厦门大学林子雨，郑海山,赖永炫 编著

《Spark编程基础（Python版）》

教材配套

机房上机实验指南

实验4 RDD编程初级实践

（版本号：2020年4月版本）

（答案）



主讲教师：林子雨

厦门大学数据库实验室

二零二零年四月

目录

[一、实验目的 1](#_Toc512500631)

[二、实验平台 1](#_Toc512500632)

[三、实验内容和要求 1](#_Toc512500633)

[1．pyspark交互式编程 1](#_Toc512500634)

[2.编写独立应用程序实现数据去重 3](#_Toc512500635)

[3.编写独立应用程序实现求平均值问题 4](#_Toc512500636)

[四、实验报告 6](#_Toc512500637)

[**附录1:任课教师介绍** 6](#_Toc512500638)

[**附录2：课程教材介绍** 7](#_Toc512500639)

[**附录3：高校大数据课程公共服务平台介绍** 8](#_Toc512500640)

厦门大学林子雨,郑海山,赖永炫 编著《Spark编程基础（Python版）》

第4章 RDD编程

教材配套机房上机实验指南

实验3 RDD编程初级实践

（答案）

**主讲教师：林子雨**

E-mail: [ziyulin@xmu.edu.cn](mailto:ziyulin@xmu.edu.cn) 个人主页：<http://dblab.xmu.edu.cn/post/linziyu>

# 一、实验目的

（1）熟悉Spark的RDD基本操作及键值对操作；

（2）熟悉使用RDD编程解决实际具体问题的方法。

# 二、实验平台

操作系统：Ubuntu16.04

Spark版本：2.4.0

Python版本：3.4.3

# 三、实验内容和要求

## 1．pyspark交互式编程

请到教材官网的“下载专区”的“数据集”中下载chapter4-data1.txt，该数据集包含了某大学计算机系的成绩，数据格式如下所示：

|  |
| --- |
| Tom,DataBase,80  Tom,Algorithm,50  Tom,DataStructure,60  Jim,DataBase,90  Jim,Algorithm,60  Jim,DataStructure,80  …… |

请根据给定的实验数据，在pyspark中通过编程来计算以下内容：

**【参考答案】**

1. 该系总共有多少学生；

|  |
| --- |
| >>> lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")  >>> res = lines.map(lambda x:x.split(",")).map(lambda x: x[0]) //获取每行数据的第1列  >>> distinct\_res = res.distinct() //去重操作  >>> distinct\_res.count()//取元素总个数  //265 |

答案为：265人

1. 该系共开设了多少门课程；

|  |
| --- |
| >>> lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")  >>> res = lines.map(lambda x:x.split(",")).map(lambda x:x[1]) //获取每行数据的第2列  >>> distinct\_res = res.distinct()//去重操作  >>> distinct\_res.count()//取元素总个数  //8 |

答案为8门

1. Tom同学的总成绩平均分是多少；

|  |
| --- |
| >>> lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")  >>> res = lines.map(lambda x:x.split(",")).filter(lambda x:x[0]=="Tom") //筛选Tom同学的成绩信息  >>> res.foreach(print)  >>> score = res.map(lambda x:int(x[2])) //提取Tom同学的每门成绩，并转换为int类型  >>> num = res.count() //Tom同学选课门数  >>> sum\_score = score.reduce(lambda x,y:x+y) //Tom同学的总成绩  >>> avg = sum\_score/num // 总成绩/门数=平均分  >>> print(avg)  //30.8 |

Tom同学的平均分为30.8分

1. 求每名同学的选修的课程门数；

|  |
| --- |
| >>> lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")  >>> res = lines.map(lambda x:x.split(",")).map(lambda x:(x[0],1)) //学生每门课程都对应(学生姓名,1)，学生有n门课程则有n个(学生姓名,1)  >>> each\_res = res.reduceByKey(lambda x,y: x+y) //按学生姓名获取每个学生的选课总数  >>> each\_res.foreach(print) |

答案共265行

('Lewis', 4)

('Mike', 3)

('Walter', 4)

('Conrad', 2)

('Borg', 4)

……

1. 该系DataBase课程共有多少人选修；

|  |
| --- |
| >>> lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")  >>> res = lines.map(lambda x:x.split(",")).filter(lambda x:x[1]=="DataBase")  >>> res.count()  //126 |

答案为126人

1. 各门课程的平均分是多少；

|  |
| --- |
| >>> lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")  >>> res = lines.map(lambda x:x.split(",")).map(lambda x:(x[1],(int(x[2]),1))) //为每门课程的分数后面新增一列1，表示1个学生选择了该课程。格式如('ComputerNetwork', (44, 1))  >>> temp = res.reduceByKey(lambda x,y:(x[0]+y[0],x[1]+y[1])) //按课程名聚合课程总分和选课人数。格式如('ComputerNetwork', (7370, 142))  >>> avg = temp.map(lambda x:(x[0], round(x[1][0]/x[1][1],2)))//课程总分/选课人数 = 平均分，并利用round(x,2)保留两位小数  >>> avg.foreach(print) |

