**The Data Crawling and Analysis of Lago.com**

**拉勾网数据爬取与分析**

2013302580150 国际2班 王孜晗

**摘要**

本文使用Python编写爬虫，通过向端口传送请求并且抓取传输过来的json字符串来获取招聘职位信息，并且分类保存为csv格式的表格文件。最后通过长时间的爬取，最终得到37.7MB的表格数据，共计314093个招聘信息。之后通过SPSS对数据进行预处理和统计，再进行深度数据分析。

**【关键词】：** 拉勾网 招聘信息 爬虫 数据挖掘 数据分析 Python SPSS

1. **前言引入**

在当下就职情况依旧紧张的情况下，高考之后的专业选择以及专业既定之后的方向选择就成为让大多数心有迷惘的问题。在经济压力之下，未来人生的职业选择早已不仅仅与兴趣理想相关，也与实实在在的金钱报酬挂钩：如果因为想要挑战自己，大学四年苦攻最艰涩的编程语言，最后等到毕业却毫无用武之地，或者工资比切切图的前端程序员工资还低，那该多遗憾呀。如果我们可以在专业学习的初始阶段，就能够了解好当下的职位招聘信息，知道哪些方向市场需求量大，哪些方向平均工资比较高，而这些自己感兴趣想要发展的方向，又有哪些需要具备的技能和工具，那大学四年的学习不就有目的多了吗。同时对于自己感兴趣的职位在各个城市的分布，也便于我们提前确定以后工作的落脚点。

同时，当前金融和IT技术都是专业选择的热门，尤其是经济与金融，成为高分考生的首选专业。那在对两个专业类型都抱有兴趣的情况下，了解一些金融与技术工作的工资、岗位需求状况有助于我们做出合适的选择。

而现在我们处于软件工程专业中，对于工科专业，能够有一门一技之长是十分重要的事情。而仍然有相当一部分同学没有认清当前的行业职位状况，在编程学习方面得过且过，想要毕业之后做一名运营人员或者产品经理，而这些职位的市场需求和工资状况，真的能够和编程工作相媲美吗？我们需要一些有说服力的数据来给自己以惊醒：如果大学阶段不好好学习和磨砺自己的技术和相关能力，毕业之后可能要在消费水平高得惊人的大城市拿着微薄的薪水过活，后悔自己不曾努力。

现在的“互联网+”理念盛行，创业公司如雨后春笋，互联网相关行业毕业生传统的想要直接进入BAT公司的观念似乎不再维持主流地位，更多的新兴公司开出了更优厚的待遇和报酬。而对于已经工作了一段时间的人来说，他们也需要了解当公司处于各个阶段时给予的薪资情况，来对于跳槽有一个自我估量。

无论如何，以上这些都需要对于当前市场中的职位招聘信息有一个分析和挖掘，而本文正是完成这一步骤，向阅读者提供了一些数据分析结果。同时提供了爬取下来的全部数据信息，便于读者对于自己感兴趣的方面自行进行探索。

接下来，本文将按照如下的结构进行展开：在第二节和第三节里分别详细介绍数据爬取和利用SPSS的整体分析数据过程，第四节展示实验结果，第五节为实验结论。

1. **数据爬取**

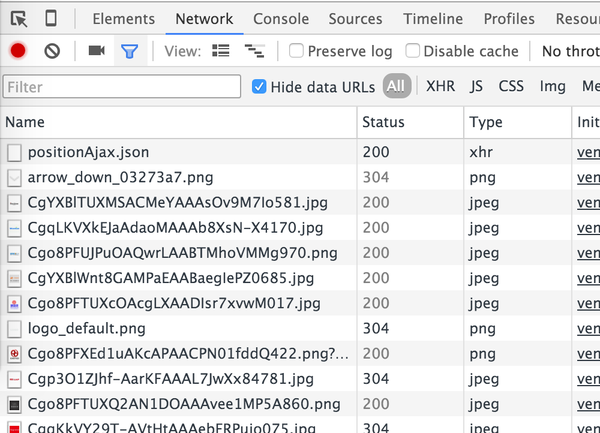
在本文中使用的是Python语言来编写爬虫程序，使用Mac OS X自带的Python 2.7环境，集成编程工具IDE为PyCharm 5。

正常情况下，爬虫需要通过BeautifulSoup解析网页并且从相应的html元素中获取想要的信息。然而对于拉勾网的抓取，我们发现HTML中并没有我们需要的数据。

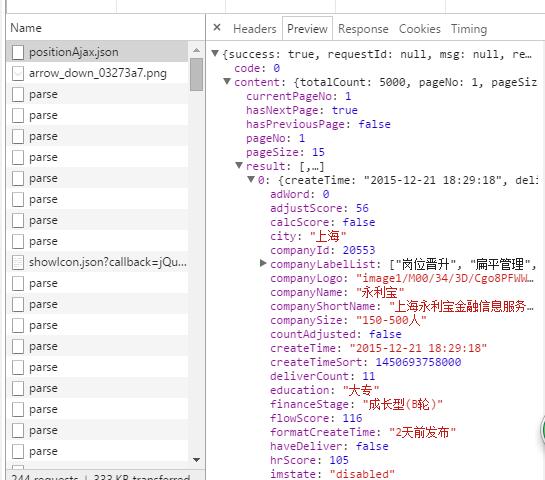
进入拉勾网点击相应的职位分类之后，会发现招聘信息的出现时间是晚于页面框架出现时间的，而且点击下一页之后页面的地址栏并没有跳转，可以知道页面是利用Ajax，把得到的数据通过js渲染之后展示到界面的。

好消息是，通过向服务器端口传递数据后得到的数据一般是json格式的，这样我们非但不需要对于页面中的各个元素进行定位，反而减轻了爬虫的任务，省去了解析HTML的功夫。

打开Chorme的开发者工具，当我们点击“下一页”之后，浏览器的Network中进行了如下活动：



注意到"positionAjax.json"这个请求，它的Type是"xhr"，全称叫做"XMLHttpRequest"，XMLHttpRequest对象可以在不向服务器提交整个页面的情况下，实现局部更新网页。那么，现在它的可能性最大了。在单击之后发现：



通过对内容的观察，返回的确实是一个json字符串，内容包括本页每一个招聘信息。

我们切换到Headers这一栏，留意三个地方：



上面的截图展示了这次请求的请求方式、请求地址等信息。



上面的截图展示了这次请求的请求头，其中需要关注的是Cookie / Host / Origin / Referer / User-Agent / X-Requested-With等参数。



