**2020年全国高校计算机能力挑战赛**

**大数据应用赛答题卡第六部分**

**队长姓名： 队长学校：**

# 第六部分：综合题（15分）

## 任务一、Spark技术栈有哪些组件，每个组件都有什么功能，适合什么应用场景。（5分）

**答： Spark中的组件有Spark core、SparkStreaming、SparkSQL、BlinkDB、MLBase、GraphX等。**

**组件功能如下：**

**Spark core：其它组件的基础，spark的内核，封装了底层通讯框架，是spark的基础。**

**SparkStreaming：是一种可以对实时数据流进行高通量、容错处理的流式处理系统，可以对多种数据源进行类似于hadoop中的map、reduce和join等复杂操作，将流式计算分解成一系列短小的批处理作业。**

**SparkSQL：能够统一处理关系表和RDD,能进行更复杂的数据分析。**

**BlinkDB:是一个用于在海量数据上运行交互式SQL查询的大规模并行查询引擎。它允许用户通过权衡数据精度来提升查询响应时间，其数据的精度被控制在允许的误差范围内。**

**MLBase：用于机器学习。**

**GraphX：用于图并行计算。**

## 任务二、本题目使用spark进行数据分析，不限制语言（可以使用pyspark，Rspark，scala等语言）（10分）

数据 

数据说明：

|  |  |
| --- | --- |
| 字段 | 字段说明 |
| positionName | 职位名称 |
| salary | 薪水 |
| workYear | 工作年限 |
| city | 城市 |
| companyShortName | 公司简称 |
| companySize | 公司规模 |
| district | 所在区 |
| financeStage | 融资阶段 |
| industryField | 所在领域 |
| thirdType | 职位类型 |
| resumeProcessDay | 简历日处理 |
| resumeProcessRate | 简历处理率 |
|  |  |

### 拆分字段salary -> min\_salary,max\_salry，并且取薪资的整数 （2分）

如果salary字段是否包含类似'10k-20k\*15薪'的值 如果存在 则把 \*15这样的类似数据给去掉，保留10K-20K; 则拆分为 min\_salary,max\_salry 分别是 10，20，

如果salary字段 类似10-20K 则拆分为 min\_salary,max\_salry 分别是 10，20，

如果 salay字段为空值或者 面议，则拆分为 min\_salary,max\_salry 分别是 -1，-1，

拆分后字段 表结构为

|  |
| --- |
| 字段 |
| positionName |
| salary |
| min\_salary |
| max\_salry |
| workYear |
| city |
| companyShortName |
| companySize |
| district |
| financeStage |
| industryField |
| thirdType |
| resumeProcessDay |
| resumeProcessRate |
|  |

