**2020年安徽省大数据与人工智能应用竞赛A卷**

**学校名称： 队伍名称： 队伍编号：**

# 第二部分：大数据程序编写部分（20分）



数据文件 某公司订单，退货数据，其中order.csv为订单数据，return是退货数据。

订单数据格式为：

订单ID 订单日期 省/自治区 产品ID 类别 销售额 数量 折扣 利润

退货数据格式为 ：

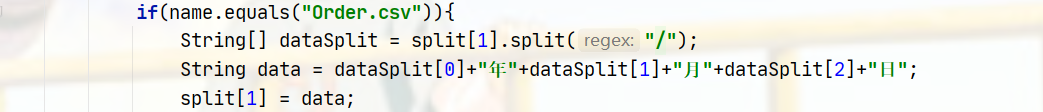
订单ID 退回

### 对订单数据中的订单日期进行格式化，并派生出年，月，日三个字段，对订单数据和退货数据两个数据进行join，合并后数据 在订单数据基础上增加 是否退货一列，并把合并后的数据输出到 dsjjs2020a（10分）

输出数据的格式为：

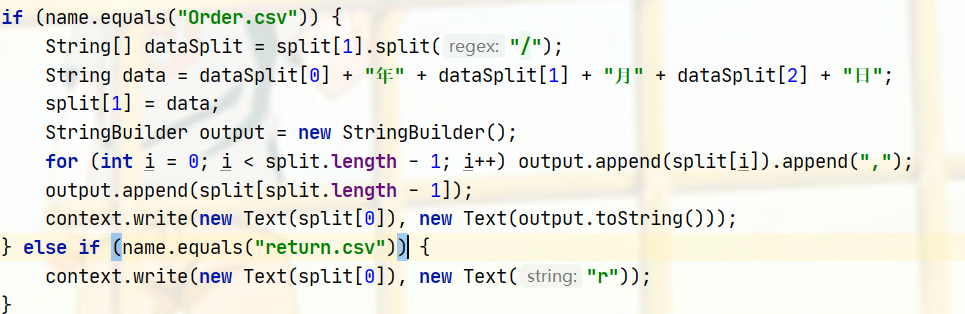
订单ID 订单日期 年 月 日 省/自治区 产品ID 类别 销售额 数量 折扣 利润 是否退货

