《大数据开发技术（Hadoop）》

实训报告

主题：招聘网站职位分析实战

组长：舒金权

成员：陈科颖 林浩祥

完成日期：2024/12/31

目 录

[第1部分 项目背景 1](#_Toc1057)

[第2部分 项目实现技术 1](#_Toc27978)

[第3部分 项目实现思路介绍 1](#_Toc20657)

[第4部分 项目实现流程介绍 1](#_Toc22979)

[第5部分 项目分析结果说明 1](#_Toc13545)

[第6部分 项目代码介绍文档 2](#_Toc1485)

# 项目背景

企业在日常招聘过程中，可能发现存在招聘周期过长、招聘成本过高、难以招到匹配度高的人才等问题。通过对招聘网站上大量目标职位的分析，能够梳理出不同职位的热门招聘渠道、求职者普遍期望的薪资范围、关键的岗位技能要求等关键信息。对招聘网站上职位的分析有助于企业了解整个行业内人才竞争的态势。从而据此调整自身的人才吸引策略，在行业人才竞争中保持优势，同时及时洞察行业内人才需求变化所反映出的技术、业务等方面的市场动态走向，提前做出应对举措。

# 项目实现技术

该项目使用pycharm、Hadoop 分布式集群环境的搭建和配置、BeautifulSoup、MySQl、Linux、PySpark、Python、Django、Echarts等技术实现。

# 项目实现思路介绍

1. 爬取招聘网站大数据职位信息，用BeautifulSoup清洗职位信息网页；
2. 使用PySpark对职位数据进行分析，通过matplotlib将分析所得数据可视化，从而更好的探索分析；
3. 用结巴分词对岗位描述进行分词，并统计关键词；
4. 创建Djingo项目，建立模型，将职位分析结果以字符云的形式可视化呈现。

# 项目实现流程介绍

1. 爬取招聘网站大数据职位信息（局域网）；
2. 使用 BeautifulSoup 清洗职位信息网页；
3. 利用 PySpark 对职位数据进行分析；
4. 对职位信息数据进行探索分析；
5. 结合分词对职位描述进行分词并将关键词统计；
6. 使用 Django 和 Echarts 将职位分析结果进行可视化；
7. 建立 LDA 模型对职位描述进行相似度计算。

# 项目分析结果说明

1. 公司规模在100-499人的公司招聘的大数据岗位最多。

公司性质为民营企业招聘的大数据岗位最多。

经验要求大部分没有明确说明，剩下的基本上集中在1-5年之间。

招聘公司主营行业主要集中在互联网、计算机、IT服务等行业。

职位类别主要侧重于数据库开发，软件工程师等岗位。

工作地点主要分布在北京，广东最少。

2. 公司招聘时，给的月工资主要集中在6000-8000元/月为35.8%，

其次是30000-50000，6000或20000-30000的也占很大比重，

500000元/月以上的所占比例最小为0.7%，几乎可以省略。

3. 薪资在1万元/月以下，经验要求由高到低的是1-3年，1-3年、无经验

薪资在1-2万元/月，经验要求由高到低的是3-5年，1-3年、5-10年

薪资在2-3万元/月，经验要求由高到低的是3-5年，5-10年、1-3年

薪资在3-5万元/月，经验要求由高到低的是5-10年，3-5年、10年以上

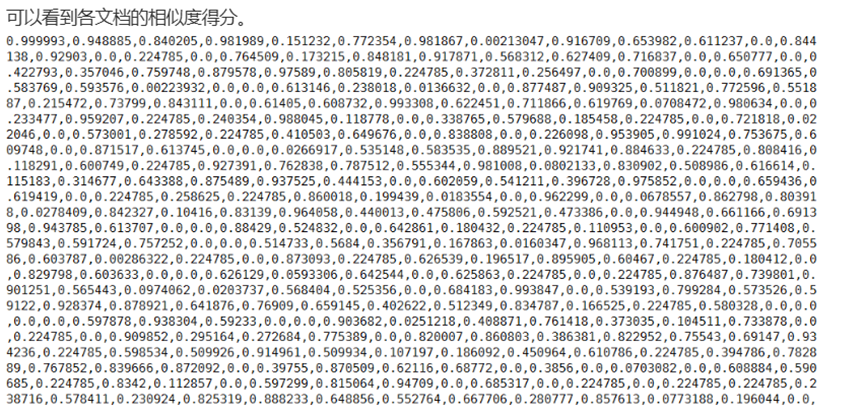
薪资在5万元/月以上，经验要求由高到低的是5-10年，3-5年、10年以上

4. 在各薪资范围内要求最低学历为本科的最多，中专、中技、高中几乎不存在。

5. 各城市公司招聘给定的薪资10000-20000元/月所占比例最大，不同薪资分布比例基本较相似。

6. 在关键词统计中hadoop,java,spark,linux的出现次数非常多。

1. 相似度统计

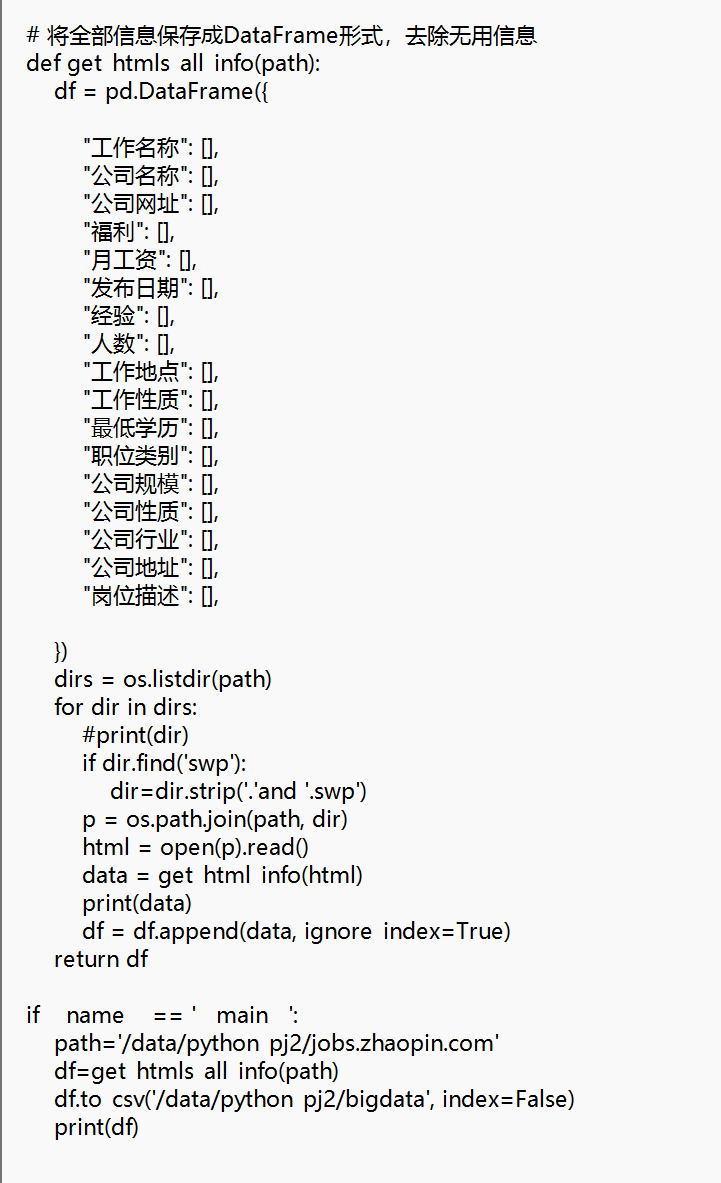


1. 职位聚类分析中大数据高级开发工程师的数据量远远大于其他聚类，所以可以得出大数据高级开发工程师需求量非常大。

# 项目代码介绍文档

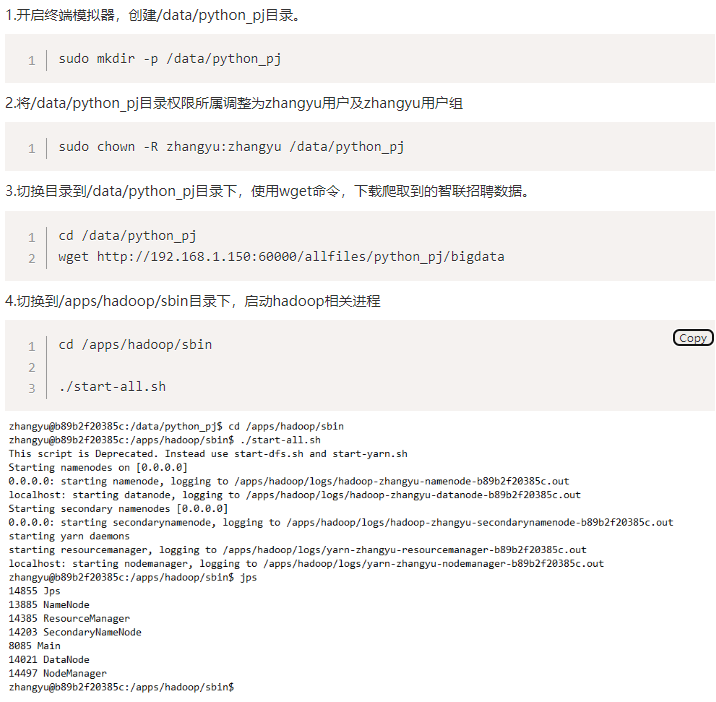
2.





获取网页的有用信息并保存成字典形式，将全部信息保存成矩阵形式，去除无用信息，在主函数中调用get\_htmls\_all\_info(path)函数，并把信息保存成.csv格式，从而实现数据清洗。

3.

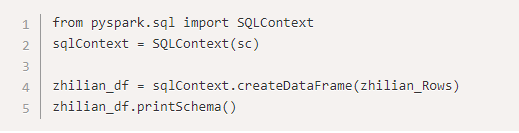




启动Spark，之后创建RDD并统计文件内容的行数。



创建DataFrame，导入row模块，通过RDD2创建DataFrame，定义DataFrame的每一个字段名与数据类型。



创建了zhilian\_Rows之后，使用sqlContext.createDataFrame()方法写入zhilian\_Rows数据，创建DataFrame，然后使用.printSchema()方法查看DataFrames的Schema。创建文件check，并将此步骤的运行结果保存至/data/目录下 check中。



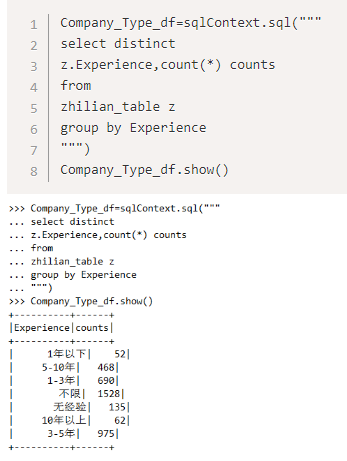
使用DataFrame统计公司性质及数量。



用registerTempTable方法将df转换为zhilian\_table表。



用sqlContext.sql()输入sql语句，使用select关键字查询文件内容行数，并使用from关键字指定要查询的表，最后使用show()方法显示查询结果。