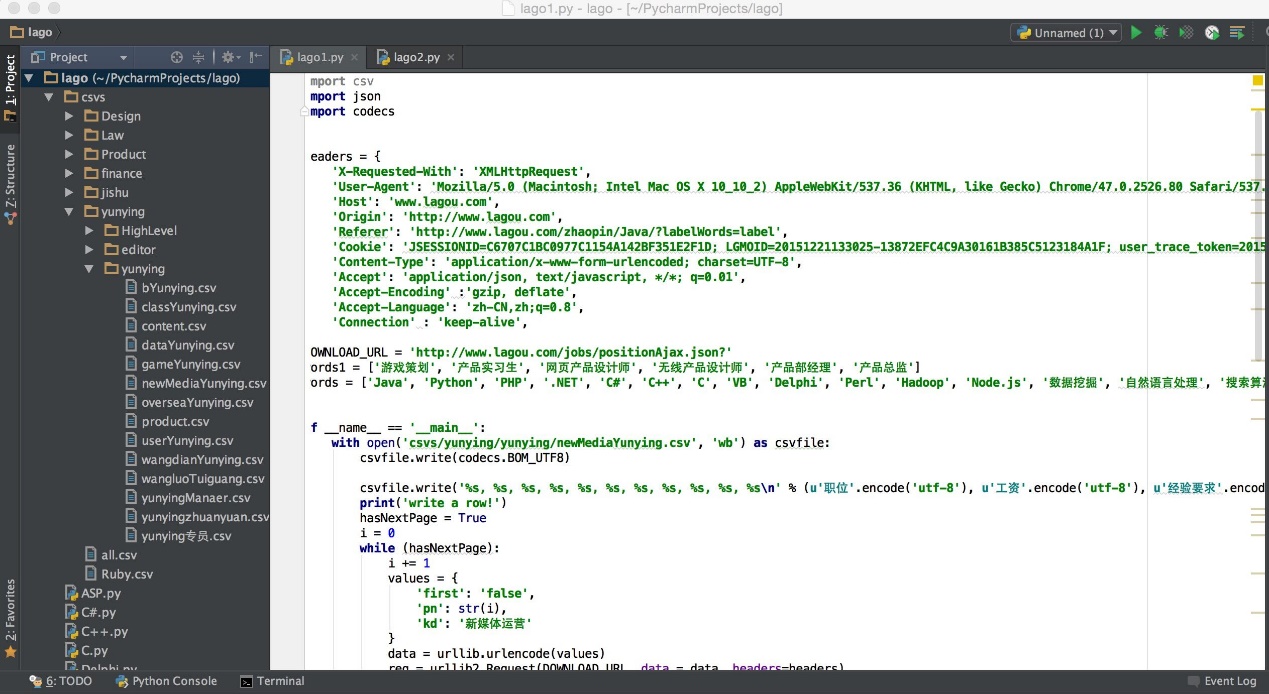
上面这张截图展示了这次请求的提交数据，根据观察，kd表示我们查询的关键字，pn表示当前页码。

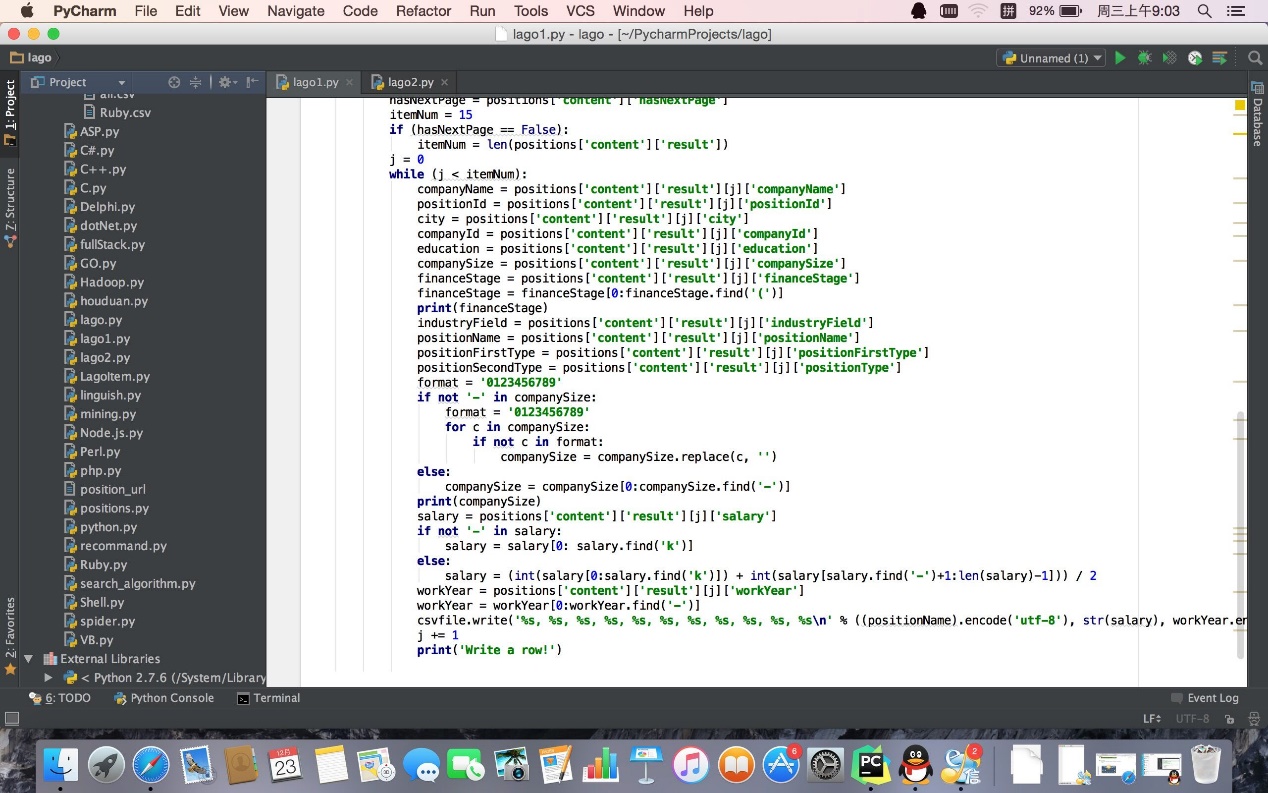
那么，我们的爬虫需要做的事情，就是按照页码不断向这个接口发送请求，并解析其中的json内容，将我们需要的值存储下来就好了。我们观察一下返回的json，格式化之后的层级关系如下：