源码：

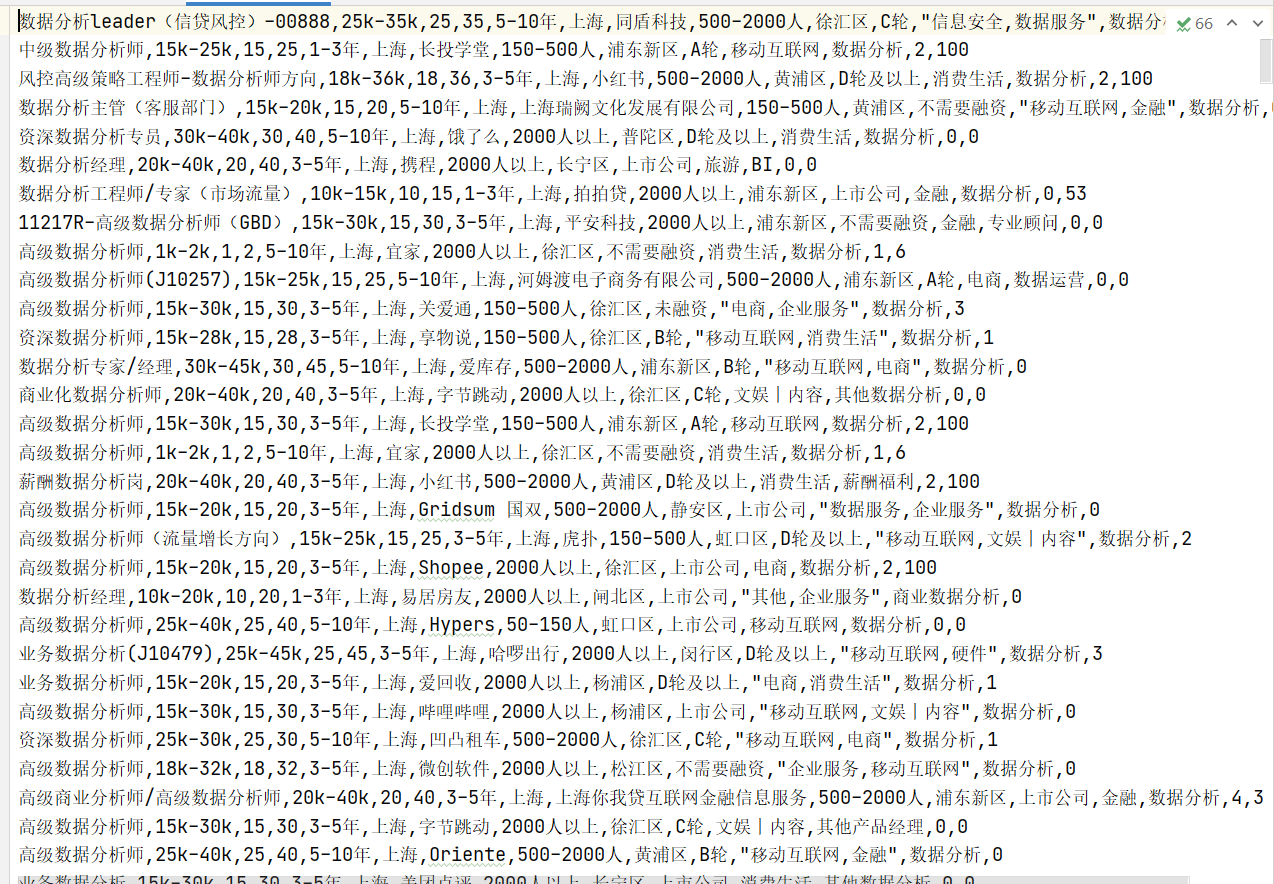


代码截图：





结果截图：



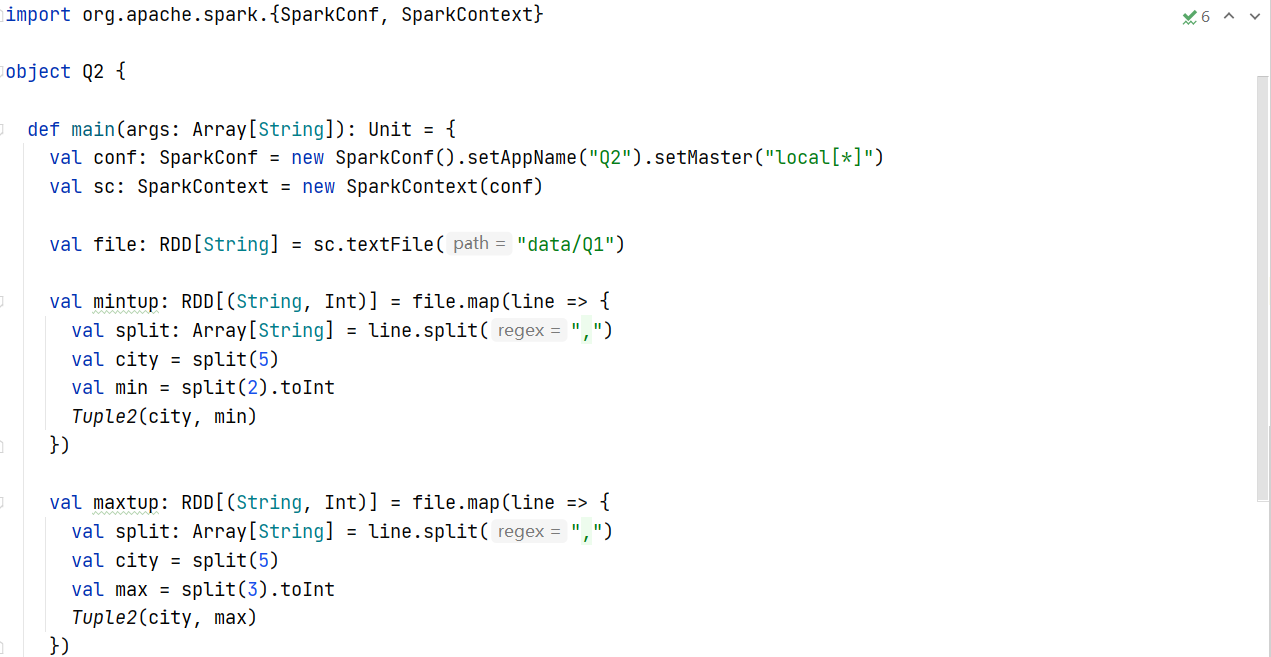