答案为：

('ComputerNetwork', 51.9)

('Software', 50.91)

('DataBase', 50.54)

('Algorithm', 48.83)

('OperatingSystem', 54.94)

('Python', 57.82)

('DataStructure', 47.57)

('CLanguage', 50.61)

（7）使用累加器计算共有多少人选了DataBase这门课。

|  |
| --- |
| >>> lines = sc.textFile("file:///usr/local/spark/sparksqldata/Data01.txt")  >>> res = lines.map(lambda x:x.split(",")).filter(lambda x:x[1]=="DataBase")//筛选出选了DataBase课程的数据  >>> accum = sc.accumulator(0) //定义一个从0开始的累加器accum  >>> res.foreach(lambda x:accum.add(1))//遍历res，每扫描一条数据，累加器加1  >>> accum.value //输出累加器的最终值  //126 |

答案：共有126人

## 2.编写独立应用程序实现数据去重

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：

20170101 x

20170102 y

20170103 x

20170104 y

20170105 z

20170106 z

输入文件B的样例如下：

20170101 y

20170102 y

20170103 x

20170104 z

20170105 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：

20170101 x

20170101 y

20170102 y

20170103 x

20170104 y

20170104 z

20170105 y

20170105 z

20170106 z

【参考答案】

实验答案参考步骤如下：

（1）假设当前目录为/usr/local/spark/mycode/remdup，在当前目录下新建一个remdup.py文件，复制下面代码；

|  |
| --- |
| from pyspark import SparkContext  #初始化SparkContext  sc = SparkContext('local','remdup')  #加载两个文件A和B  lines1 = sc.textFile("file:///usr/local/spark/mycode/remdup/A")  lines2 = sc.textFile("file:///usr/local/spark/mycode/remdup/B")  #合并两个文件的内容  lines = lines1.union(lines2)  #去重操作  distinct\_lines = lines.distinct()  #排序操作  res = distinct\_lines.sortBy(lambda x:x)  #将结果写入result文件中，repartition(1)的作用是让结果合并到一个文件中，不加的话会结果写入到两个文件  res.repartition(1).saveAsTextFile("file:///usr/local/spark/mycode/remdup/result") |

（2）最后在目录/usr/local/spark/mycode/remdup下执行下面命令执行程序（注意执行程序时请先退出pyspark shell，否则会出现“地址已在使用”的警告）；

|  |
| --- |
| $ python3 remdup.py |

（3）在目录/usr/local/spark/mycode/remdup/result下即可得到结果文件part-00000。

## 3.编写独立应用程序实现求平均值问题

每个输入文件表示班级学生某个学科的成绩，每行内容由两个字段组成，第一个是学生名字，第二个是学生的成绩；编写Spark独立应用程序求出所有学生的平均成绩，并输出到一个新文件中。下面是输入文件和输出文件的一个样例，供参考。

Algorithm成绩：

小明 92

小红 87

小新 82

小丽 90

Database成绩：

小明 95

小红 81

小新 89

小丽 85

Python成绩：

小明 82

小红 83

小新 94

小丽 91

平均成绩如下：

(小红,83.67)

(小新,88.33)

(小明,89.67)

(小丽,88.67)

**【参考答案】**

实验答案参考步骤如下：

（1）假设当前目录为/usr/local/spark/mycode/avgscore，在当前目录下新建一个avgscore.py，复制下面代码；

|  |
| --- |
| from pyspark import SparkContext  #初始化SparkContext  sc = SparkContext('local',' avgscore')  #加载三个文件Algorithm.txt、Database.txt和Python.txt  lines1 = sc.textFile("file:///usr/local/spark/mycode/avgscore/Algorithm.txt")  lines2 = sc.textFile("file:///usr/local/spark/mycode/avgscore/Database.txt")  lines3 = sc.textFile("file:///usr/local/spark/mycode/avgscore/Python.txt")  #合并三个文件的内容  lines = lines1.union(lines2).union(lines3)  #为每行数据新增一列1，方便后续统计每个学生选修的课程数目。data的数据格式为('小明', (92, 1))  data = lines.map(lambda x:x.split(" ")).map(lambda x:(x[0],(int(x[1]),1)))  #根据key也就是学生姓名合计每门课程的成绩，以及选修的课程数目。res的数据格式为('小明', (269, 3))  res = data.reduceByKey(lambda x,y:(x[0]+y[0],x[1]+y[1]))  #利用总成绩除以选修的课程数来计算每个学生的每门课程的平均分，并利用round(x,2)保留两位小数  result = res.map(lambda x:(x[0],round(x[1][0]/x[1][1],2)))  #将结果写入result文件中，repartition(1)的作用是让结果合并到一个文件中，不加的话会结果写入到三个文件  result.repartition(1).saveAsTextFile("file:///usr/local/spark/mycode/avgscore/result") |