这样通过对json字符串的操作就可以轻松地拿到我们想要的数据了。对于拿到的每一条数据，都需要通过file操作写入csv表格文件中。其中的编码又是一个难题，因为csv的默认编码格式是ASCII码，而中文编码是UTF-8。这里我们在文档头部设置全文件coding编码格式，并且在写入文件时，对于从json中读取到的UTF-8格式的数据执行encode再使用，最终解决了文件写入时的编码问题。

最终的爬虫代码如下：

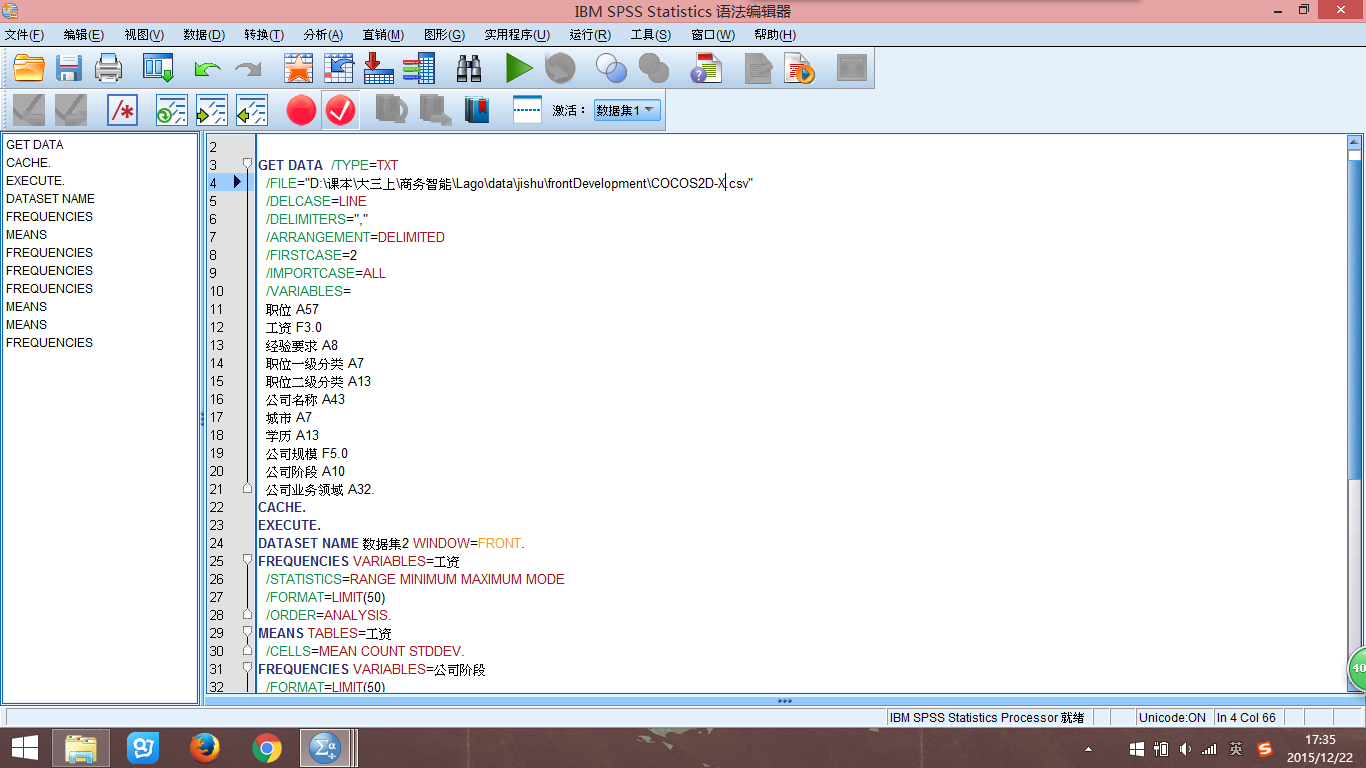




1. **SPSS整体数据分析**

在Windows环境下，使用SPSS 22.0软件对于获取到的全部csv文件逐个进行数据预处理，包括职位数目统计、职位平均工资与工资众数计算、职位学历要求与公司阶段情况统计、工作经验要求统计等等，并且将结果输入到单个excel表格文档中，便于之后更深入的数据统计与分析。

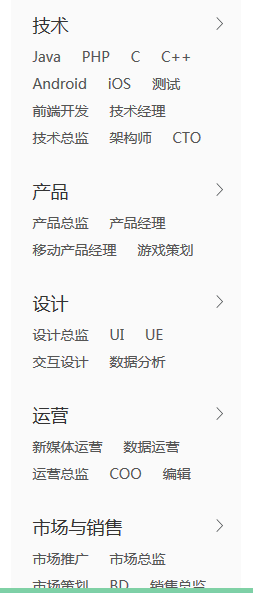
SPSS数据预处理代码如下：



之后对于数据预处理之后得到的excel表格，在SPSS中进行深入的分析，包括构建数据描述、根据比例描绘饼状图、以工作领域为时间序列构建时序图、构建茎叶图、双变量线性相关检验等等多种分析方式，数据分析得到的结果图将在第五节“实验结果”中进行展示。