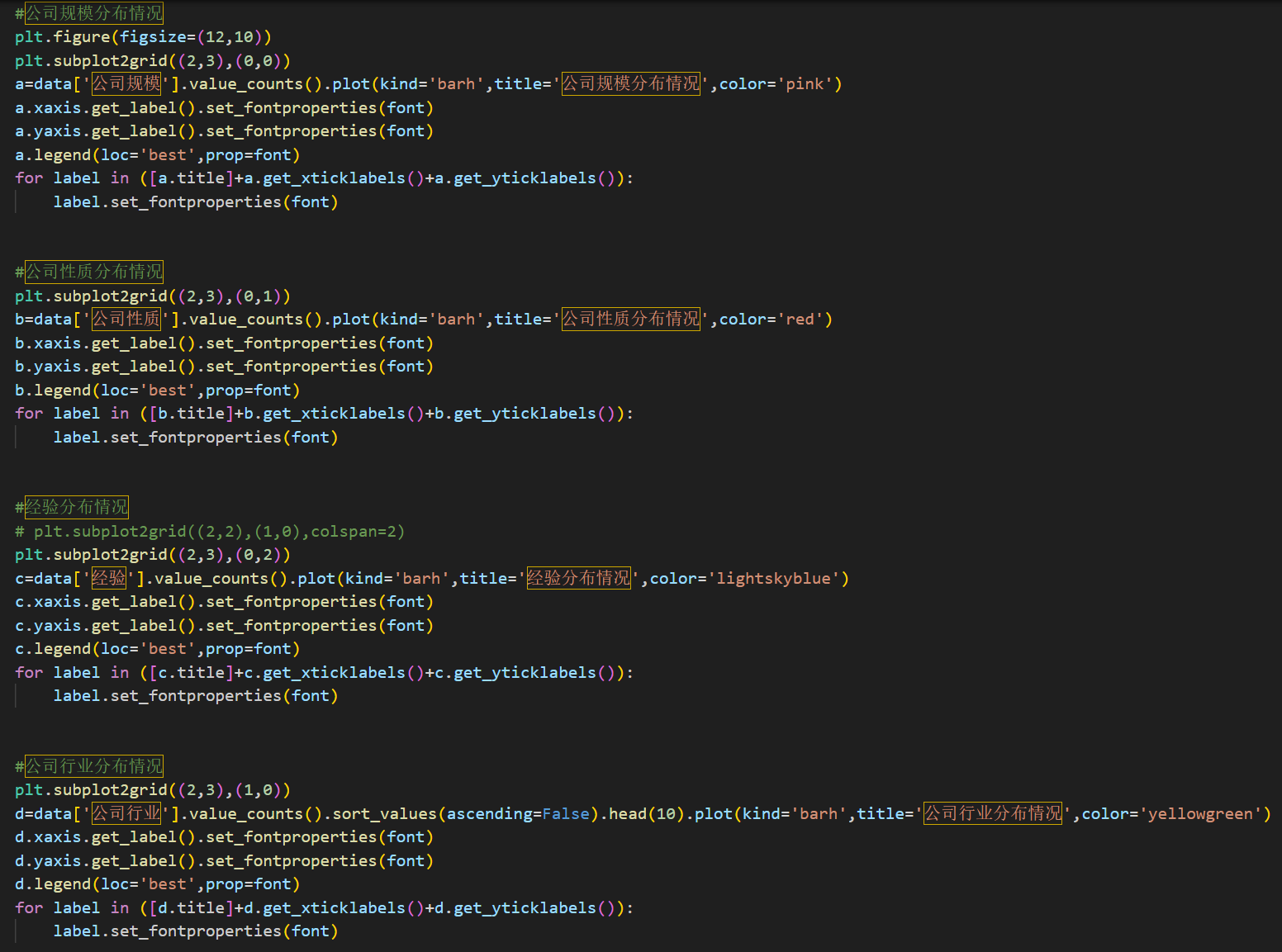


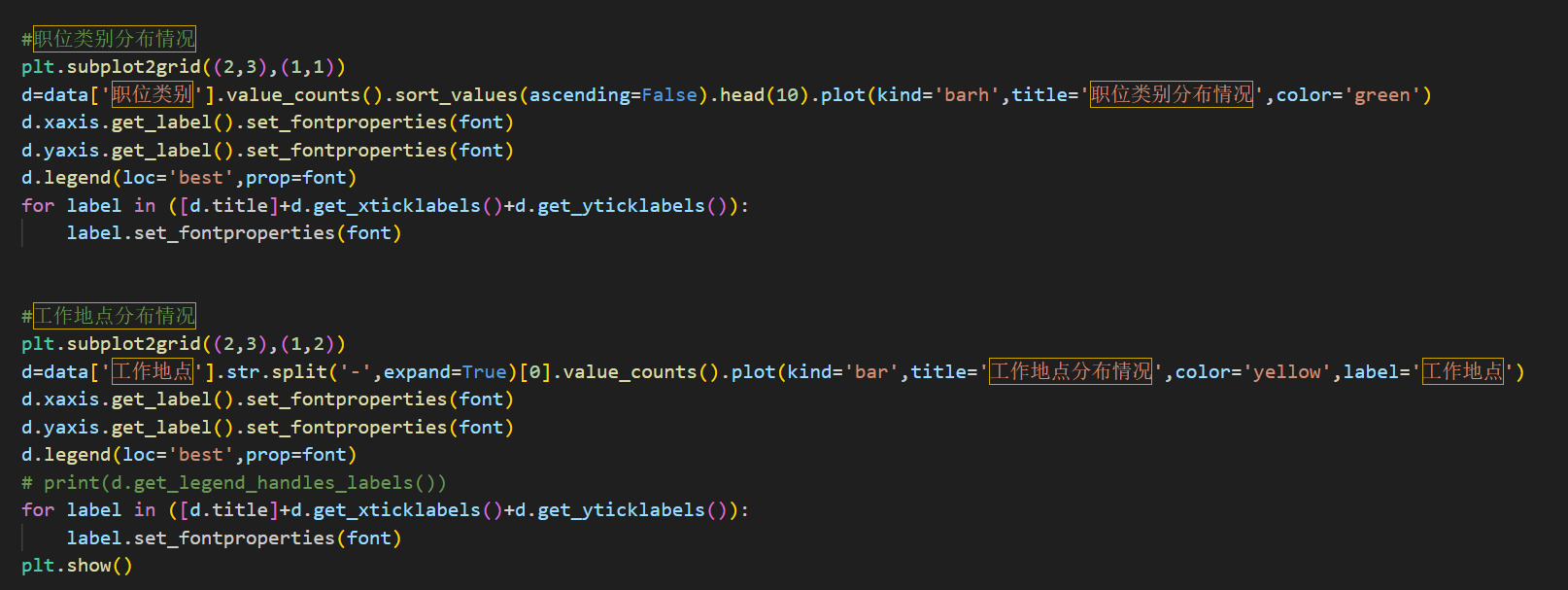
用PySpark SQL统计经验要求及数量。



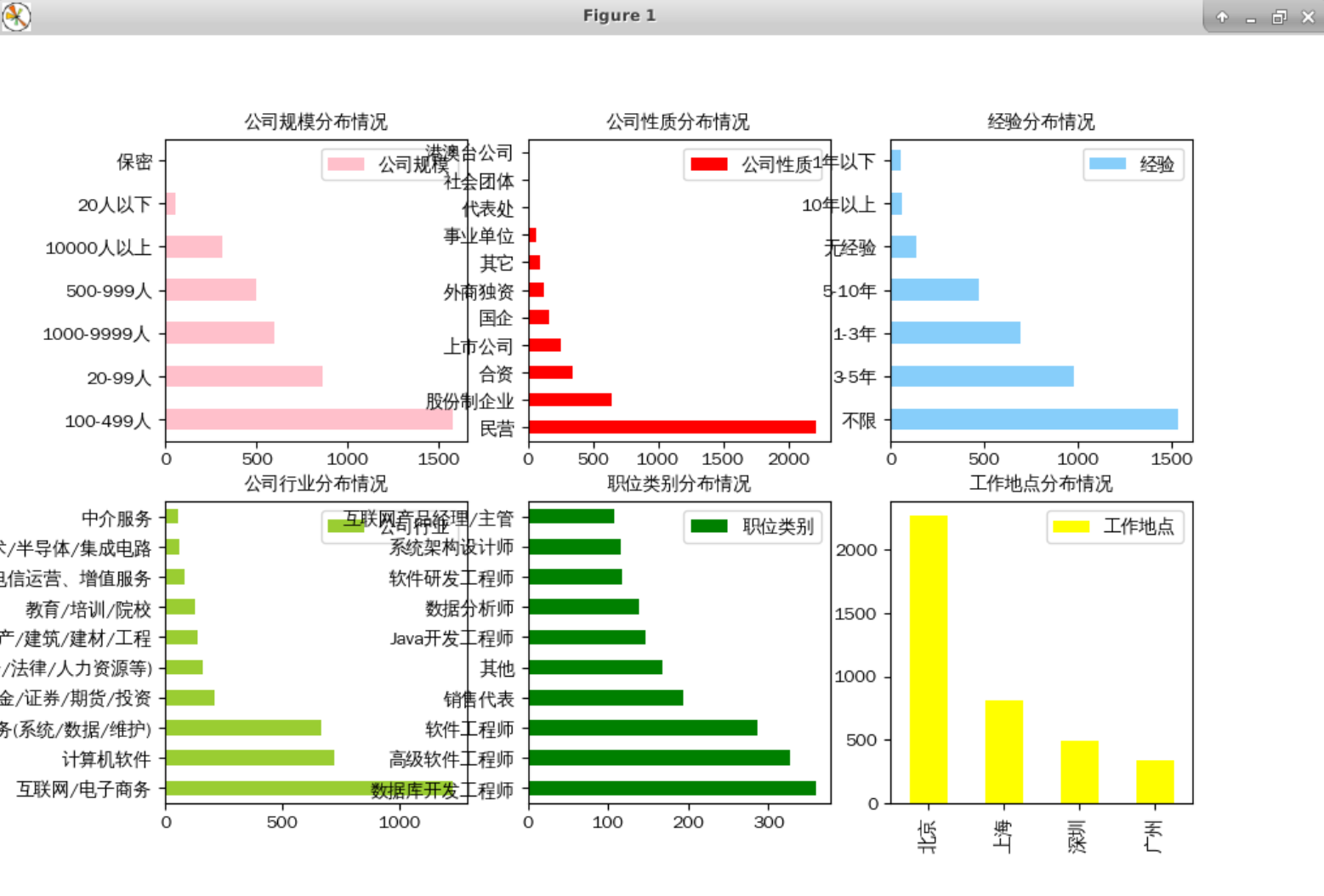
用PySpark SQL统计公司规模及数量，以便更好地分析，为接下来的数据可视化提供所需数据。

1. 利用之前创建的python\_pj3里面的数据对招聘职位信息进行探索分析，先导入相关的库做准备。接着分析公司规模分布情况，公司性质分布情况，经验分布情况，公司行业分布情况，职位类别分布情况，工作地点分布情况，最后进行可视化。
   1. 相关代码：