### 分析不同城市薪资概况（工资的最小平均值，最大平均值）（2分）

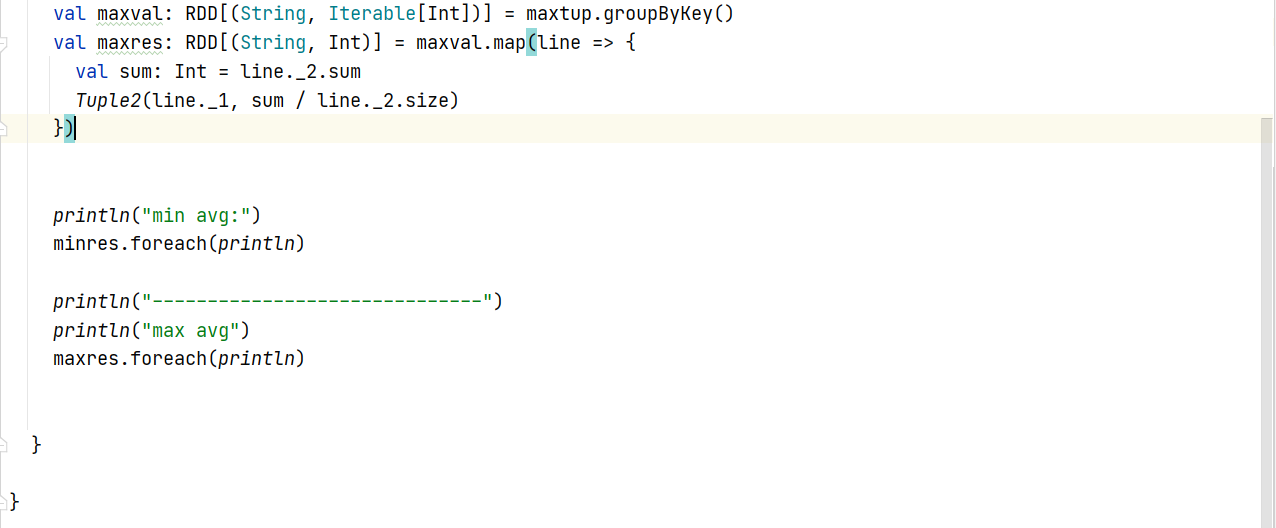
返回结果包含 城市名称，薪资最小平均值avg（min\_salry），最大平均值avg（max\_salry）