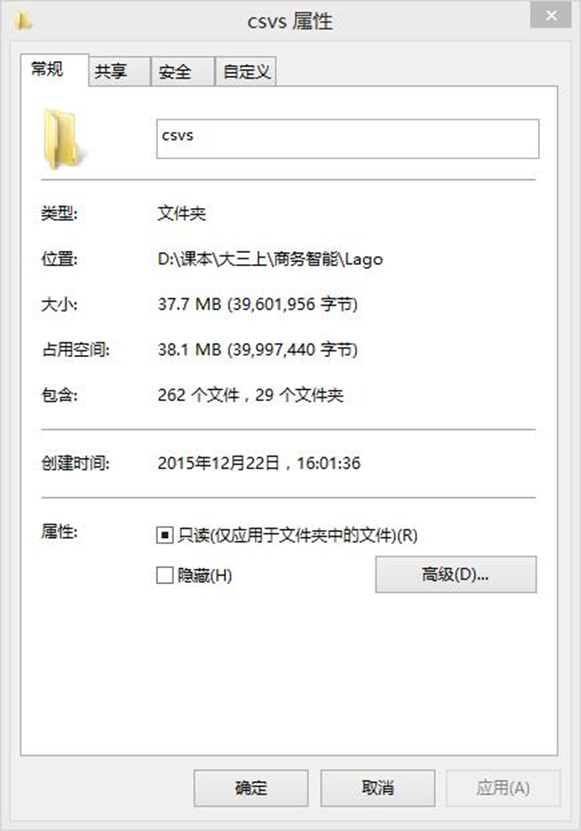
1. **实验结果**

拉勾网上的职位分类如下所示：

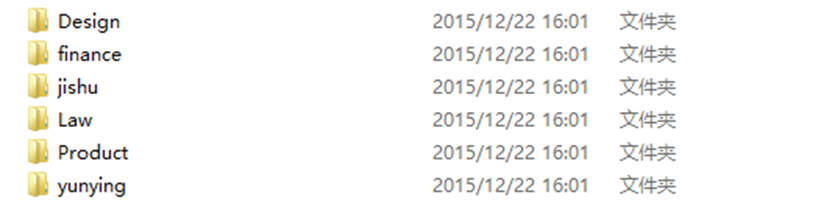


因为我们的主要研究为互联网及金融相关，所以对于本专业涉及不到及兴趣无关的行业领域的数据，我们并没有选择去爬取，即没有爬取“市场销售”及“职能”两个领域的职位数据，但是仍然爬取了“职能”数据下的“法务”数据。

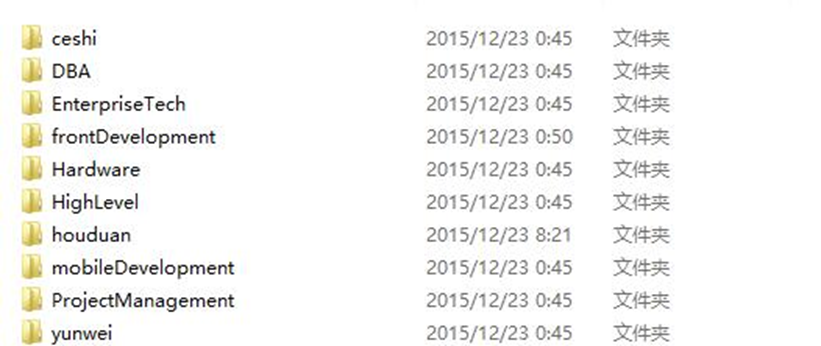
通过运行爬虫，最终我们爬取到了37.7MB的csv表格数据，如下面的截图所示：



对于爬取到的csv表格，我们按照领域分类和领域下的方向分类进行归类存放：



上面显示的是第一 领域分类，即大方向上的领域分类，分别为：设计、金融、技术、法务、产品和运营。



上面显示的是技术领域中的第二领域分类，即在各个领域下的职位方向分类。

在技术领域中，方向分别为：测试、DBA（数据库管理员）、企业软件、前端开发、硬件开发、高端职位、后端开发、移动开发、项目管理和运维。

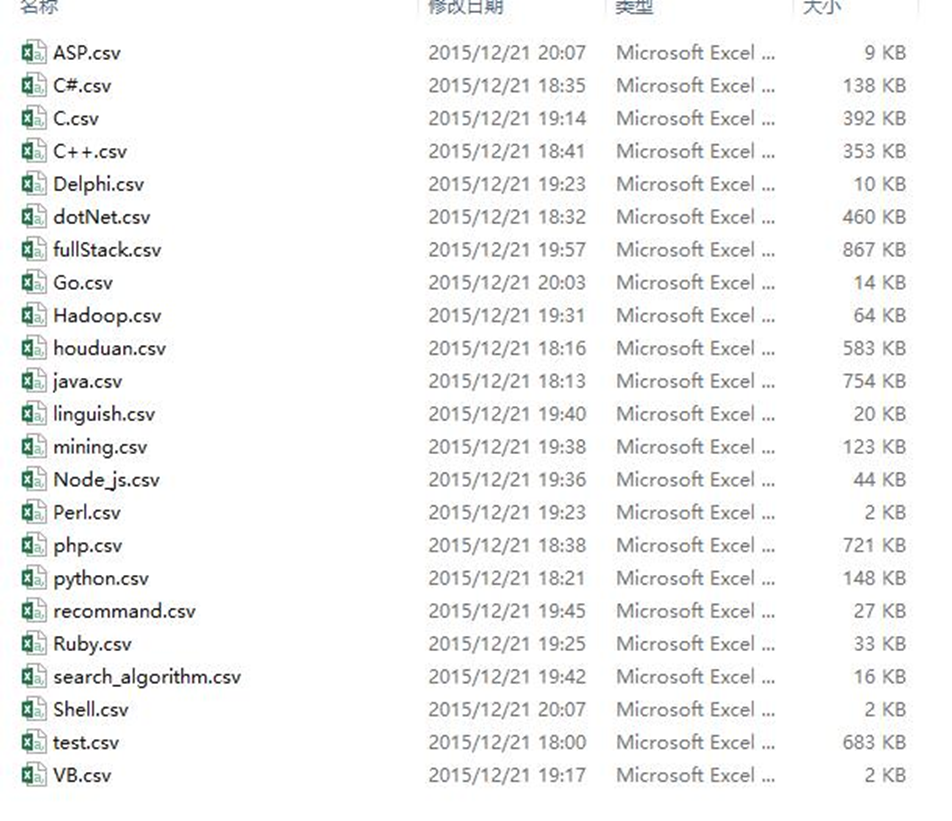
在设计领域中，方向分别为：高端职位、交互设计、用户研究和视觉设计。

在金融领域中，方向分别为：风控、高端职位、审计和投融资。

在法务领域中，只有一个方向，即法务。

在产品领域中，方向分别为：高端职位，产品设计和产品经理。

在运营领域中，防线分别为：运营、编辑和高端职位。



上面展示的是技术领域后端开发方向中的具体职位种类，包括ASP、C#、C、C++、Delphi、.NET、全栈工程师、Go、Hadoop、Java、Node.js、Perl、PHP、Python、Ruby、Shell、VB、自然语言处理、数据挖掘、精准推荐、搜索算法、测试等等。



上面展示的技术领域后端开发方向中Java开发职位招聘的原生数据。其中工资是对于单条招聘信息中给出的工资最低工资和最高工资的平均数，单位为“千元/月”；经验要求为单挑招聘信息中对于经验要求的最低要求，单位为“年”；公司规模单位为“人”。