* 1. 可视化：



由上图可知：

公司规模在100-499人的公司招聘的大数据岗位最多。

公司性质为民营企业招聘的大数据岗位最多。

经验要求大部分没有明确说明，剩下的基本上集中在1-5年之间。

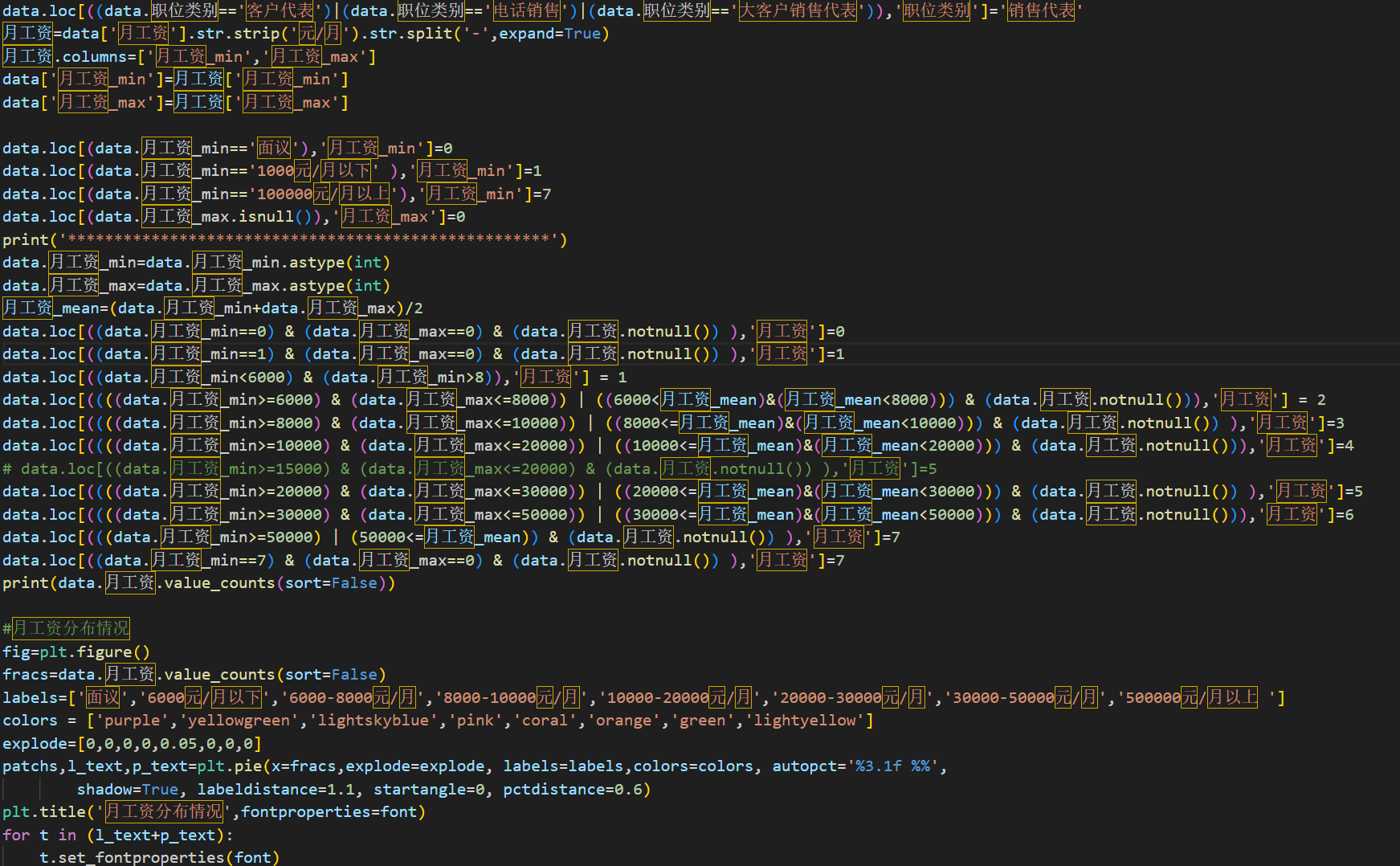
招聘公司主营行业主要集中在互联网、计算机、IT服务等行业。

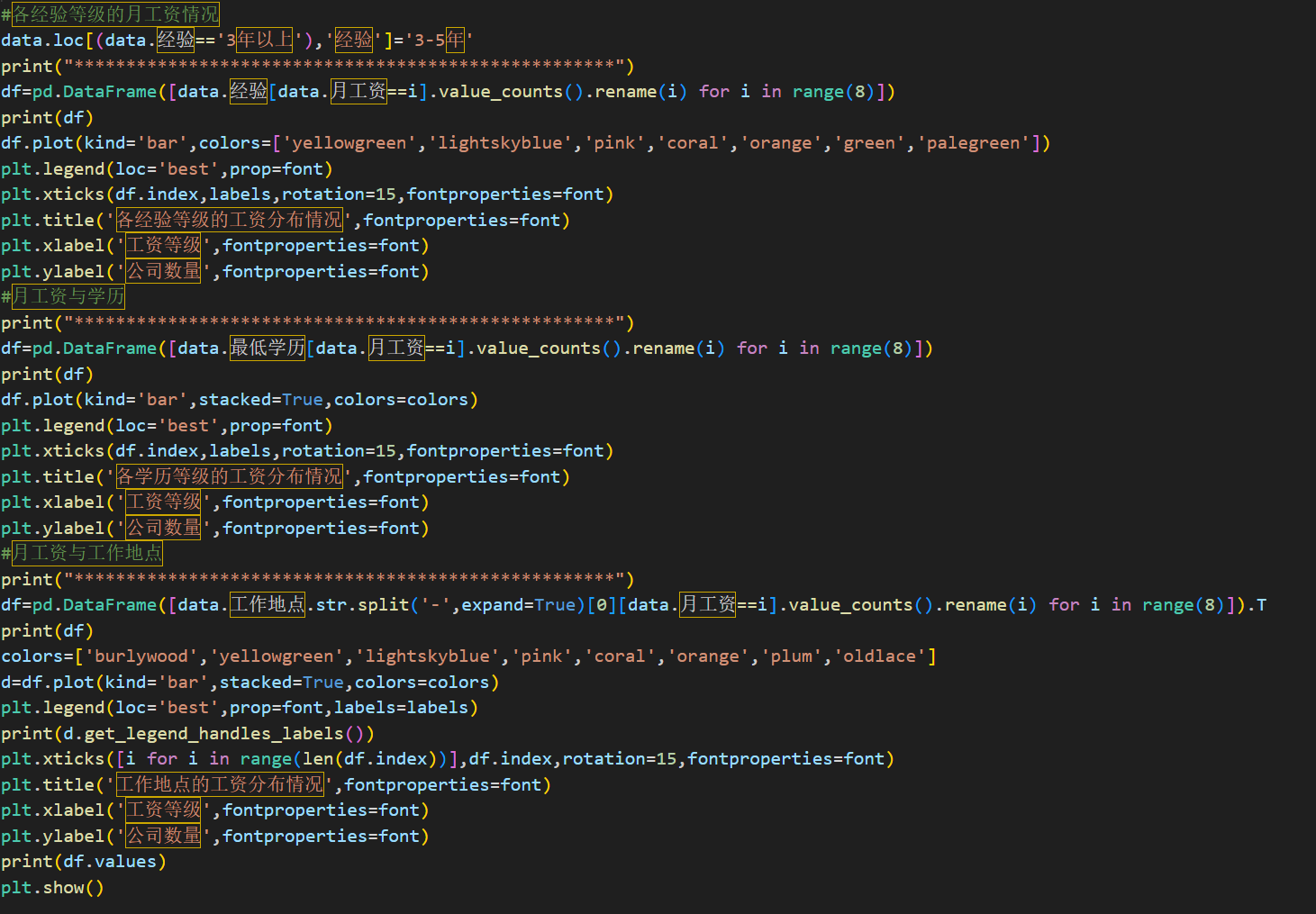
职位类别主要侧重于数据库开发，软件工程师等岗位。

工作地点主要分布在北京，广东最少。

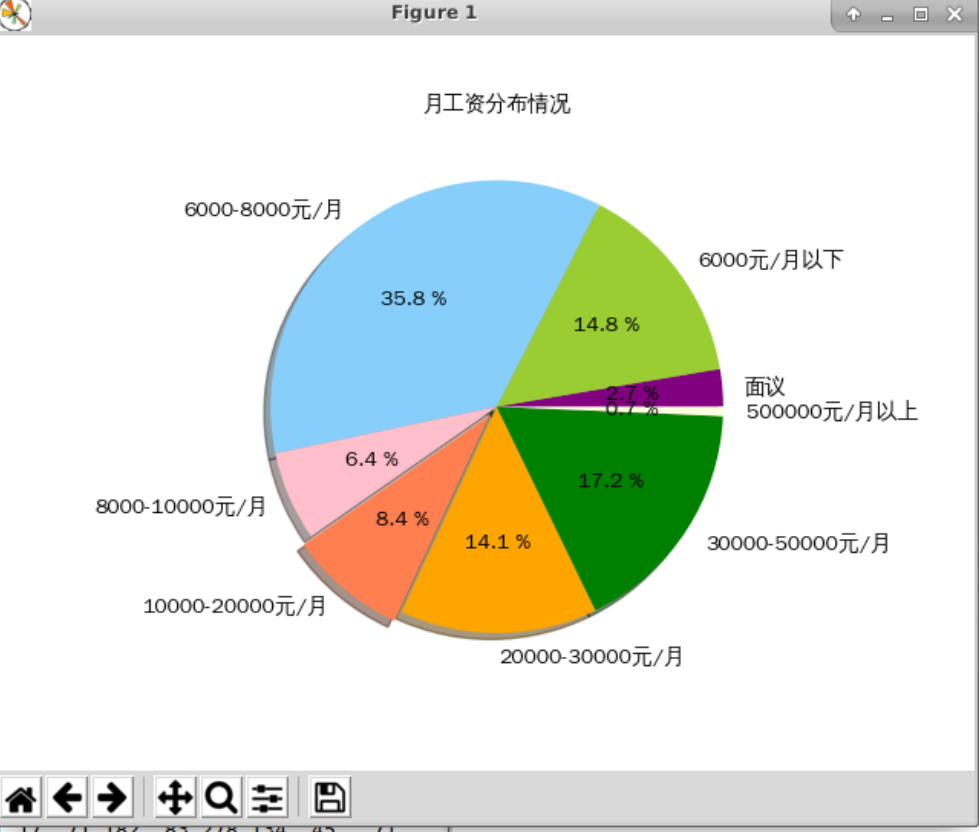
然后分析月工资分布情况与月工资相关属性的情况。

1. 相关代码

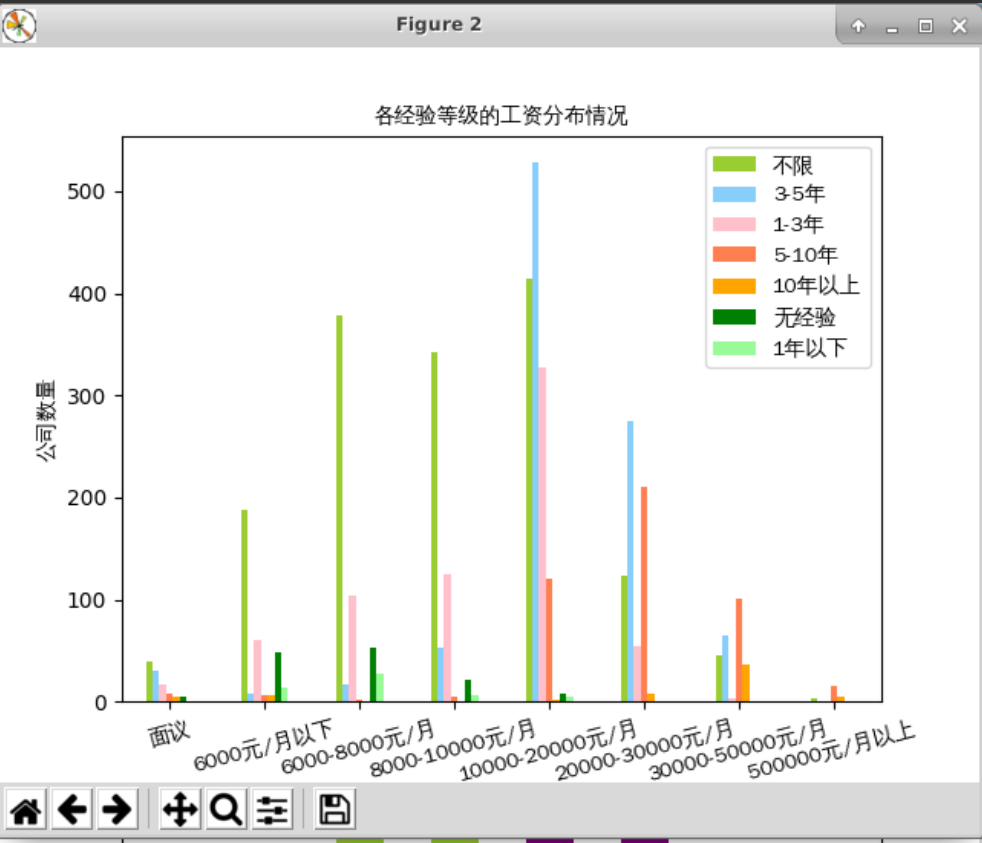




1. 可视化



由上图可知：公司招聘时，给的月工资主要集中在10000-20000元/月为35.8%，其次是20000-30000，10000-8000或6000-8000的也占很大比重，50000元/月以上的所占比例最小为0.7%，几乎可以省略。



