基于Python的招聘数据分析与可视化

闫美娟
山西省财政税务专科学校

摘　要：大数据时代，有效提取互联网大数据价值已成为当前数据挖掘领域的研究热点。基于Python的招聘数据分析与可视化，以大型招聘网站数据为分析对象，使用正则表达式实现了“Java开发工程师”这一岗位的招聘数据的爬取，通过xlwt模块实现数据存储，然后进行数据预处理，通过Matplotlib绘制图形对诸如薪资水平、工作经验要求、地区及所需Java开发工程师数量情况等数据进行可视化分析。分析结果表明提出的大数据爬取算法和数据可视化方法可以真实、完整、有效地反映对应信息，为Java开发工程师求职意向者提供了丰富的参考信息。

关键词：Python；数据分析；可视化
【DOI】10.12252/j.issn.2096-6261.2022.10.131

一、引言

当今世界是一个互联网迅速发展的时代，网络渗透在我们日常生活的方方面面，它为我们提供了极大的便利，比如求职者可以在各大网络招聘平台中了解到丰富的招聘信息，但是随着社会的智能化程度越来越高，网络平台的招聘信息不断增加，求职者对求职的岗位技能要求、学历要求及薪资水平等关键信息掌握不够全面，导致在求职过程中出现信息偏差，怎样高效利用这些招聘信息，从中获取求对求职者有价值的招聘信息，已然成为迫切的需求[1]。本文使用正则表达式爬取主流招聘网站的招聘数据，然后通过Python语言的第三方库pandas对数据进行有效清洗和分析，最后通过matplotlib库对统计分析后的招聘数据进行可视化展示，最终根据不同学历、工作经验、地区等条件对Java开发工程师薪资的影响进行了详细的分析，为Java开发工程师求职意向者提供了丰富的参考信息。

二、相关技术介绍

（一）网络爬虫

网络爬虫是一种通过设置可以自动抓取网络数据的程序，它主要通过链接关系完成分析和寻找网页，进而完成网页内容的读取[2]。

针对静态网页，在爬虫使用时需要解决三个主要问题：

（1）对抓取目标的详细描述或定义；

（2）对数据的详细分析与过滤；

（3）关于URL搜索策略的设定[3]。

（二）数据预处理

实际问题中我们收集的数据大多是不一致、不完整的“脏”数据，比如网络爬虫得到的数据，这些“脏”数据需要进行专门的处理，才能进行数据分析，否则将会导致错误的分析结论或者不能完成分析任务[4]。而且有价值的数据分析结果依赖于高质量的数据，所以对收集到的数据，我们要进行针对性的处理，这些特有的处理称为数据预处理。数据预处理包括：清洗不满足要求的数据以保证数据质量、对数据进行规范化处理以使数据形式满足分析要求。数据预处理可通过数据清理、数据集成和变换、数据归约、数据离散化等[5]多种方法完成。

（三）数据分析

数据分析重要的一个应用方面就是，针对事先确定的分析目标或者要解决的问题，选择恰当的分析方法和工具，对收集来的数据进行分析，提取隐藏的规律，形成结论报告，进而给使用者提供决策依据[6]。

统计分析方法中的描述统计是常用的数据分析方法，它采用整理、归类、和绘制图表的方法，来描述数据的特征，进而归纳、总结数据之间的关系。

Python是一种高级动态编程语言，是一个强大的数据分析工具，依托于丰富的扩展库，可以轻松完成高级任务，其中，pandas用于数据的基础处理，matplotlib提供数据可视化的丰富函数，Spyder用于网络数据的爬取、Scikit-learn用于进行数据分析等，这些扩展库给数据分析带来了极大的便利。

（四）数据可视化

数据可视化技术是为了给决策者提供一目了然的分析结果，而将数据分析结果以直观形象的形式表现出来，并能够进行交互处理的技术。为了更加直观地展示对“Java工程师”这一岗位数据分析的结论，本文采用在可视化方面有重要意义的Python第三方库matplotlib提供的数据分析工具，根据不同的数据规律用不同的图表进行了展示，可用的图表包括：常规的散点图、折线图，用于统计的直方图、饼图，还有漏斗图、箱图等。