这样，我们一共爬取到了314093个招聘信息，其中包含183种具体职位。去掉三个虽然有职位挂名但是没有招聘信息的职位：F5(技术-运维)、热传导（技术-硬件开发）、精益工程师（技术-硬件开发）后，得到的职位的总平均工资为13.1963k/月。

在各个领域中，金融平均工资最高，技术其次；运营平均工资最低，法务其次。

在各个职位方向中，高端职位平均工资最高，后端开发及DBA其次；工资最低，运营及审计税务其次。

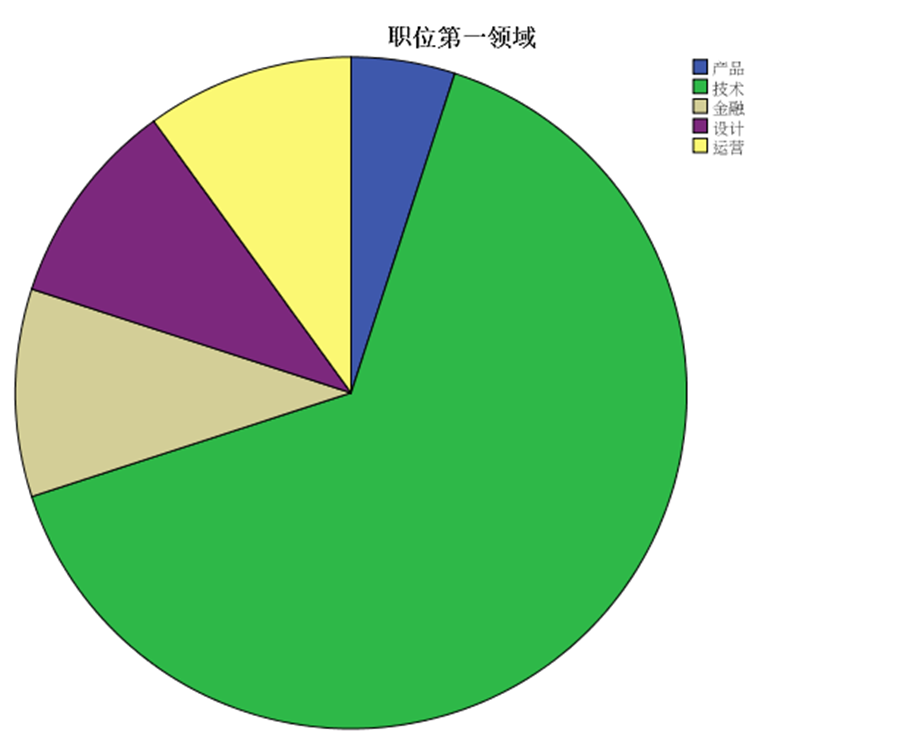
平均工资最低的职业为：运营-编辑-内容编辑；平均工资最高的职业为：金融-高端职位-副总裁。

工资最高的20种职位为：

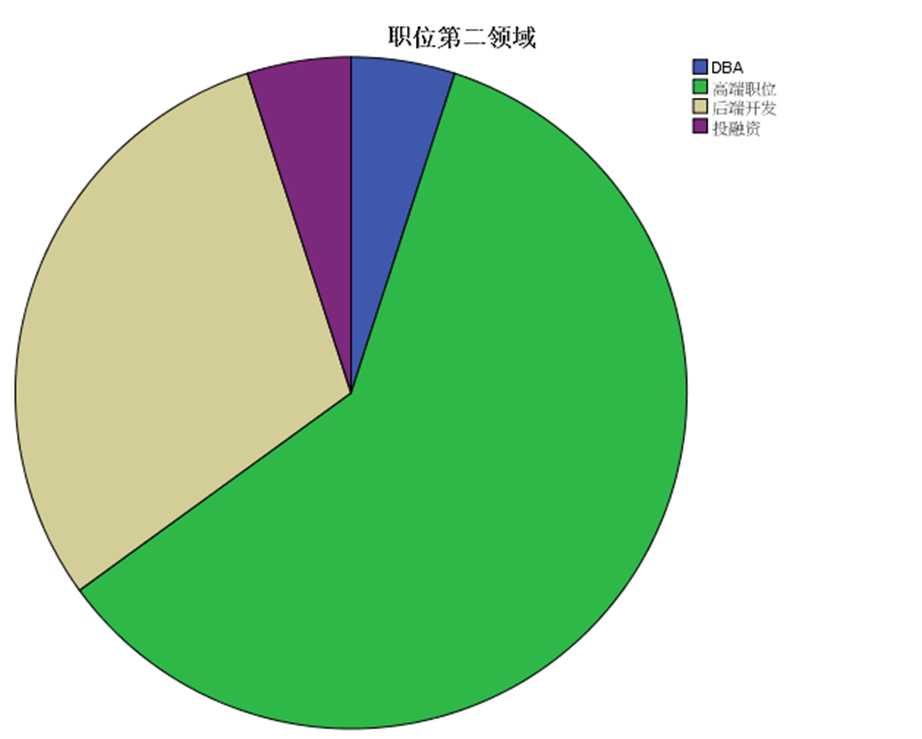


除了高端职位独占鳌头之外，算法及数据分析、数据库相关工作的工资也相当之高，包括搜索算法、精准推荐、自然语言处理、数据挖掘、MongoDB等等。初次之外，后端开发的报酬也比较可观。