* + 1. 订单日期进行格式化，派生出 年，月，日 三个字段的代码截图：



* + 1. 订单数据与退货数据进行join的代码截图

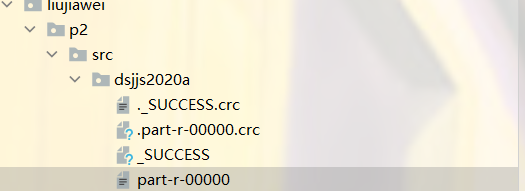
Map



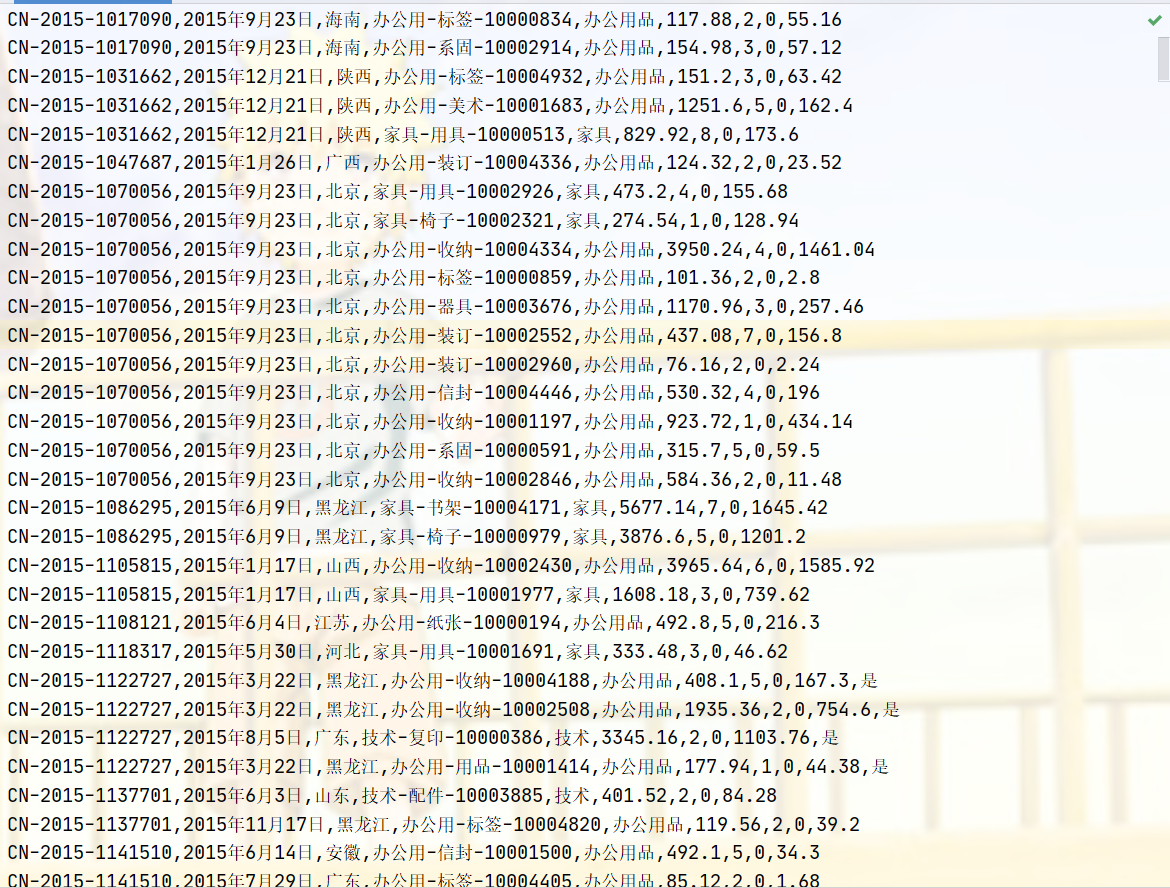
Reducer



* + 1. 输出目录的截图



4）结果截图



### 统计 分组分析（10分）

1. 计算order.csv中 各省的总销售额和总利润 要求销售额降序（5分）