由上图可知：

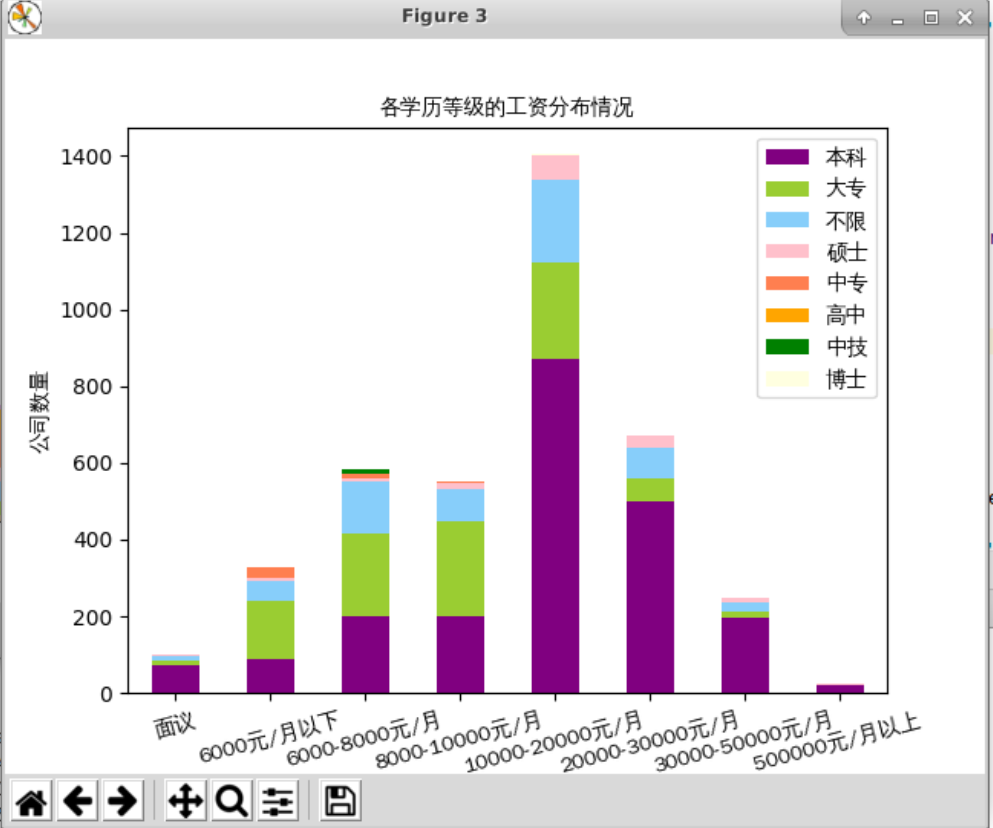
薪资在1万元/月以下，经验要求由高到低的是1-3年，1-3年、无经验

薪资在1-2万元/月，经验要求由高到低的是3-5年，1-3年、5-10年

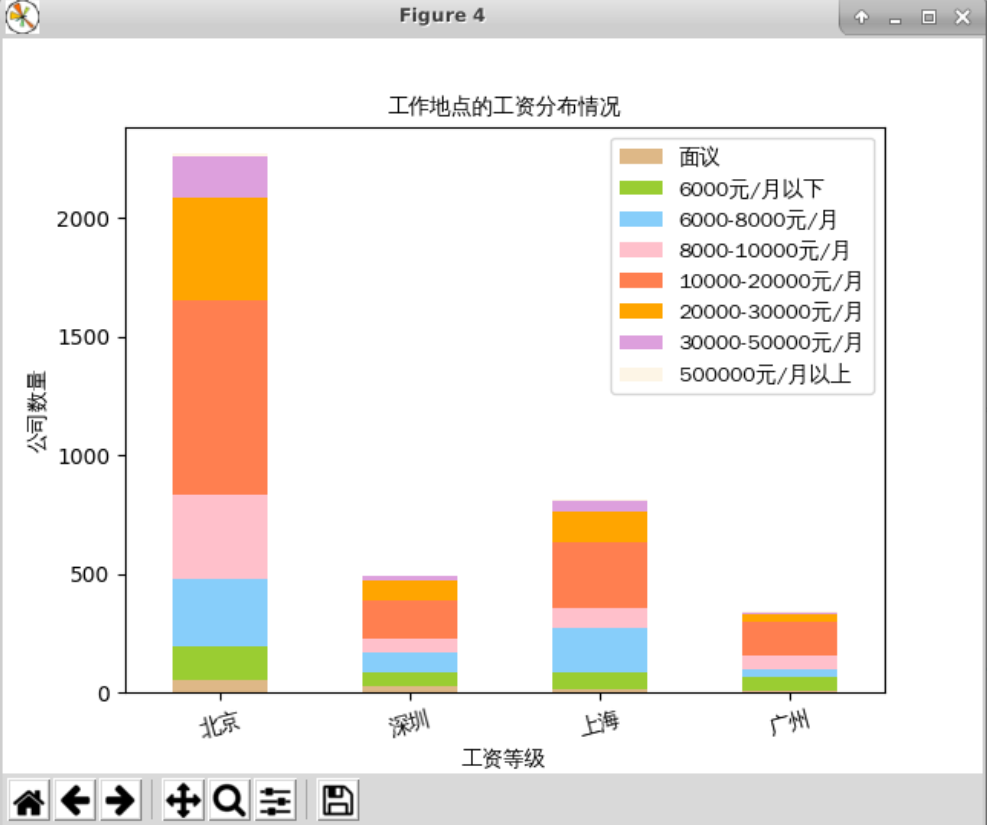
薪资在2-3万元/月，经验要求由高到低的是3-5年，5-10年、1-3年

薪资在3-5万元/月，经验要求由高到低的是5-10年，3-5年、10年以上

薪资在5万元/月以上，经验要求由高到低的是5-10年，3-5年、10年以上。

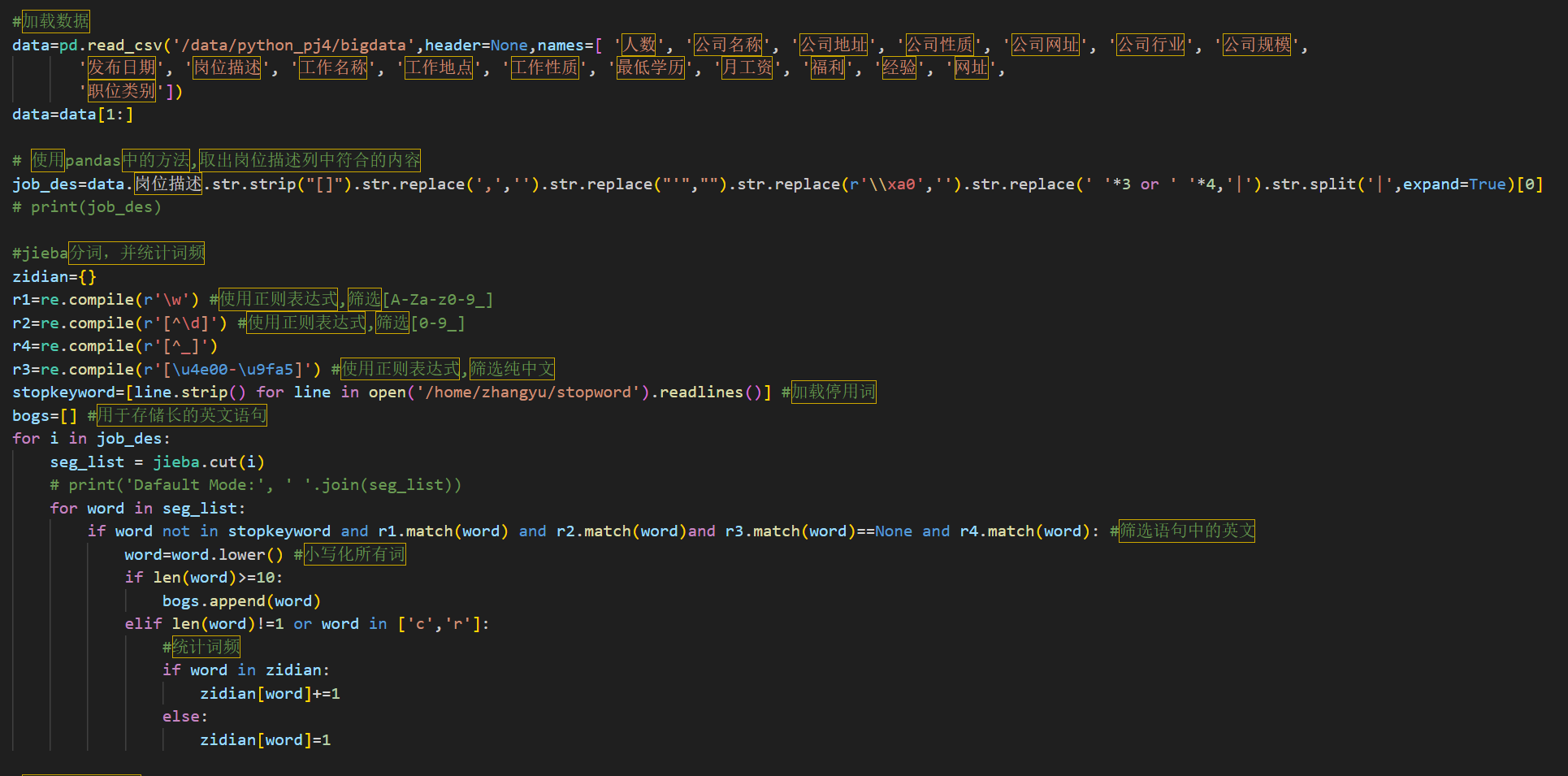


由上图可知：在各薪资范围内要求最低学历为本科的公司最多，中专、中技、高中几乎不存在。



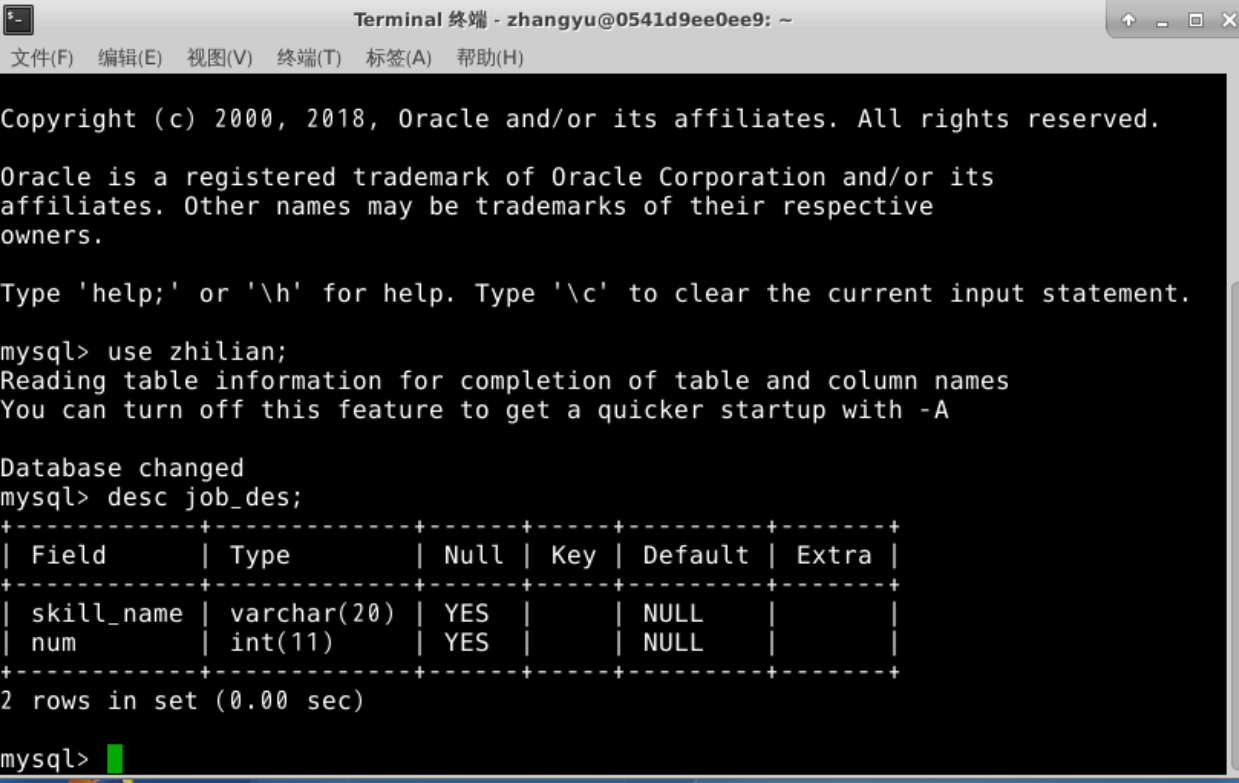
由图可知：各城市公司招聘给定的薪资10000-20000元/月所占比例最大，不同薪资分布比例基本较相似。