在工资最高的前20名中，职位领域及方向分布如下面的饼状图所示：



可以看到，工资最高的TOP20职位里面，技术领域的职位最多。从事技术行业能够拥有更丰厚的薪资回报。

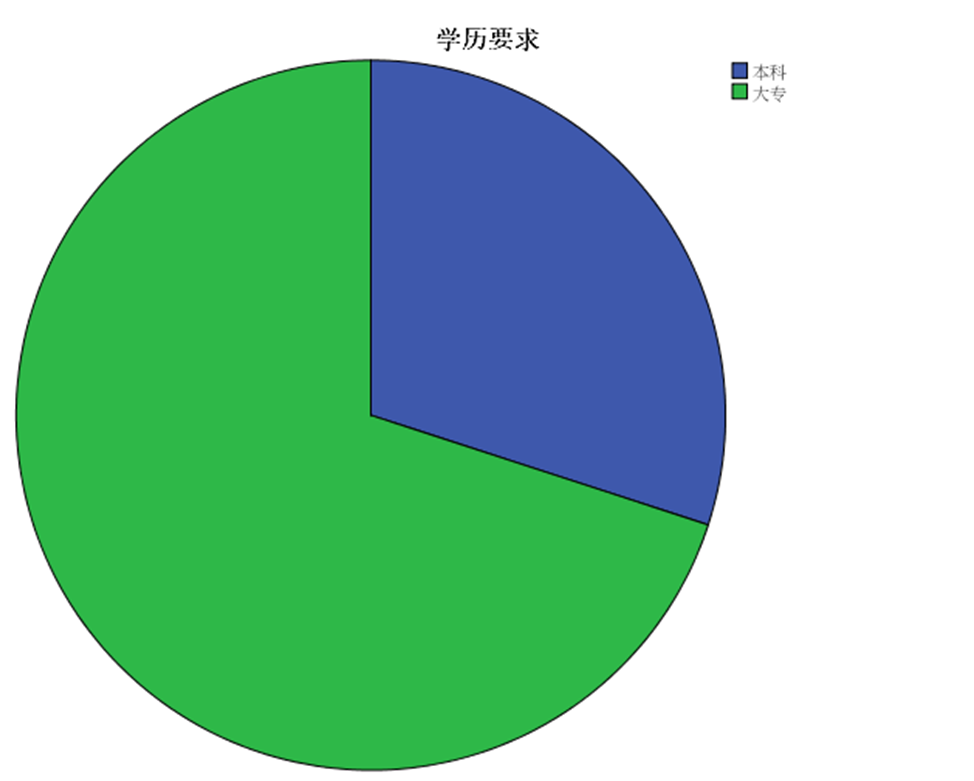


在工资最高的TOP20职位里面，按方向比例来划分，其中高端职位即管理岗位或者全面技术调控岗位占了大多数，其次便是后端开发、DBA和投融资。也就是说这四种职位方向往往会获得较高报酬。

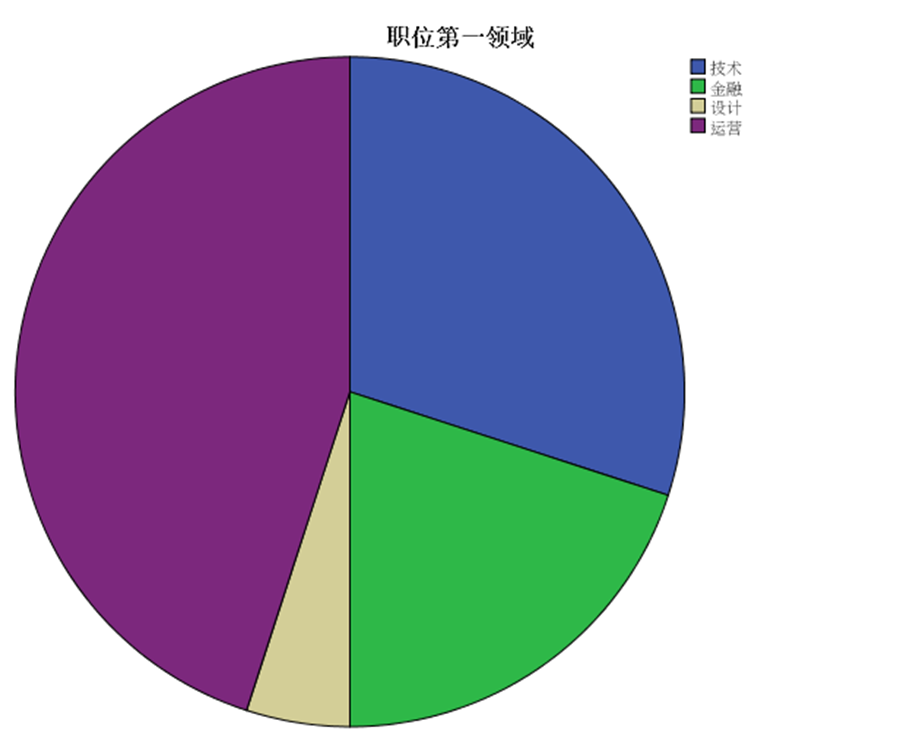
工资最低的20种职位为：



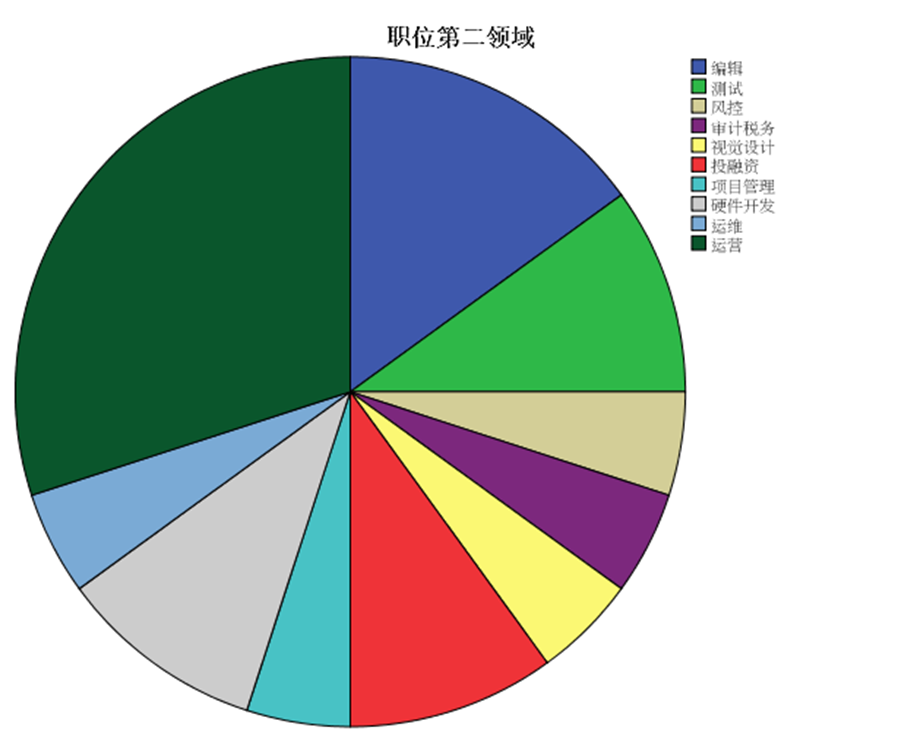
可以看到，刚才工资最高的20种职位中，学历要求都是本科，而工资最低的20种职位中，大专则占了相当的比例，即工资较低的职位对于学历的要求并不高。