返回格式（省，销售额，利润）

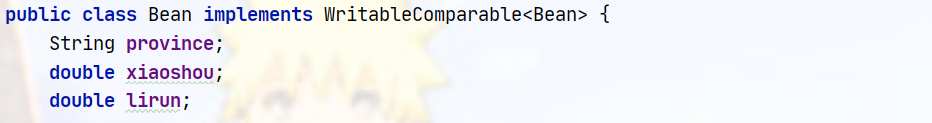
结果截图：



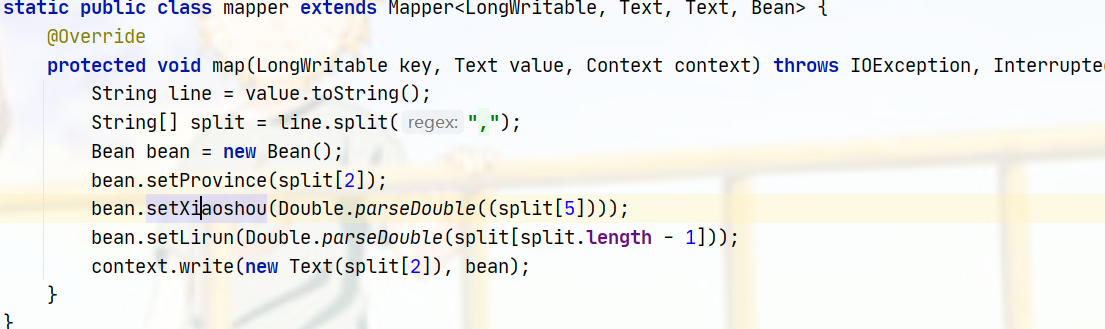
代码截图：

本题使用了一个Bean来包裹输出结果，并实现了WritableComparable以实现排序

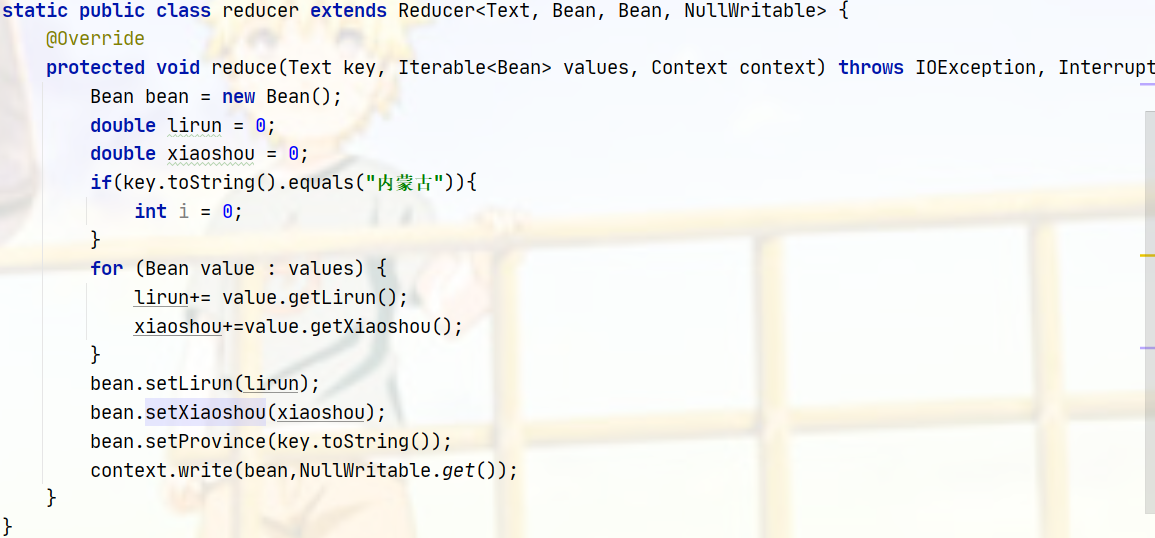
Bean