三、数据采集及预处理

（一）爬虫模块设计

1.网页爬取

本次爬虫设计的目标是获取“Java开发工程师”这一岗位的相关信息，主要信息包括职位、工作城市、薪资、学历要求、工作经验、公司规模等。

2.内容的提取

本次提取数据采用的是正则表达式，通过正则表达式筛选所需要的信息。并且设定对其中的100页进行了爬取，一共爬取了近一万条数据。

（二）存储模块设计

Python中常用的存储方法有很多，本文使用xlwt模块将爬取到的数据存储到excel表格中。存储结果如下表所示（这里截取了部分数据）：

表1 Java开发工程师岗位信息

职位	公司名称	公司地点	薪资	学历要求	工作经验	公司规模
Java开发工程师	浙江省建筑科学设计研究院有限公司	杭州	0.8-1.3万/月	本科	2年经验	1000-5000人
Java开发实习生	成都九威全然科技有限公司	成都	4.5-6千/月	大专	无需经验	50-150人
Java开发实习生	成都荆或科技有限公司	成都	2-3千/月	大专	在校生/应届生	50-150人
Java开发工程师	广州极云信息科技有限公司	广州	4.5-6千/月	招2人	1年经验	少于50人
Java开发工程师	南京途酷信息科技有限公司	合肥	0.9-1.5万/月	本科	2年经验	少于50人
Java开发工程师	上海德慧信息技术有限公司	上海	0.6-1.2万/月	若干人	1年经验	150-500人
Java初级开发工程师	成都珂芙故荟科技有限公司	成都	3-4千/月	大专	在校生/应届生	50-150人
Java开发工程师	西安普锐软件科技有限公司	西安	7-10万/年	本科	1年经验	少于50人
中级JAVA开发工程师	厦门中软海晟信息技术有限公司	异地招聘	1-1.5万/月	大专	2年经验	500-1000人
java高级软件工程师	江苏银丰信息技术有限公司	西安	1.5-2万/月	本科	8-9年经验	50-150人
Java资深开发工程师	端云信息技术（上海）有限公司	上海	2-2.5万/月	本科	5-7年经验	50-150人
java初级程序员+双休	武汉维动汇星科技有限公司	武汉	5-8千/月	招5人	大专	50-150人

（三）数据预处理

本文对爬取到的数据进行了数据清洗、数据规范化等处理，具体内容包括：

1.去除无用数据

在爬取到的数据中，公司性质、公司类型这两个字段与本文的分析目标无关，所以将这两列删除。

2.去除噪声数据

在学历要求、工作经验两个字段中，均存在噪声数据，因这两个字段均为分类属性，所以我们采用众数来填充。

3.去重

使用Pandas库的drop_duplicates函数来删除重复记录。

4.数据规范化

薪资字段单位有很多种格式，比如：千/月、万/年、元/天，为了进行统一分析，使用正则表达式筛选出详细的薪资信息，然后通过替换，将薪资统一成相同格式。

四、数据分析及可视化

本文采用pandas模块对对预处理后的数据进行统计分析，然后采用matplotlib库对数据进行可视化分析，主要内容如下：

（一）学历分析

不同学历不仅会影响求职者所选的工作单位，还会影响求职者入职后的薪资，首先对Java开发工程师在全国范围的学历要求进行了统计分析，在所有求职信息中要求为本科学历的占67.79%，硕士学历占0.42%，大专及以下学历占29.01%，硕士以上学历占2.78%，本科及以上学历对“Java工程师”这一职位明显有很大优势。然后对不同学历的薪资情况进行了具体分析和可视化展示，设置x轴为学历，y轴为薪资，再通过plot（kind='bar'）绘制折线图，结果表明硕士学历在“Java工程师”这一职位中薪资有明显优势，结果如下图所示：

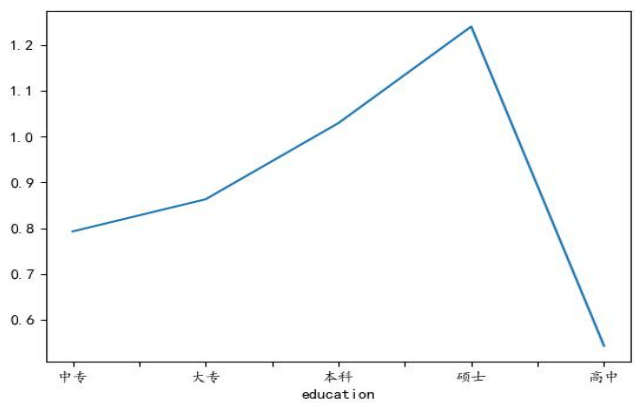

图1 学历与薪资的关系图

（二）工作经验分析

工作经验不仅对是否能够求职成功有影响，更与入

职后的薪资息息相关，所以本文进行了不同工作经验人数和不同工作经验段之间薪资差异的分析。首先采用漏斗图对Java开发工程师的不同工作经验人数进行了可视化展示，结果显示具有3-4年工作经验的人数最多，依次为2年工作经验，5-7年工作经验等，具有10年以上工作经验的人数最少，表明“Java开发工程师”这一职位比较年轻化。