在工资最高的前20名中，职位领域及方向分布如下面的饼状图所示：

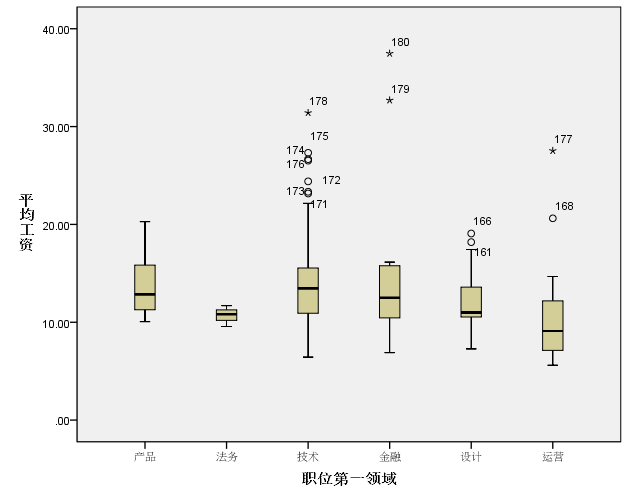


可以看到，工资最高的TOP20职位里面，按领域划分，运营和技术所占的比例最多，设计所占比重最小。



按方向划分，运营、编辑、测试、投融资、硬件开发所占的比例最大，入行需谨慎。

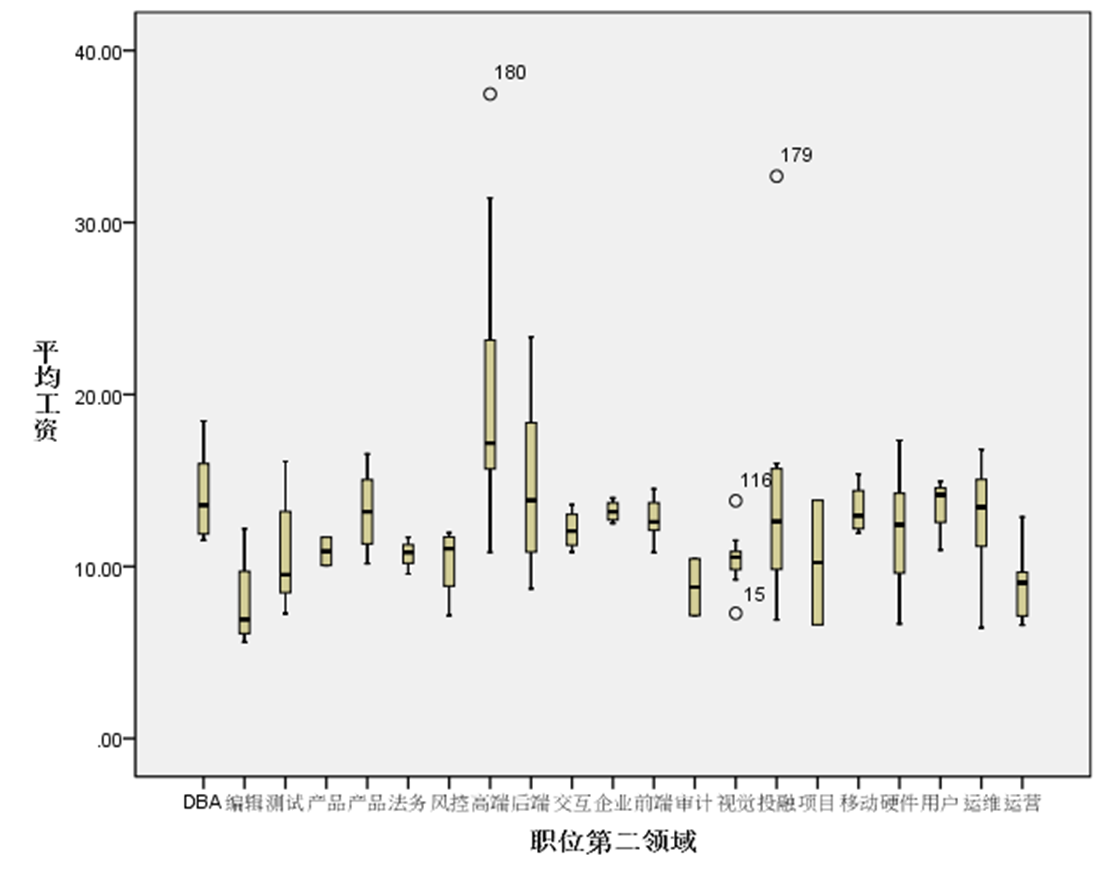
对于各个职位的平均工资，按照行业领域进行划分，统计构建茎叶图，可以得到如下的结果：



其中方框上边缘表示公司平均数，中间黑线表示中位数，而独立的点比如“175”则表示孤立的、数值特别的职位种类。每个数字表示一种职位种类。

可以看到，产品、技术和金融的平均工资都比较高，不相上下，而金融领域的平均要稍微高一点。但是工资中位数则是技术领域的工资最高。设计领域虽然平均工资似乎还不错，但是中位数却较低。报酬最惨淡的是运营领域。

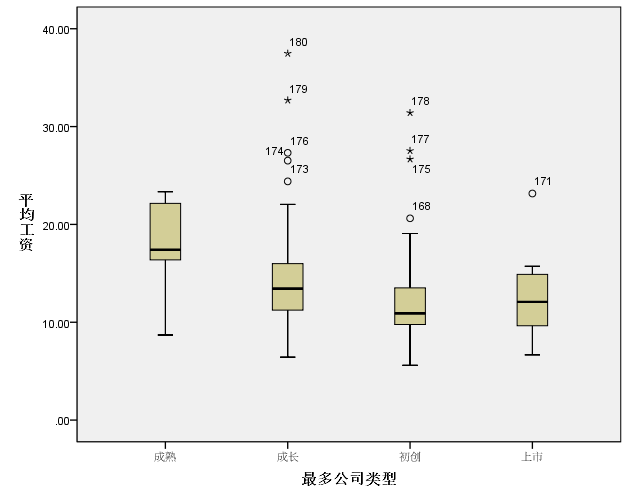
对于各个职位的平均工资，按照职位方向进行划分，统计构建茎叶图，可以得到如下的结果：