源码：

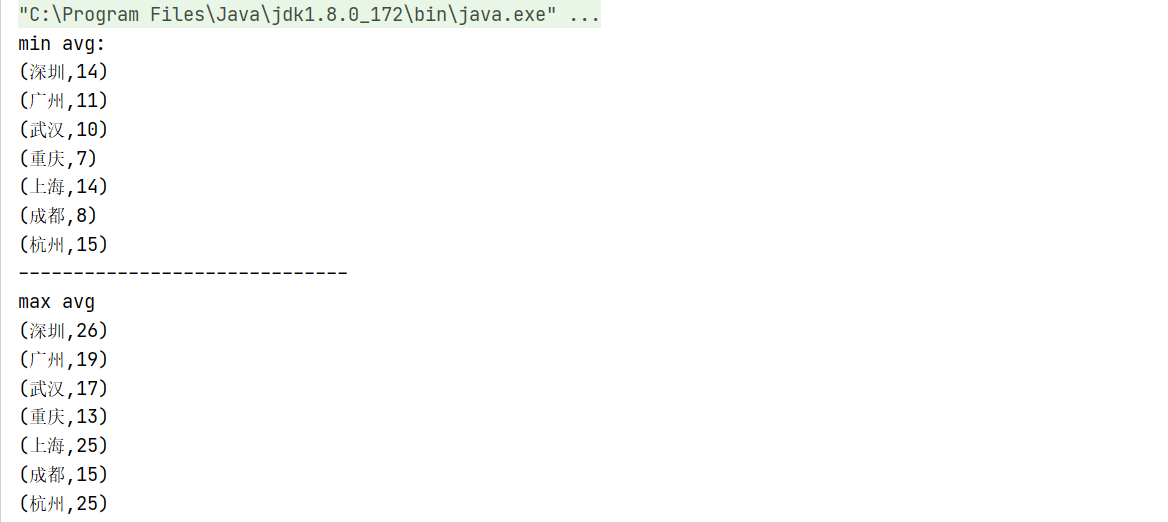


代码截图：





结果截图：



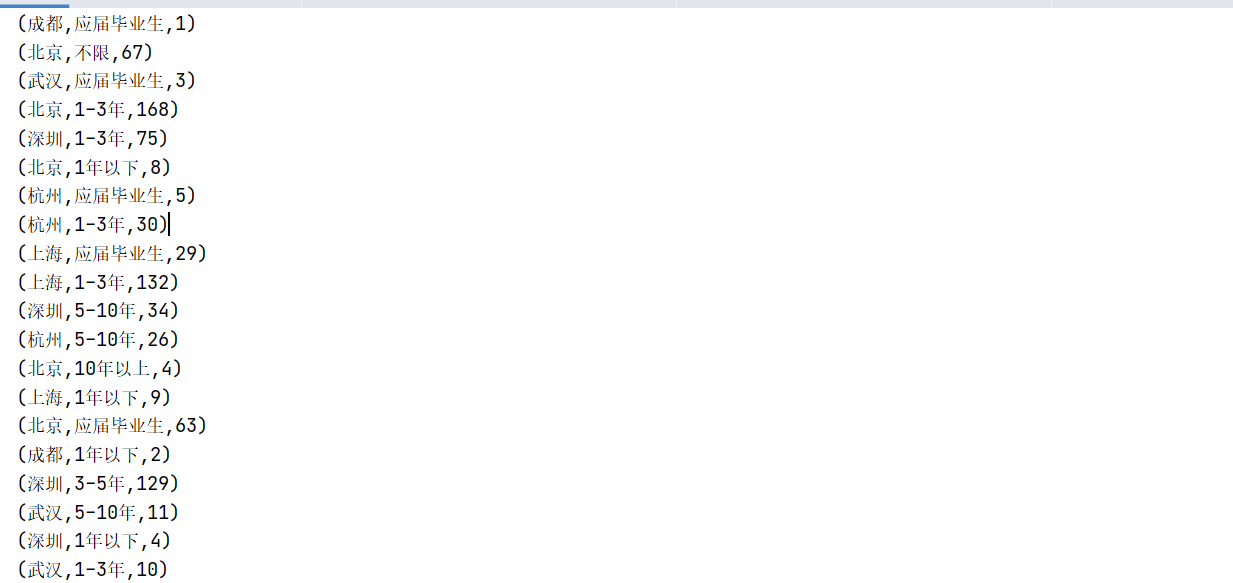
### 分析不同城市 不同工作年限公司职位需求数量（2分）

返回结果包含 城市名称，工作年限，需求数量，

源码：



代码截图：



结果截图：



### 分析获得每个城市，在 职位名称中包含 “数据分析“” 职位需求 公司名称 按照需求数量倒序排列（4分）

返回结果包含 城市，公司名称，需求数量，

源码：

代码截图：

结果截图：