1. 使用结巴分词对岗位描述进行分词并将关键词统计
2. 结巴分词

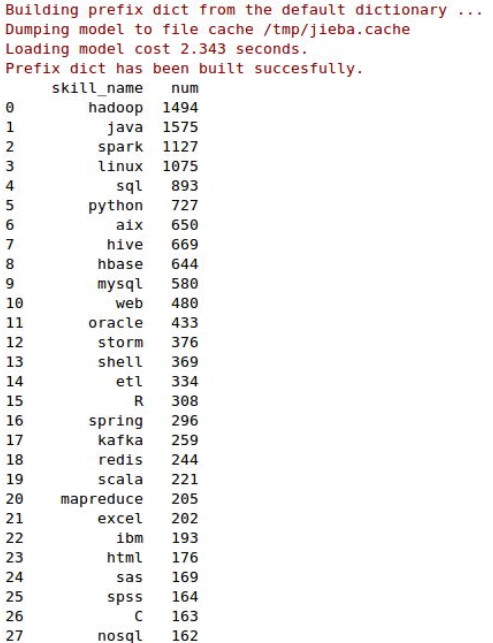




（2）查看job\_des表数据结构



（3）关键词统计



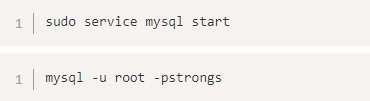
1. Django 和 Echarts 将职位分析结果进行可视化：

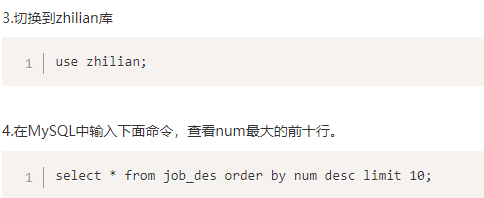
先在终端模拟器上构建Django环境



在pycharm创建Django项目命名为word文档构建可视化网址的具体组成部分，编辑其主要代码

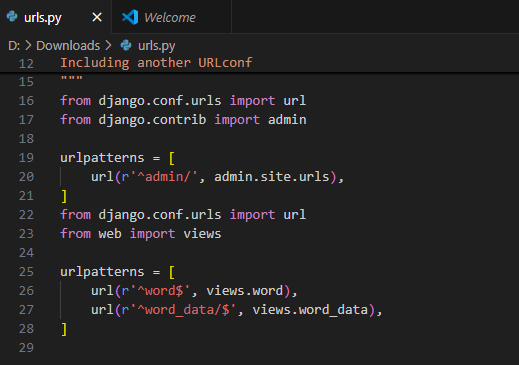
用终端模拟创建Django Model，并进入MySQL

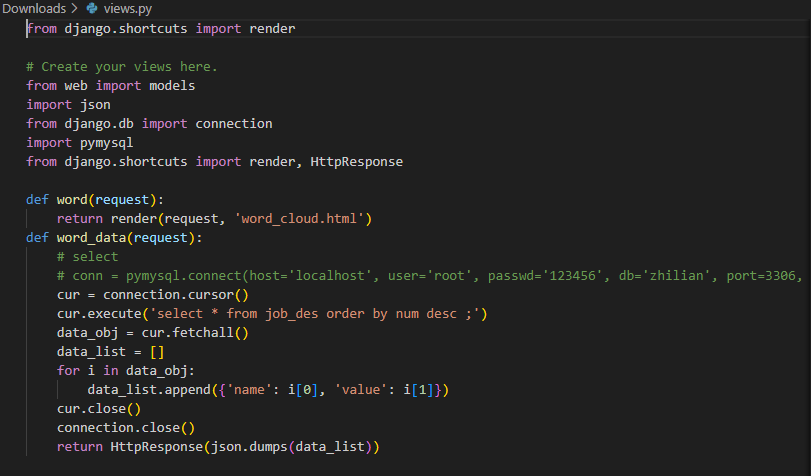




网页构建：

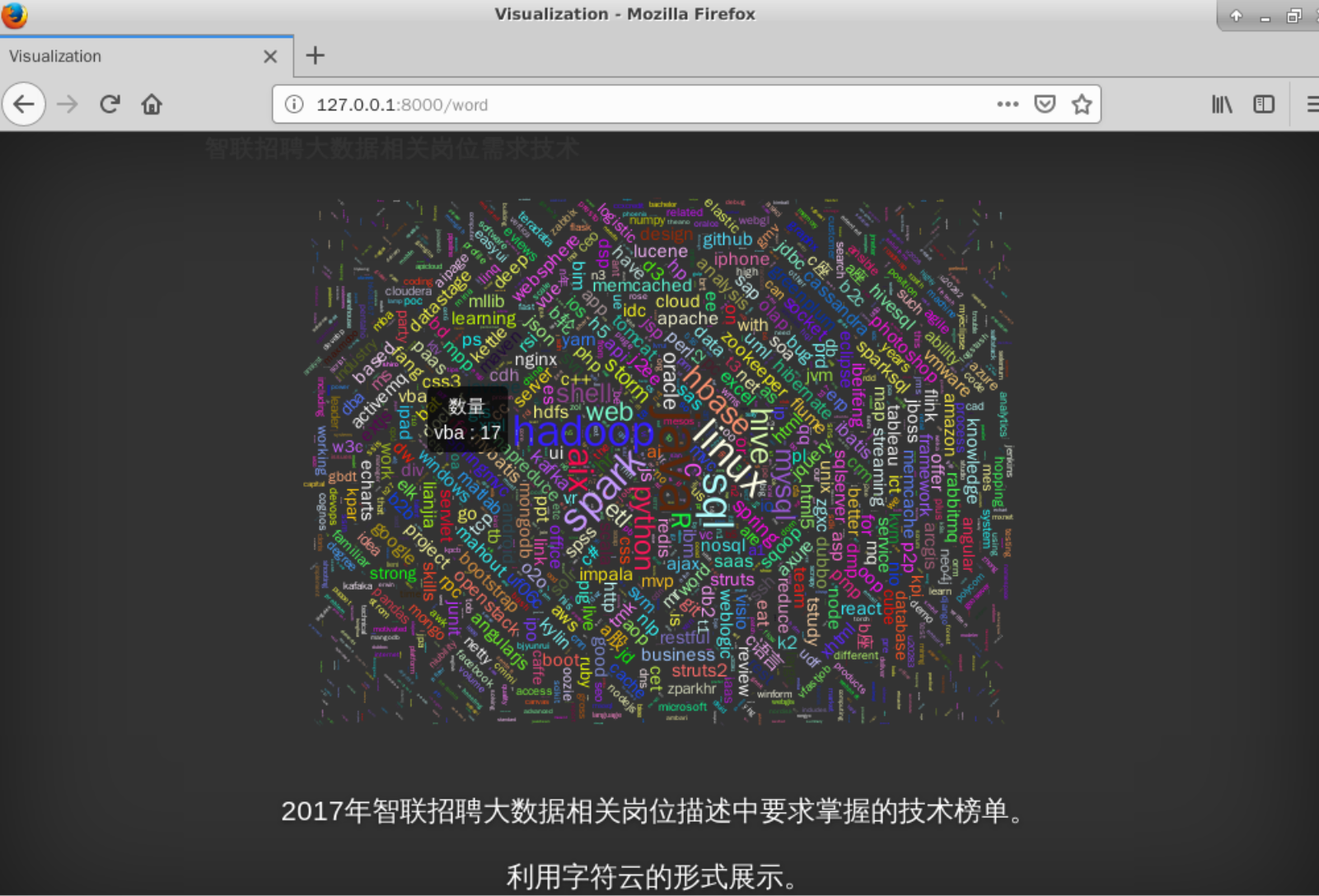
