针对不同的职位 jobname = ´安卓´
4. 设置要爬取的内容：公司，地点，薪水，要求等

```
company = scrapy.Field()
work_place = scrapy.Field()
salary = scrapy.Field()
joblink = scrapy.Field()
```

将数据保存在 job.csv 文件中，接着跟进其中的 joblink 链接，爬取对应岗位的详细信息保存在 basejob.csv。然后将两个文件(job.csv 和 basejob.csv)合并，得到最终文件 zhaoping.csv。

3.2 数据清洗

数据清洗(Data cleaning)是对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。使用 Python 程序在招聘网上爬取的原始数据会存在许多问题，需要将这些数据进行清洗，然后再进行分析并可视化。

首先使用 padas 框架读取数据并清洗薪资在不同企业招聘过程中的表示方式各有不同，例如薪资，如：1–1.5 万/月，要将数据格式统一为 10K 这种格式，进行格式化操作为例进行数据的清洗展示。

1. 引入 padas 框架 import pandas
2. 读取数据 pd.read_csv(´data/job.csv´)
3. 格式化薪水格式
4. df[´salary´]=df[´salary´].apply(get_salary)

3.3 数据分析与可视化

数据清洗完成后，可以针对数据进行分析，本文以不同岗位的薪资水平对比为例实现数据分析和可视化。

根据不同岗位的薪资水平进行对比

1) 设置不同岗位
2) lang = [´python´,´java´,u´大数据´,u´安卓´,´android´]
3) 获取对应的平均薪水
4) avg_salary = avg_salary[:-2]+[sum(avg_salary[-2:])/len(avg_salary[-2:])]
5) 设置显示的图标格式
6) p = plt.bar(lang,avg_salary)
7) plt.title(u´python、java、大数据和安卓职业薪资待遇对比´)
8) 采用柱状图显示
9) def autolabel(rects):

最后运行之后得到的柱状图如图 1 所示。

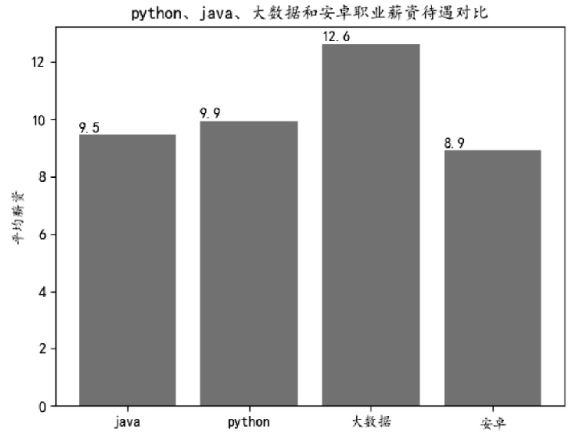

图 1　不同岗位平均薪资水平对比图

在软件企业人才需求、岗位分布及要求的分析过程还进行了如图 2，图 3 所示的大数据和 Java 在不同地区薪资的热力图。图 4，图 5，图 6，图 7 所示的大数据和 Java 的岗位分析词云图。

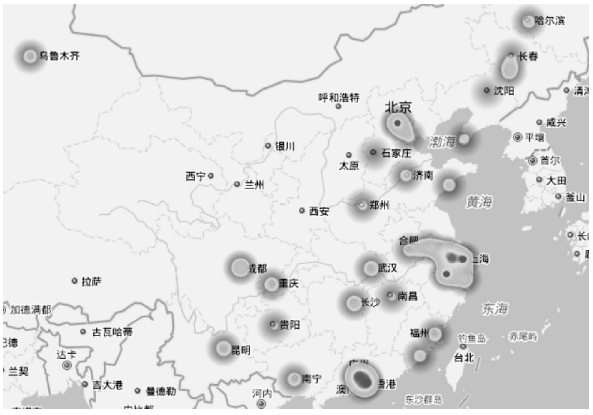

图 2　大数据岗位不同区域薪资热力图

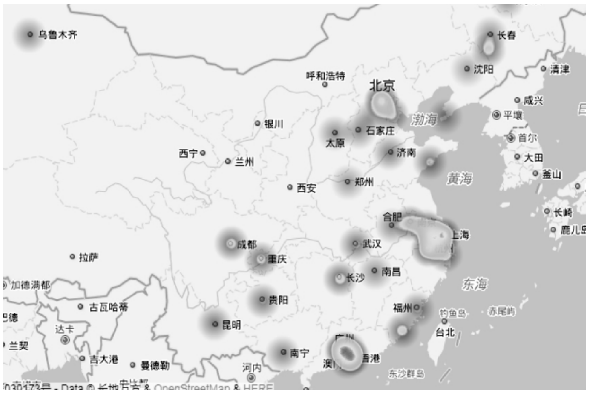

图 3　Java 岗位不同区域薪资热力图

根据图 1 到图 3 所示大数据无论是工作地点还是薪资均高于其他三种岗位；Python 薪资也不低；Java 系列的岗位软件行业的招聘主力，其工作地点和薪资也仅次于大数据行业；由此观之，大数据的发展空间是最大的，目前前途也是最好的，但 Java 仍然是很不错的选择。

图 4 和图 5 展示的内容可以体现出目前招聘的企业对 Java 和大数据两个不同岗位的人才需求的具体方向，另外如图 4 所示不仅出现大数据的开发分析岗位，还出现了市场营销、房产中介、置业顾问等岗位，这个从另外一个方面反应了目前不仅软件开发行业对大数据人才有需求，其他行业对大数据人才也有需求。

图 4　大数据岗位分析词云图

图 5　Java 岗位分析词云图

四、结束语

过去对行业和企业的人才需求的获取一般采用座谈、实地考察、问卷等方法调查，获得的信息比较有限。本文利用 Python 语言编程爬虫程序和分词工具获取了招聘网站上招聘信息的地域、行业和工作职责的文本特征，并对数据进行清洗，分析并根据不同的要求进行可视化，基于这些特征研究了四种不同类型岗位在企业中对人才需求以及招聘企业在国内的区域分布情况。这些数据对于高校的软件人才培养有一定的参考价值。

参考文献：

[1] 工业和信息化部. 工业和信息化部关于印发大数据产业发展规划（2016-2020 年）的通知 [EB/OL]. (2017-01-17) (2018-08-30).http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5464999/content.html.

[2]中国产业信息网. 2018 年中国软件行业总体发展情况分析[EB/OL]. (2018-06-05) (2018-09-01). http://www.chyxx.com/industry/201806/646630.html.

[3]Hacker News Hiring Trends. Top 10 Programming Languages. [EB/OL]. (2018-07-10) (2018-08-31). https://www.hntrends.com/2018/jul-top-ten-programming-languages.html.

作者简介：

王涛(1984-)，男，汉，江西永修人，讲师，硕士研究生，主要研究领域为职业教育、大数据分析和数据挖掘。