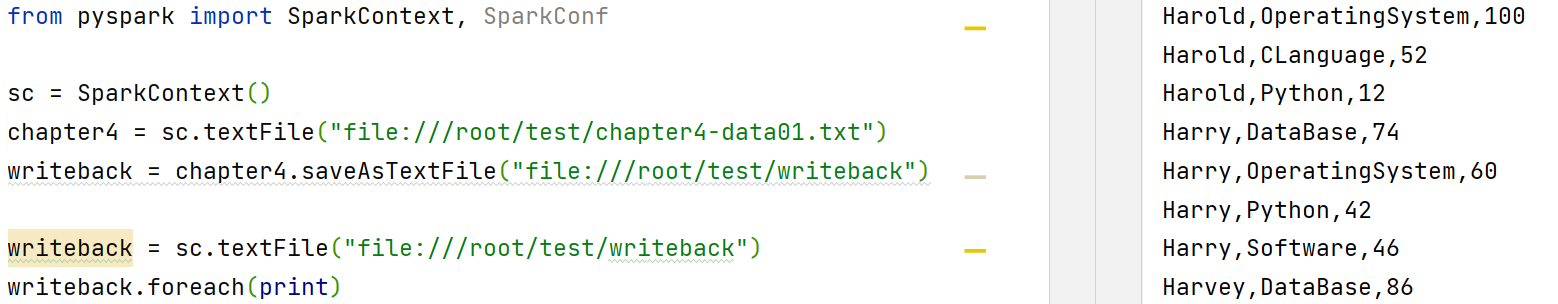
**实验五 RDD与DataFrame**

**【实验目的】**

**1．掌握使用rdd转换与文件存储；**

**2．掌握使用DataFrame的创建和基本操作。**

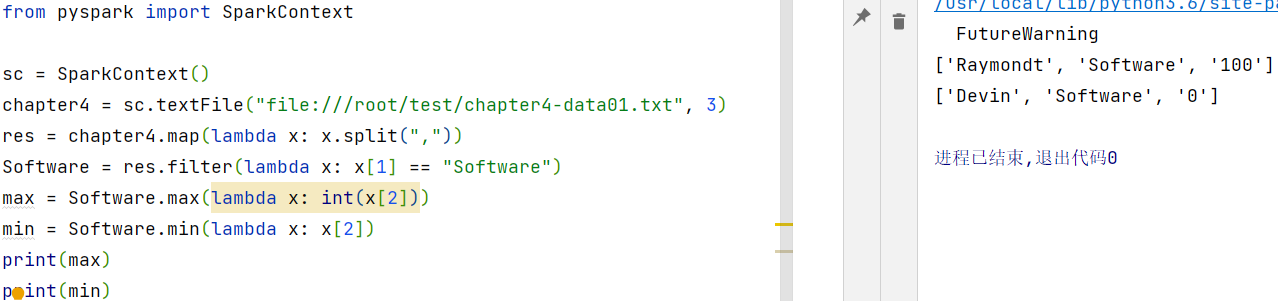
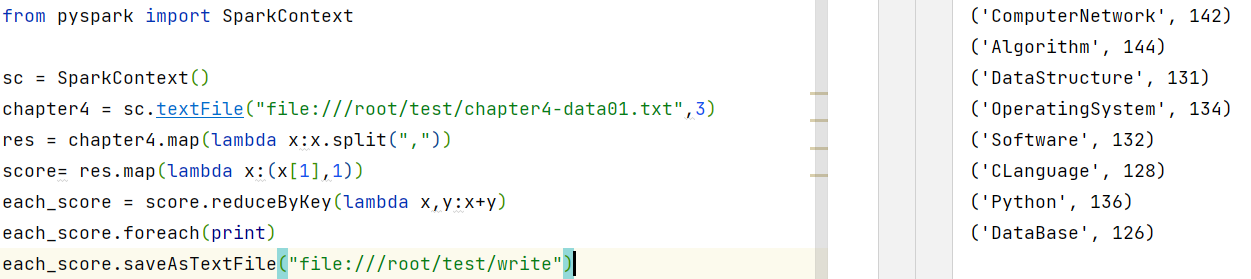
**【实验内容】**

**1.将本地路径/root/test下的文件chapter4-data01.txt**

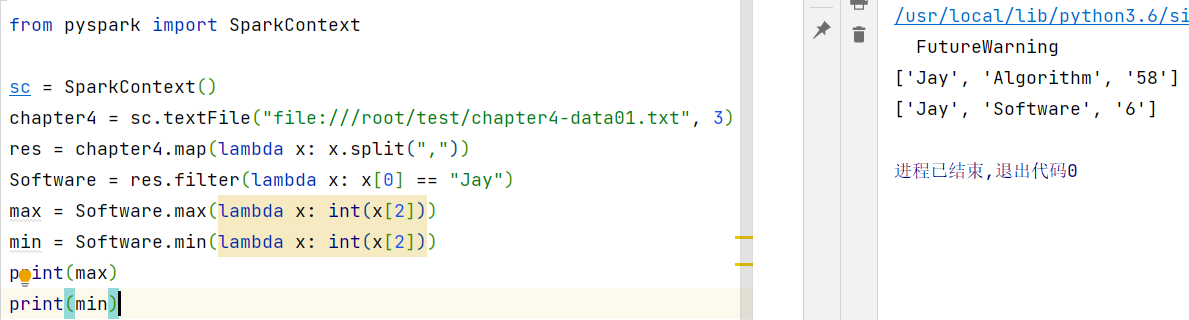
**（1）转换为rdd，并将结果命名为writeback保存到root/test下,显示输出；**

**（2）计算每个人选修了几门课，并将结果命名为read保存到root/test下；**

**（3）计算每门课有几个人选修，并将分区设置为3后，命名为write保存到root/test下；**

**（4）显示选修Software课程的最高分和最低分；**

1. **显示Jay的最高分和最低分；**



**2.课本P113页实验**

**Spark SQL基本操作**

**本地路径/root/test下有文件employee.json。**

|  |
| --- |
| **{ "id":1 , "name":" Ella" , "age":36 }**  **{ "id":2, "name":"Bob","age":29 }**  **{ "id":3 , "name":"Jack","age":29 }**  **{ "id":4 , "name":"Jim","age":28 }**  **{ "id":4 , "name":"Jim","age":28 }**  **{ "id":5 , "name":"Damon" }**  **{ "id":5 , "name":"Damon" }** |

**为employee.json创建DataFrame，并写出Python语句完成下列操作：**

1. **查询所有数据；**
2. **查询所有数据，并去除重复的数据；**
3. **查询所有数据，打印时去除id字段；**
4. **筛选出age>30的记录；**
5. **将数据按age分组；**
6. **将数据按name升序排列；**
7. **取出前3行数据；**
8. **查询所有记录的name列，并为其取别名为username；**
9. **查询年龄age的平均值；**
10. **查询年龄age的最小值。**