相关url和views设计代码



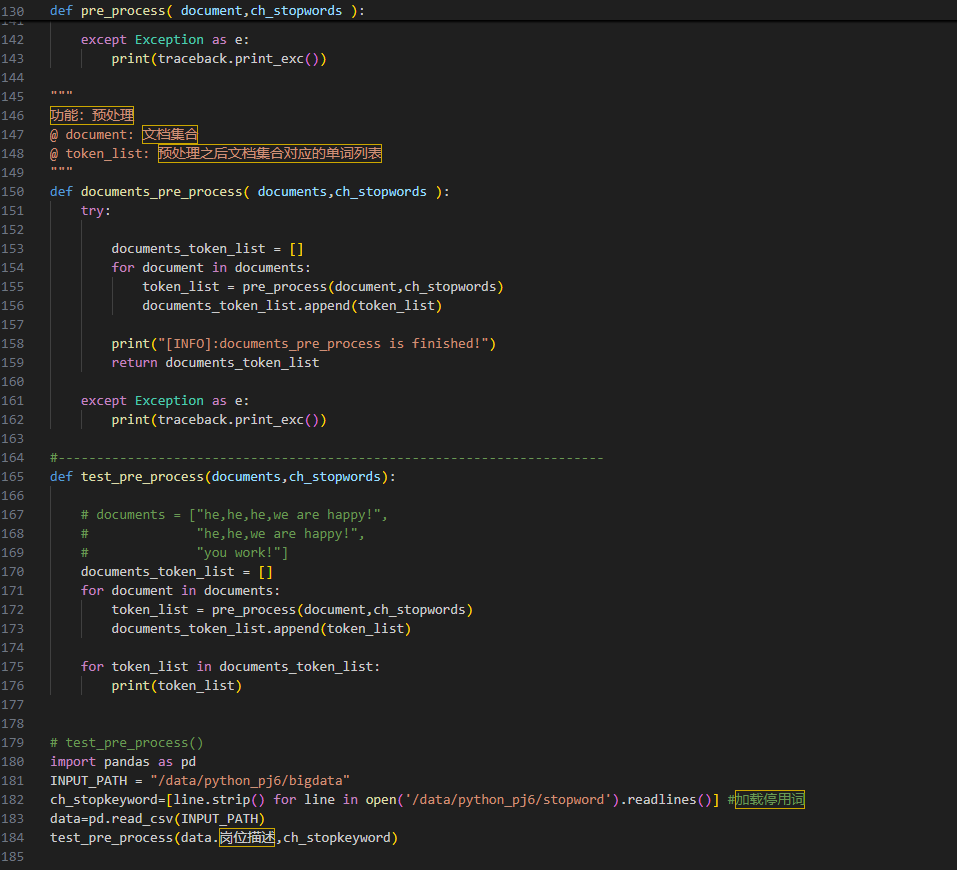
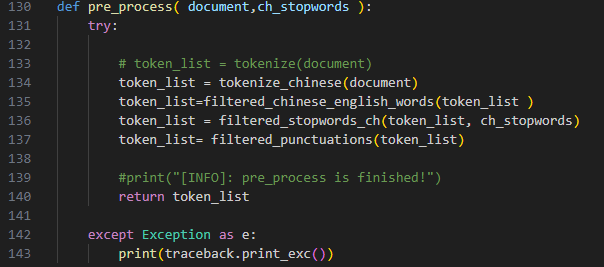
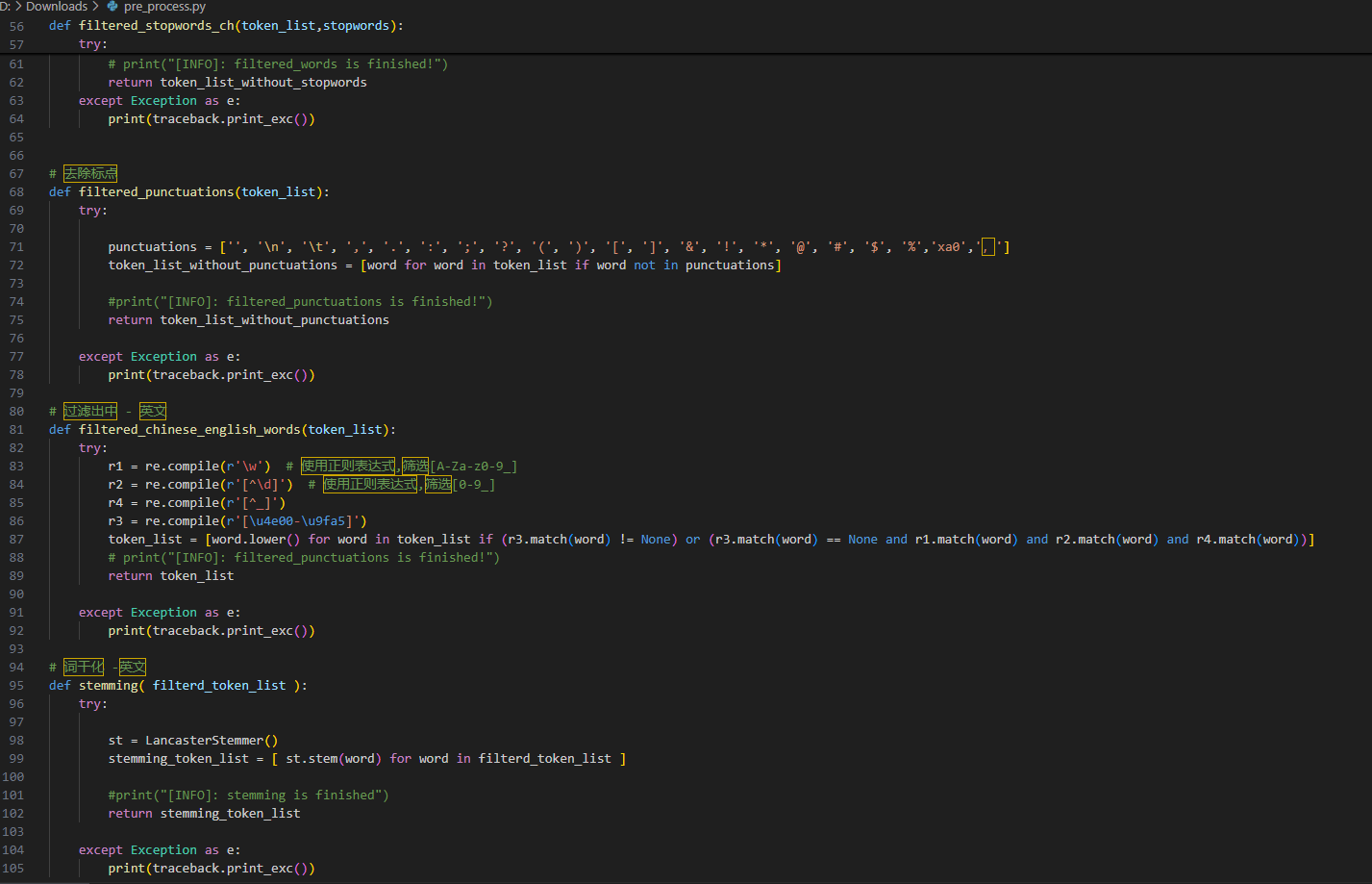
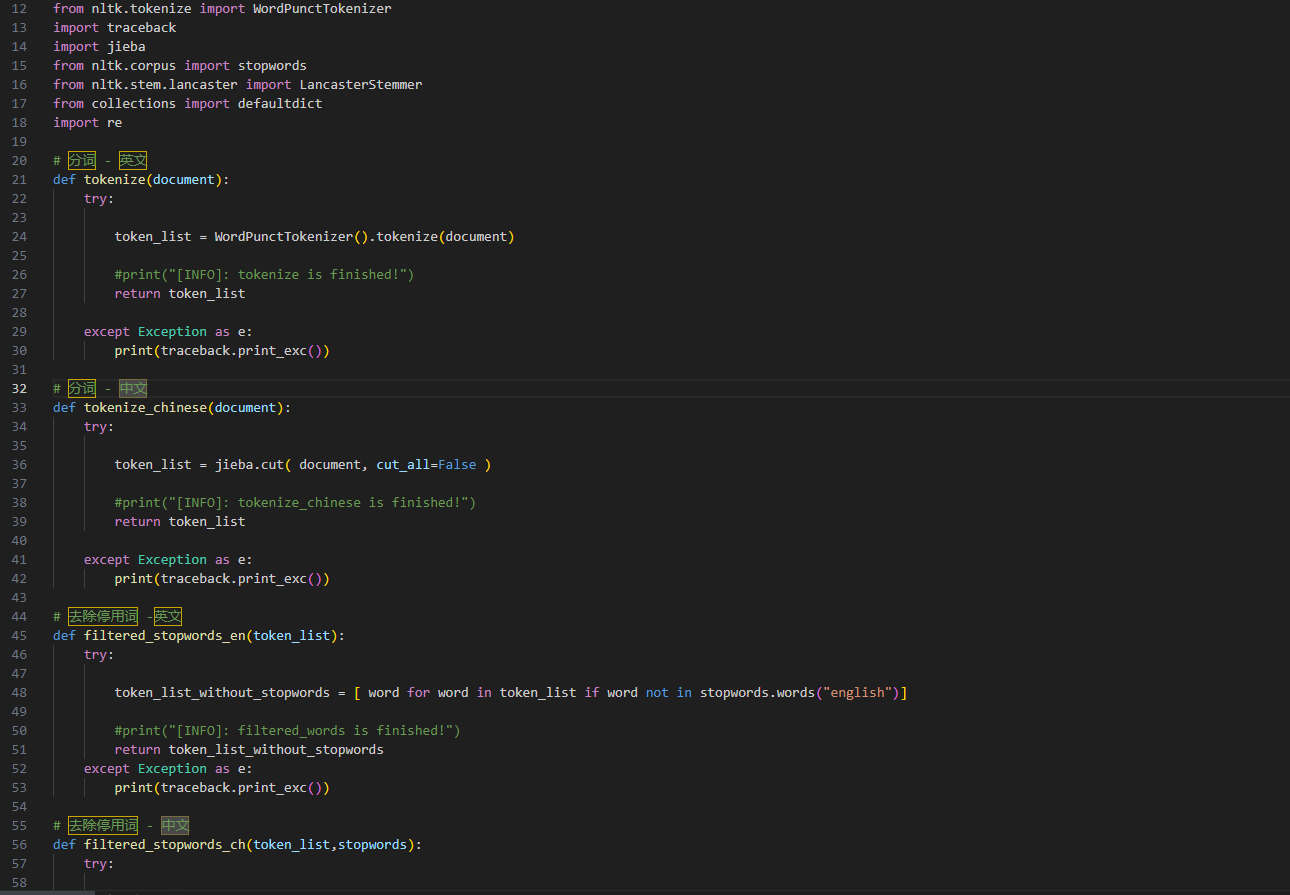


创建templates和static并解析word\_cloud.html

运行项目解析出地址http://127.0.0.1:8000/word

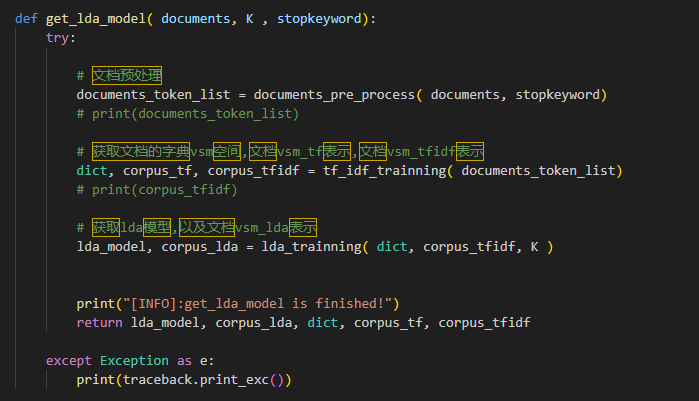


1. 建立LDA模型对职位描述进行相似度计算

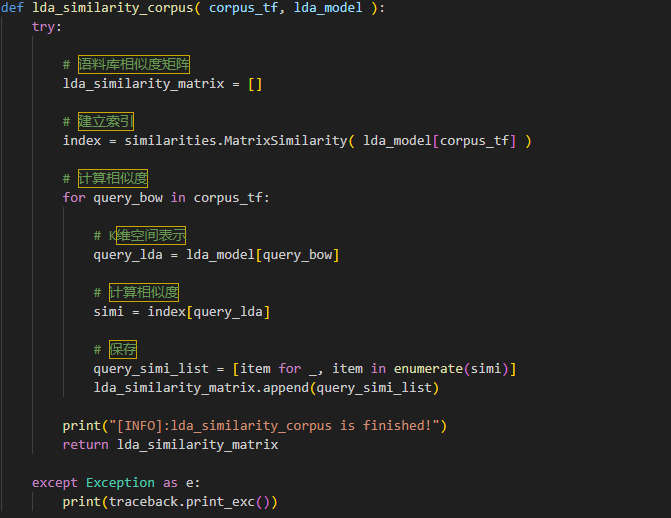
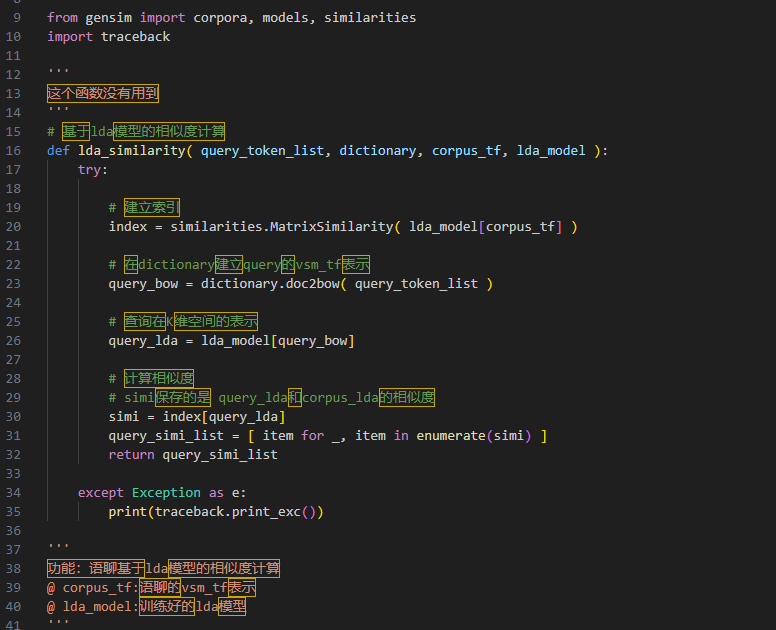
对招聘岗位信息中的岗位描述进行数据预处理