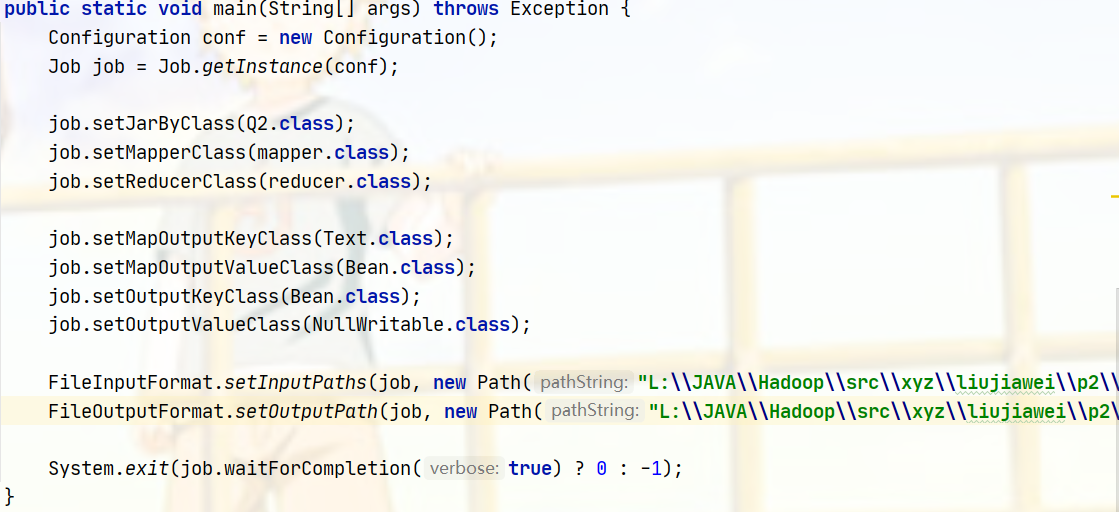
Mapper



Reducer



Main



1. 统计order.csv中 每个类别利润最好的前3条记录，信息（5分）

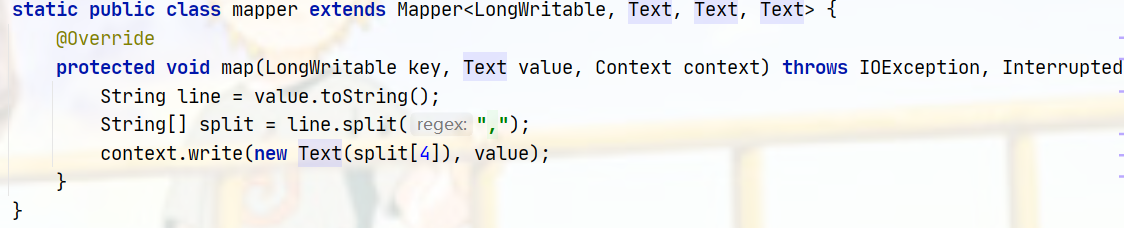
返回格式（订单日期，产品id，类别，销售额，利润）

结果截图：

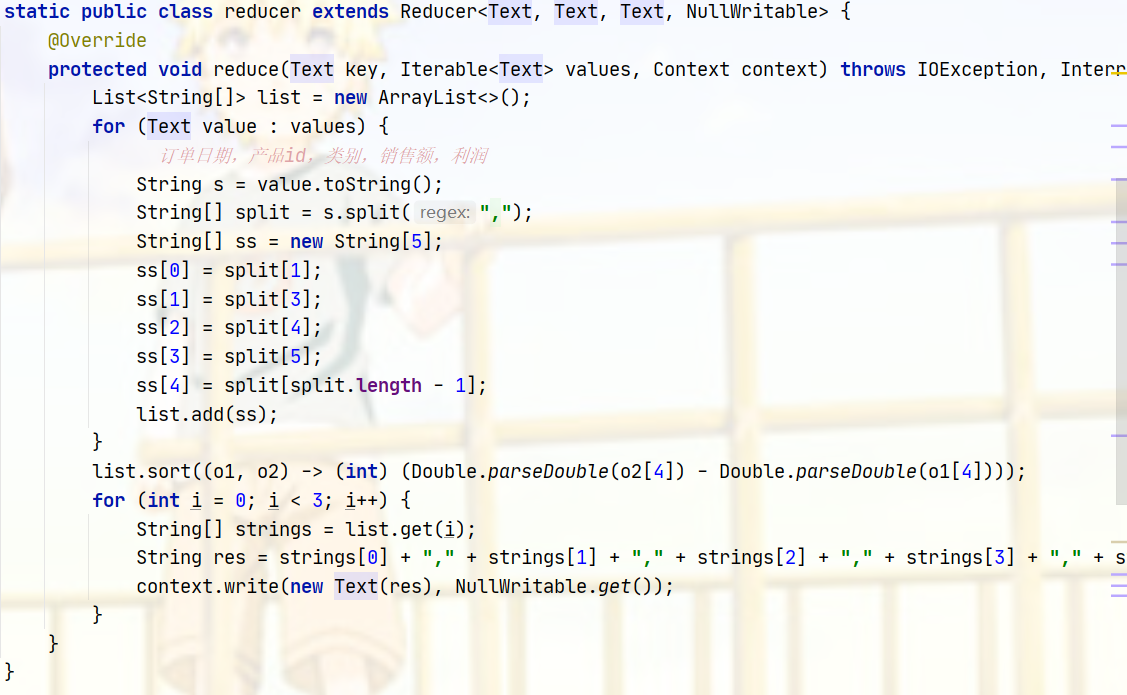


代码截图：

Mapper



Reducer



Main

