**作业6**

**RDD编程初级实践**

1. **编程题目**

使用Scala语言编写Spark应用程序实现数据去重（需要使用编译打包工具Maven或sbt进行编译打包）。

对于两个输入文件A和B，编写Spark独立应用程序，对两个文件进行合并，并剔除其中重复的内容，得到一个新文件C。下面是输入文件和输出文件的一个样例，供参考。

输入文件A的样例如下：

20180901 x

20180902 y

20180903 x

20180904 y

20180905 z

20180906 z

输入文件B的样例如下：

20180901 y

20180902 y

20180903 x

20180904 z

20180905 y

根据输入的文件A和B合并得到的输出文件C的样例如下：

20180901 x

20180901 y

20180902 y

20180903 x

20180904 y

20180904 z

20180905 y

20180905 z

20180906 z

**二、实验报告**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验报告 | | | | |
| 题目 | RDD编程初级实践 | 姓名 | 吴雨娟 | 日期：  2021.10.28 |
| 实验环境：VMware workstation 16 pro，Ubuntu-20.04操作系统, JDK1.8,scala-2.12.10, Spark-3.1.2,maven-3.8.3,ideaIU-2021.2.2 | | | | |
| 实验内容与完成情况：  **一、编程题目**  使用Scala语言编写Spark应用程序实现数据去重（需要使用编译打包工具Maven或sbt进行编译打包）。  对于两个输入文件A和B，编写Spark独立应用程序，对两个文件进行合并，并剔除其中重复的内容，得到一个新文件C。  1.建立输入文件A和B。        2.编写scala程序。  直接在原来的WordCount.scala程序里修改了，所以程序名还是WordCount。    3.在idea里运行程序。    4.打开C.txt文件查看结果。    5. 使用编译打包工具Maven进行编译打包。    6.运行JAR包。  由于程序中没有编写输出在命令行的代码，所以结果要在C.txt文件中查看。 | | | | |
| 出现的问题：  1.问题一：发现无法使用saveAsTextFile()函数来保存。 | | | | |
| 解决方案（列出遇到的问题和解决办法，列出没有解决的问题）：  1.问题一解决方案：改为在foreach里进行写入文件操作，修改后可以得到正确的结果，修改后的代码如下所示。 | | | | |

（备注：相关实验过程可以截图放入实验报告）