开始训练IDA模型

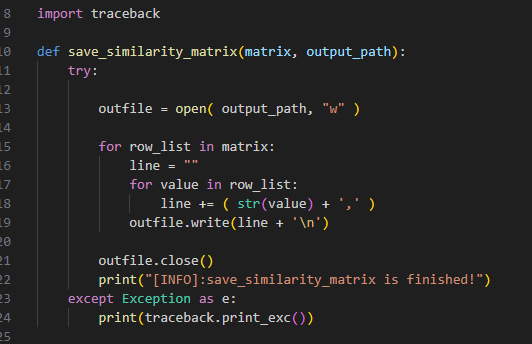


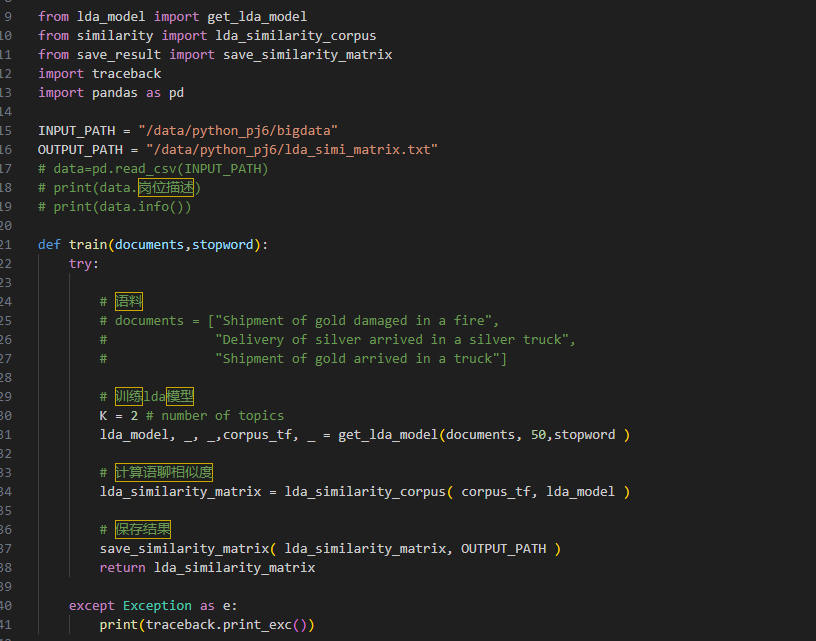


对训练完的lda模型进行预测



将结果进行保存

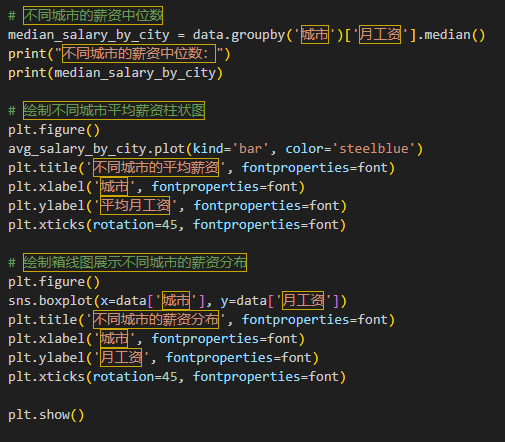
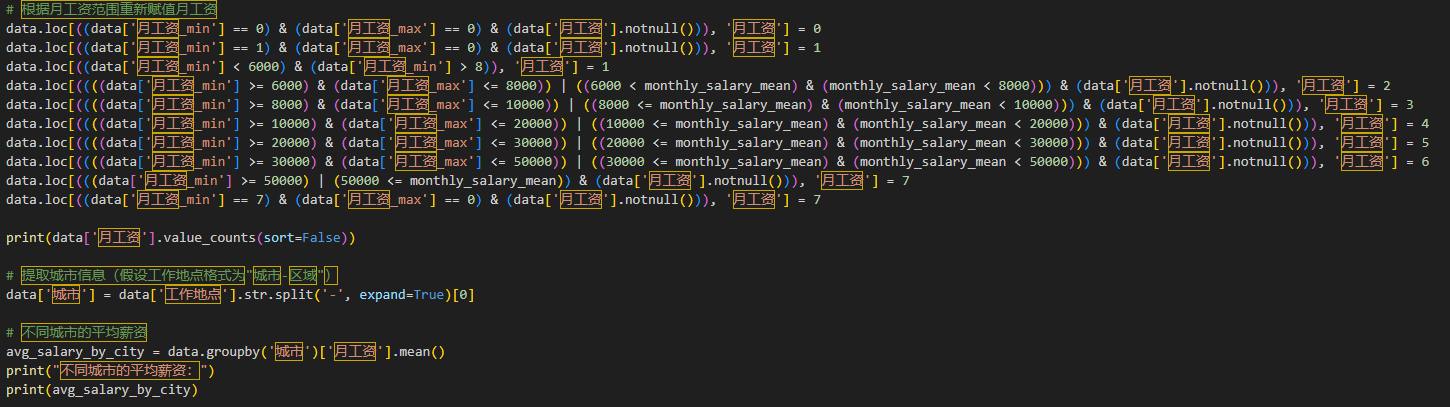


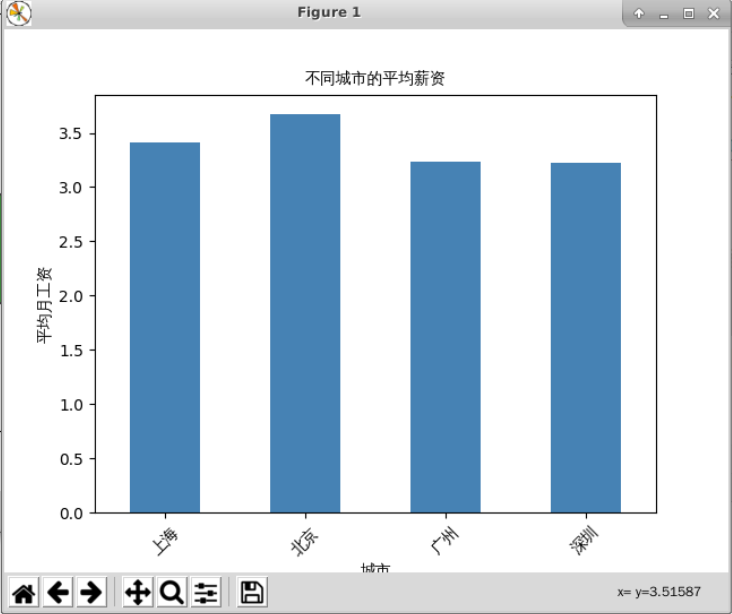
对汇总数据进行相似度的计算



额外的分析任务：

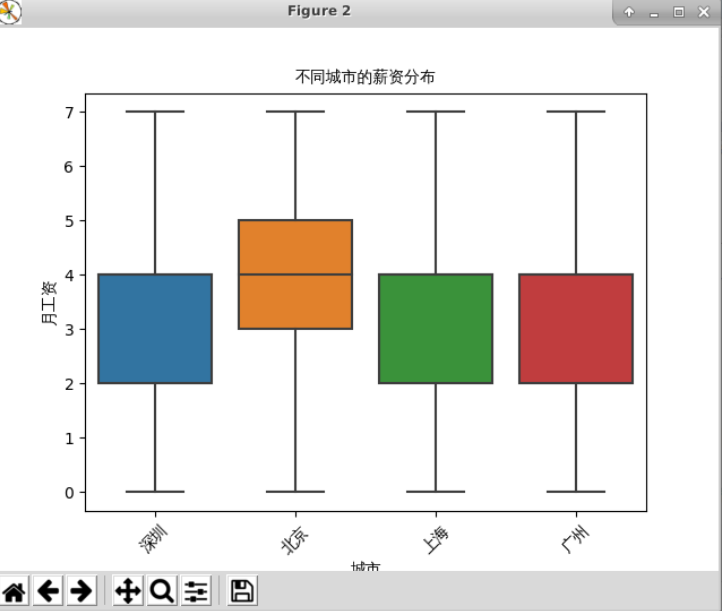
地区的不同是否会影响薪资的多少

代码如下

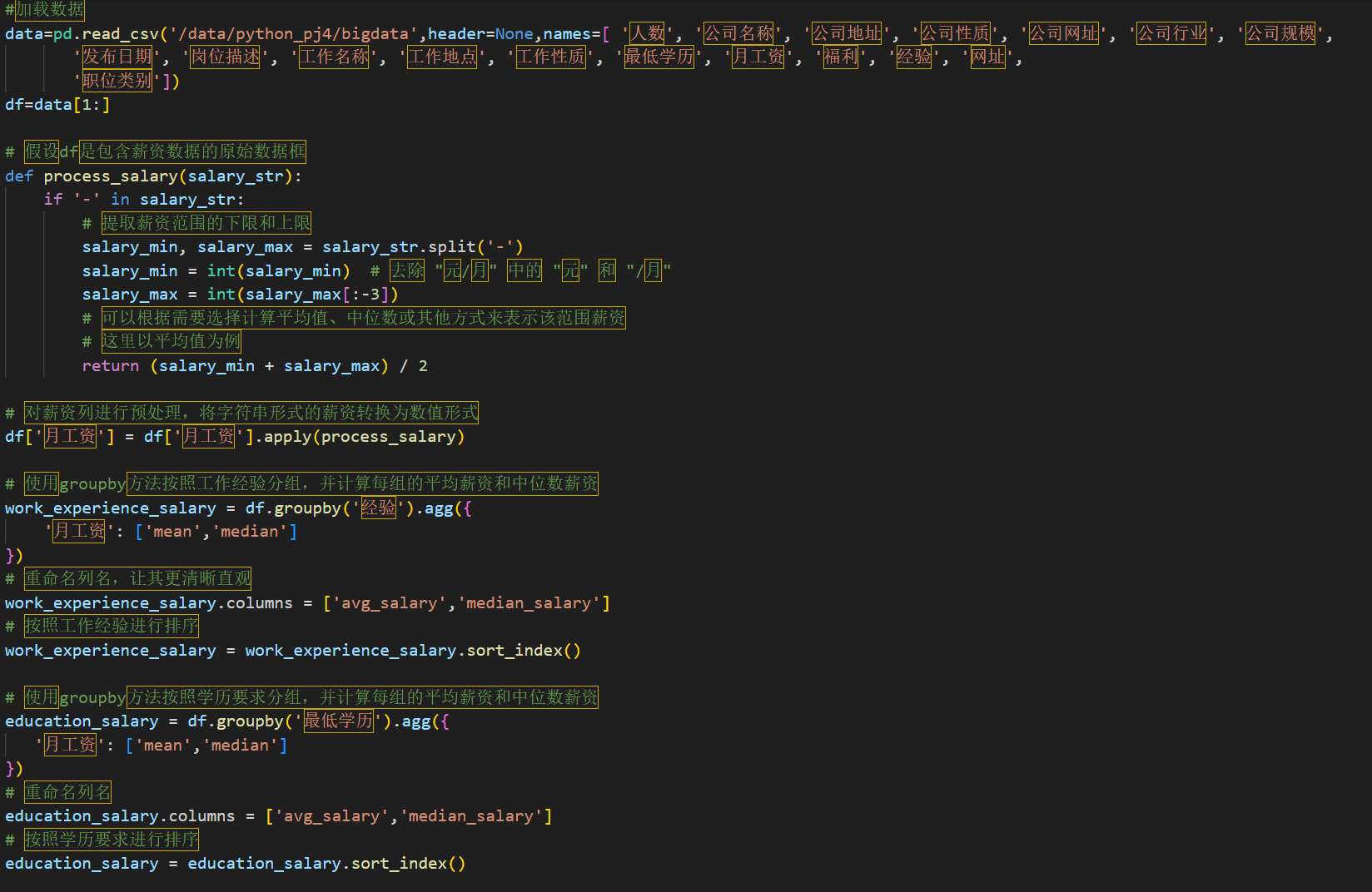
绘出直观图

由表一此可以看出在选择工作的时候，不同城市之间的薪资也是有一定的差距，北京为首，北京和上海的平均薪资相对较高，且两者之间差异不大。广州和深圳的平均薪资稍低一些，但彼此之间差距较小。平均薪资在一定程度上反映了城市的经济发展水平。北京和上海作为中国的重要经济中心，其平均薪资较高，可能是因为这些城市有更多高薪行业和工作机会。

由表二可以看出北京的薪资分布显示出较高的中位数和较宽的中间薪资范围，表明北京有较多的中高薪资岗位。深圳的薪资分布相对较低，可能反映出深圳的整体薪资水平略低于其他三个城市。