其中方框上边缘表示公司平均数，中间黑线表示中位数，而独立的点比如“175”则表示孤立的、数值特别的职位种类。每个数字表示一种职位种类。

可以看到，高端职位的工资遥遥领先，当然这是很符合常识的，毕竟高管确实是高薪职位。其次后端开发、投融资、DBA等方向的薪资也十分可观。而运营和编辑无疑是最惨的。

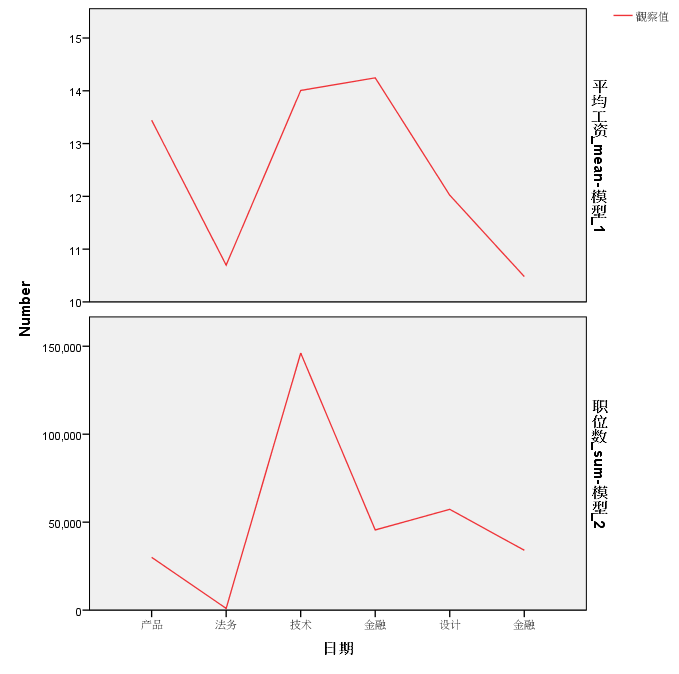
对于各个职位的平均工资，按照公司阶段进行划分，统计构建茎叶图，可以得到如下的结果：



其中方框上边缘表示公司平均数，中间黑线表示中位数，而独立的点比如“175”则表示孤立的、数值特别的职位种类。每个数字表示一种职位种类。

可以看到，与往常所理解的尝试不同，上市公司的平均工资反而还没有成长型公司给予的薪资丰厚。而成熟公司的工资的中位数甚至比其他公司阶段的平均数还要高，遥遥领先。初创型公司的条件就比较艰苦。

接下来，我们以行业领域为时间序列，构建时序图，探寻行业领域的平均工工资与职位数量的关系，得到如下的结果：



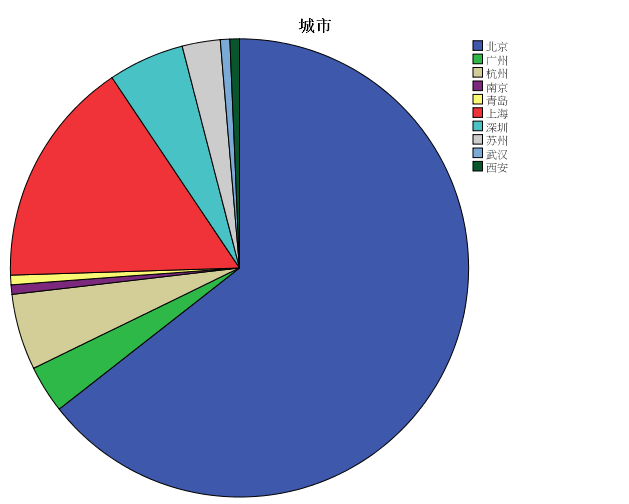
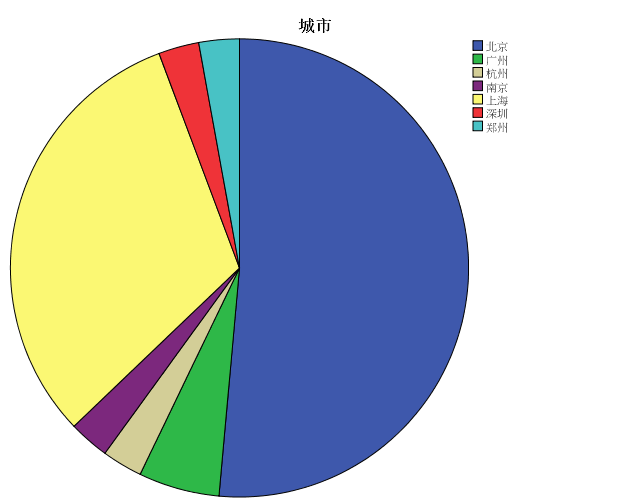
可以看到金融工作的工资最高，技术和产品其次。而技术工作的市场需求量是最大的。

作为软件工程专业的学生，我们特别关注的是技术领域的工资方向分布，经过统计计算可以得到如下结果：



高端职位暂时不在我们应届毕业生的考虑之列，那么我们能够锻炼的能力就主要集中在了后端开发和DBA上。测试和项目管理的工资最低。而现在当前火热的移动开发在薪资上表现得也不是那么给力，和企业软件、前端开发差别并不是很大。

对于我们比较关心的技术和金融领域，对比一下他们的城市分布，可以得到如下的结果：



可以看到相较而言，北京对技术人才的需求量更大，而上海对于金融人才的需求量更大。

1. **实验结论**

其实对于这个选题，我感觉我做得有点晚了，大三才对市场职业需求和薪酬进行了一次分析，而没有在大一的时候就认识市场从而调整学习方向。比如我先前比较感兴趣并且一直在学在写的前端、网站后台和安卓，其实在市场并不具备较高的报酬，同时工作也往往较为枯燥。另外，在我们心中认为的编程老大C++，无论在竞争力还是薪资报酬上都已经比不上了后起之秀Java。而在学生生涯里一直被嘲讽的PHP，在市场上却是风生水起。Python作为一款小而精的语言，在学校学习里面往往得不到足够的重视，然而市场的需求量和薪资报酬都非常可观。

同时，对于一些比较冷门的语言，比如VB、Perl等等，不仅市场量需求非常小，薪资也较低。以这些语言为钻研兴趣的同学，需要将毕业之后的生活压力考虑到心中。

而如果这份报告能出来得再早一点，就能够给大一大二的同学们一点警示：如果你在专业技术技能上经过四年学习浑浑噩噩一无所得，那么运营运维这样的工作薪资是少得可怜的，在大城市中生活将十分艰难。

另外技术行业的门槛特别低，很多招聘信息都愿意接受大专学历的简历，而金融法务等行业则没有这么幸运，本科毕业是必须。

我将奉上爬取下来的数据，供大家对自己感兴趣的方面进行更加深入详细的挖掘分析与探讨。