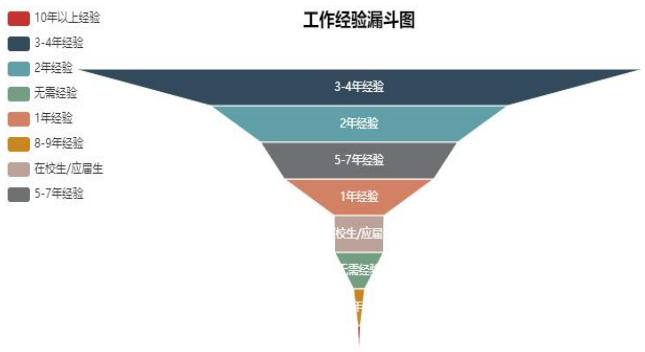

图2 工作经验漏斗图

本文还对不同工作经验段之间薪资差异进行了统计分析，在校生/应届毕业生和无工作经验人员薪资水平较低，都在6000元以下，随着工作经验的增长，薪资从8000逐渐上涨，当有8-9年工作经验时，可以涨到20000元甚至更多，表明“Java开发工程师”这一职位的薪资与工作经验息息相关。

（三）区域分析

不同地区所需要的人才数量也不同，对Java开发工程师需求量进行了统计分析，用分布图展示全国各个地区对Java工程师的需求分布状况，结果表明需求主要集中在长三角、珠三角、山东半岛、华中地区等。

然后进一步对不同城市所需的Java开发工程师数量进行分析，通过plt（bar）绘制如下柱形图，设置x轴为城市，y轴为Java开发工程师数量，结果表明需求量排名前三的城市依次是上海、深圳、广州。

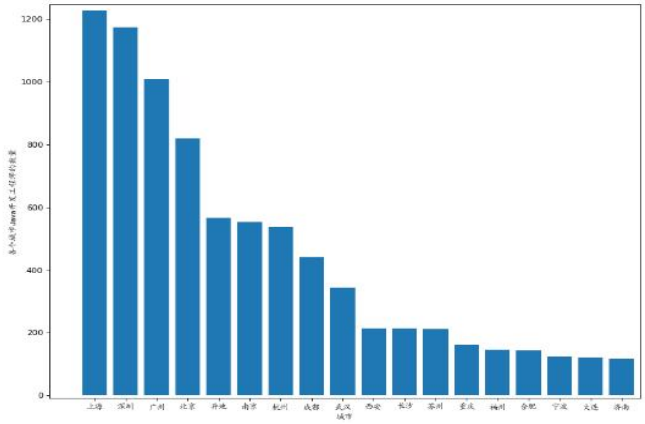

图4 不同城市Java开发工程师需求数量统计图

五、结论

学历、工作经验、城市都对求职者的薪资有所影响，通过城市的可视化分析，可以得出上海对Java开发工程师的需求量相对来说比较大，岗位比较多；大连和济南需要的Java开发工程师的需求比较少。同时通过Java开发工程师需求地域分布图可以看到经济发展快速的沿海地区对Java开发工程师需求比较大。

通过学历的可视化分析，可以得出学历越高，薪资越高。最低薪资是每月六千以下，最高薪资则达到每月1.2万左右。而且学历要求为本科居多，因此对于本科毕业生来说，就业机会较大，但由于本科毕业生基数较大，所以竞争也相对较大。

通过对工作经验的分析，可以得出在一定范围内，工作经验越长，薪资越高，通过工作经验漏斗图分析，Java工程师这一岗位的工作经验在3-4年居多，工作经验在10年及以上人数最少，同时工作经验对薪酬影响较大，且薪酬随工作经验的递增有着明显的递增趋势，尤其是具备3-4年的工作经验后，薪酬增长速度将变的更快。

本文对Java开发工程师的招聘信息进行了全方位、具体的分析，给不同需求的求职者提供了丰富的参考信息，以帮助求职者更好地了解就业的关键信息，求职者可以根据不同薪资水平类别的分布情况选择更符合自身期望的岗位。

参考文献

[1]丁文浩，朱齐亮.基于Python的招聘数据爬取与分析[J].网络安全技术与应用，2022（01）：43-45.

[2]杨应浩.基于Python的电影信息爬取与数据可视化分析[J].新型工业化，2021，11（07）：71+73.

[3]依力·吐尔孙，艾孜尔古丽.基于Python的美食数据爬取及可视化研究[J].电脑知识与技术，2021，17（10）：19-20+29.

[4]简悦，汪心瀛，杨明昕.基于Python的豆瓣网站数据爬取与分析[J].电脑知识与技术，2020，16（32）：51-53.

[5]刘晓知.基于Python的招聘网站信息爬取与数据分析[J].电子测试，2020（12）：75-76+110.

[6]葛琳，杨娜.Python招聘数据分析[J].计算机与网络，2020，46（16）：62-65.