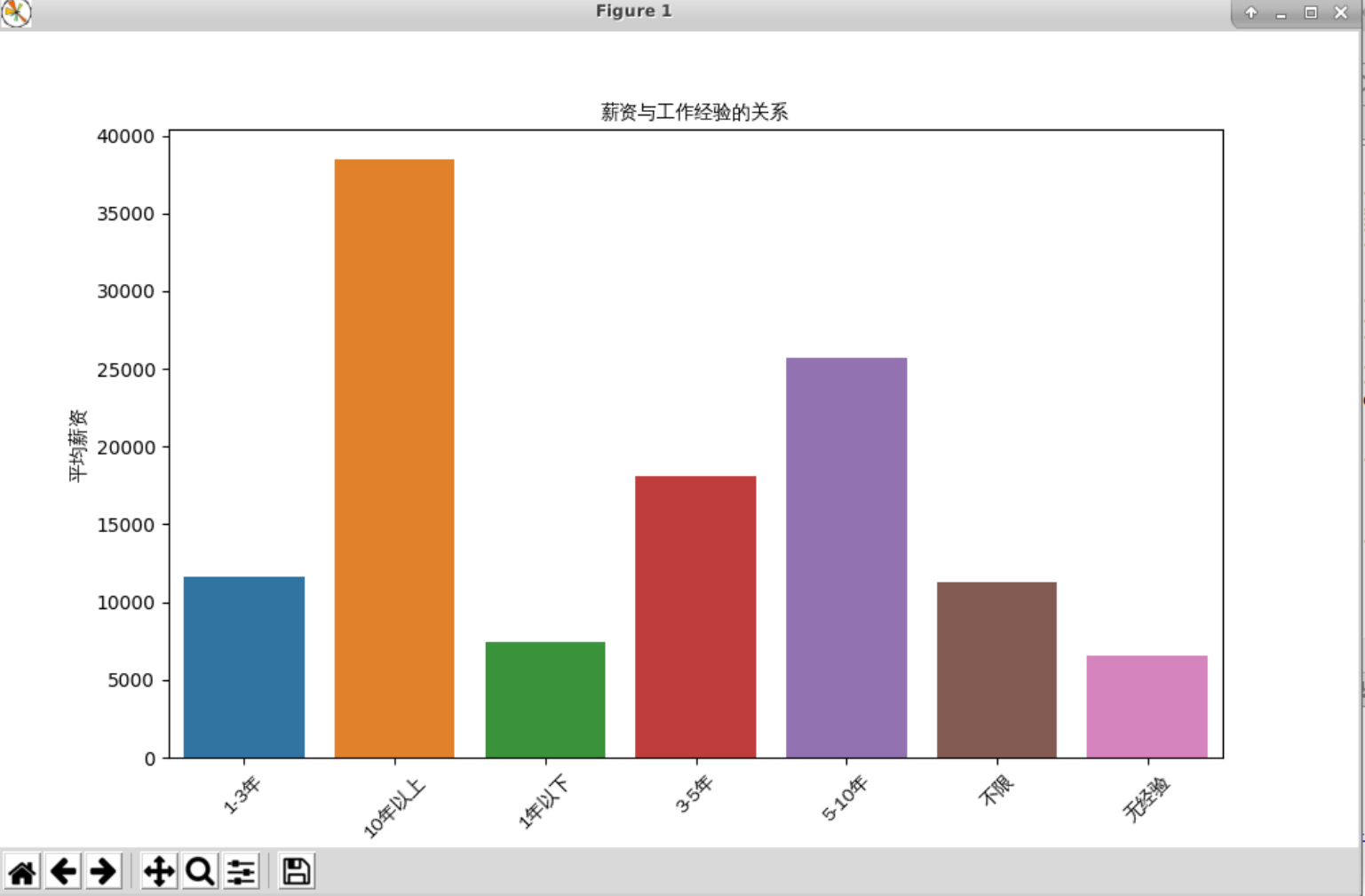


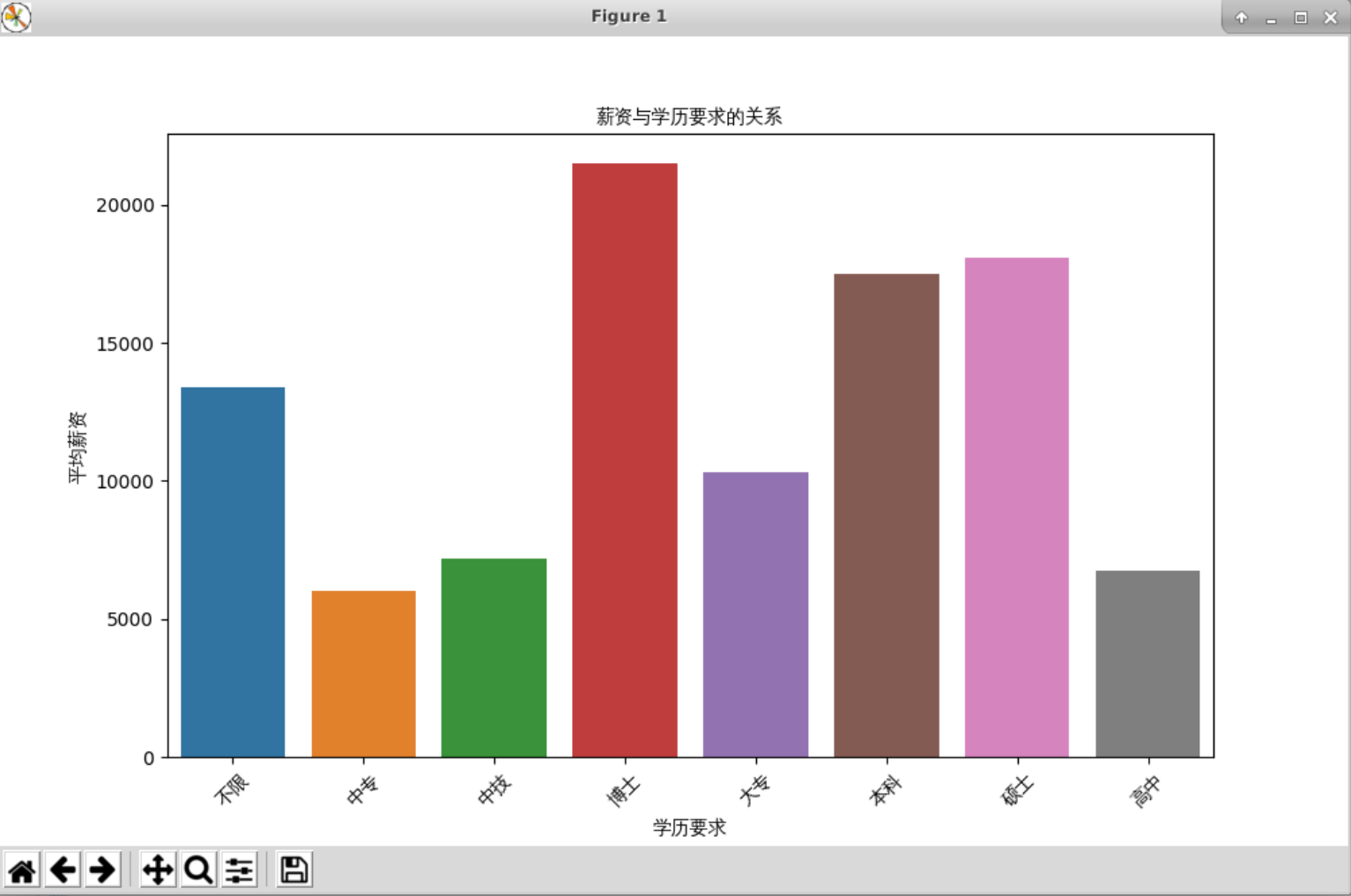
工作经验和学历对工资的影响





可视化

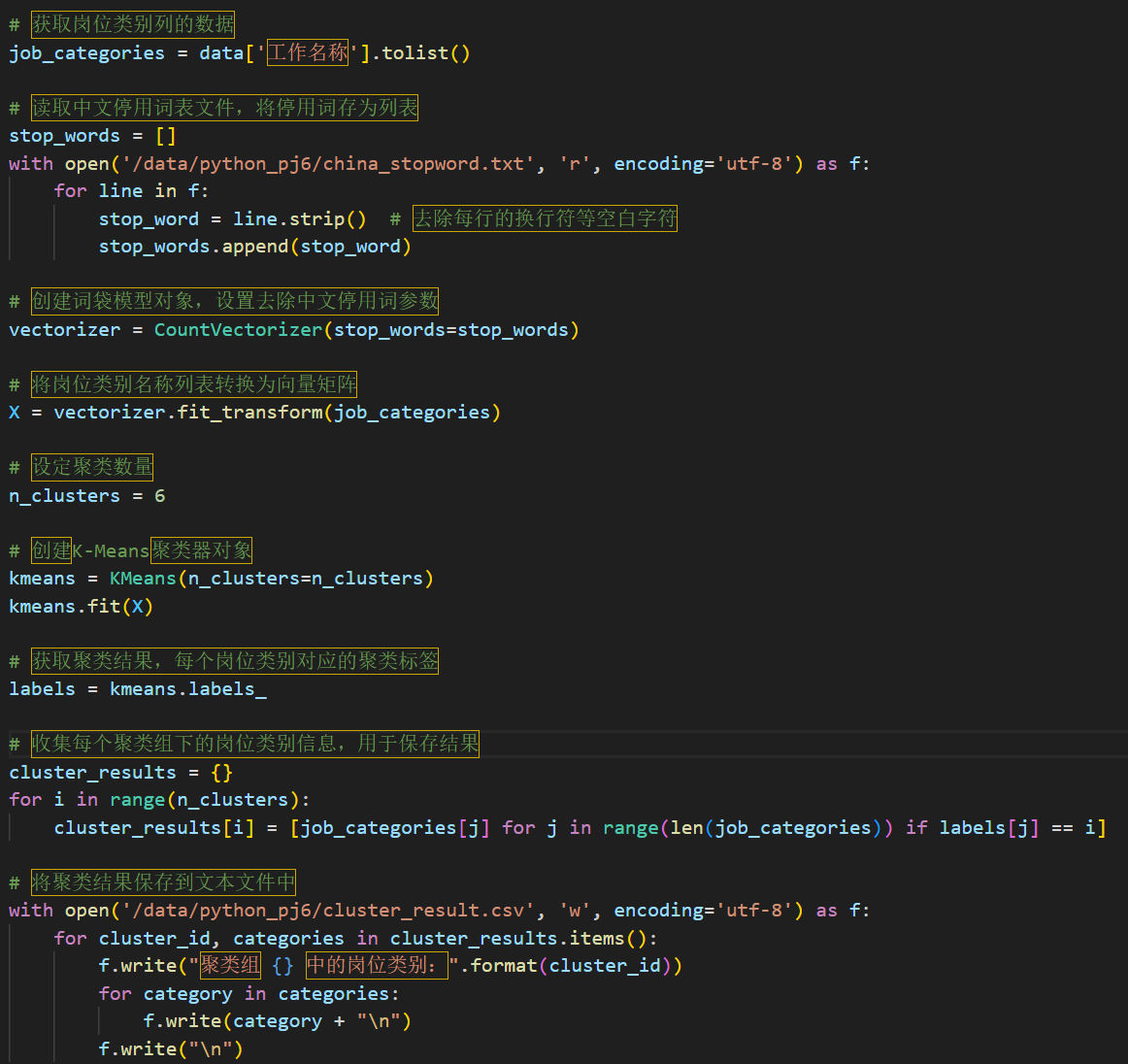




由上图可知：学历与工作经验越高对工资也就越高，所以我们要提升自己的学历，然后进入工作后不断增加自己工作经验。

对python6中的数据文件和中文停用词进行利用，对爬下来的工作名称进行聚类分析，对工作进行分类，对岗位的需求进行更细致的分析。

详细代码：



结果输出：

聚类组 0 中的岗位类别：大数据高级开发工程师

聚类组 1 中的岗位类别：数据分析师

聚类组 2 中的岗位类别：大数据开发工程师

聚类组 3 中的岗位类别：大数据分析 数据运营 接受应往届及转行人员

聚类组 4 中的岗位类别：移动网络开发6/大数据方向

聚类组 5 中的岗位类别：java软件开发工程师实习+转正

在文件中大数据高级开发工程师的数据量远远大于其他两个聚类，所以可以得出大数据高级开发工程师需求量非常大，可以往这方面发展。

# 项目总结

在本次项目过程中，我们将课堂上学到的数据分析方法、文本处理技巧等专业知识真正运用到实际操作中，不仅加深了对专业知识的理解，还提升了运用知识解决实际问题的能力，明白了理论与实践相互促进的关系。小组成员分工明确，共同完成数据收集、分析以及报告撰写等工作环节，在过程中大家相互沟通、协作配合，遇到问题时一起讨论解决方案，学会了如何在团队中发挥各自的优势，提高工作效率。本次实训数据分析方法相对较为基础，在处理一些复杂的文本信息和挖掘深层次关联时，显得不够精准和全面，没能充分挖掘出所有有价值的信息。