（2）最后在目录/usr/local/spark/mycode/avgscore下执行下面命令执行程序（注意执行程序时请先退出pyspark shell，否则会出现“地址已在使用”的警告）。

|  |
| --- |
| $ python3 avgscore.py |

（3）在目录/usr/local/spark/mycode/avgscore/result下即可得到结果文件part-00000。

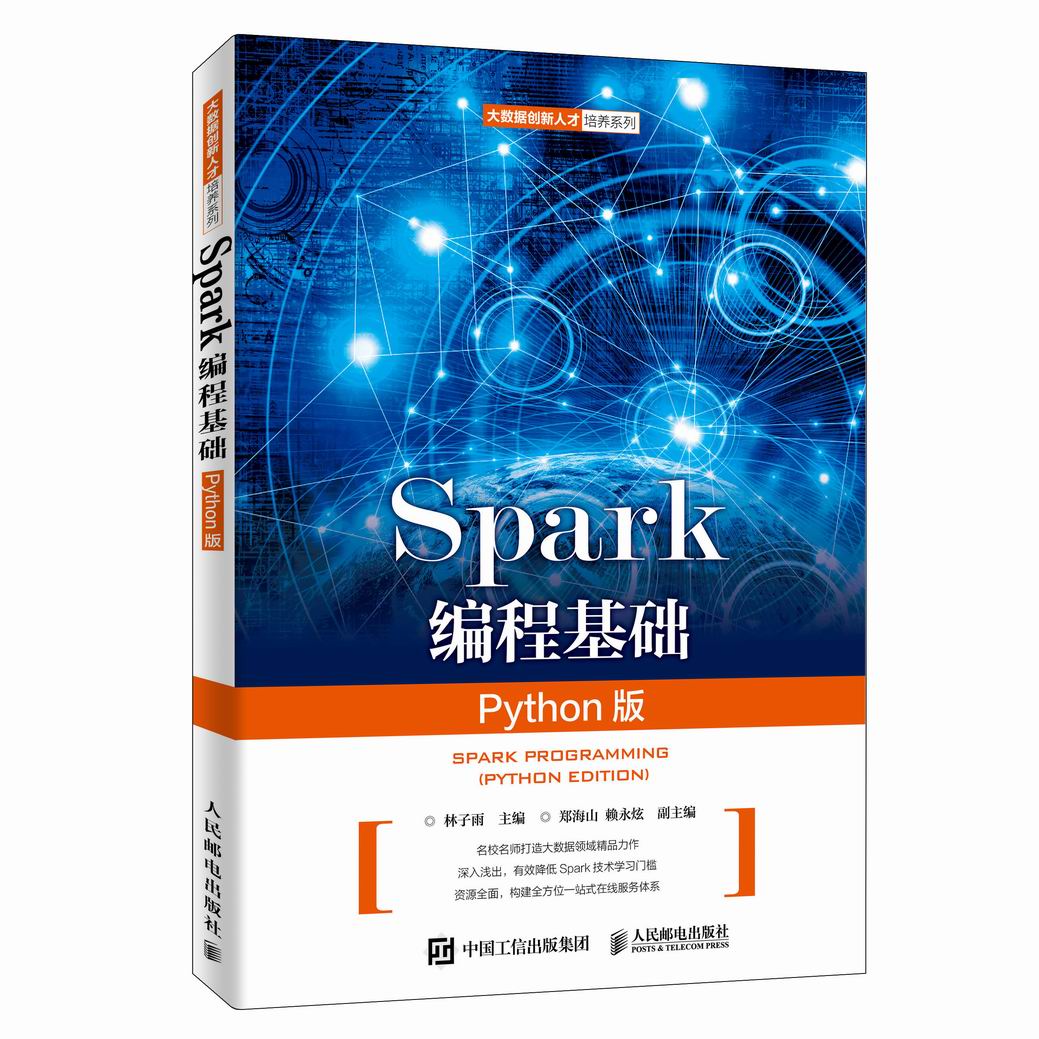
# 四、实验报告

|  |  |  |
| --- | --- | --- |
| 《Spark编程基础（Python版）》实验报告 | | |
| 题目： | 姓名： | 日期： |
| 实验环境： | | |
| 实验内容与完成情况： | | |
| 出现的问题： | | |
| 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）： | | |

**附录1:任课教师介绍**

|  |  |
| --- | --- |
| [http://www.cs.xmu.edu.cn/database/linziyu/images/linziyu2016.jpg](http://dblab.xmu.edu.cn/post/linziyu/) | 林子雨（1978－），男，博士，国内高校知名大数据教师，厦门大学计算机科学系副教授，厦门大学云计算与大数据研究中心创始成员，厦门大学数据库实验室负责人，中国计算机学会数据库专委会委员，中国计算机学会信息系统专委会委员，中国高校首个“数字教师”提出者和建设者。2013年开始在厦门大学开设大数据课程，建设了国内高校首个大数据课程公共服务平台，平台累计网络访问量超过1000万次，成为全国高校大数据教学知名品牌，并荣获“2018年福建省教学成果二等奖”，主持的课程《大数据技术原理与应用》获评“2018年国家精品在线开放课程”。  E-mail: ziyulin@xmu.edu.cn  个人主页：http://dblab.xmu.edu.cn/post/linziyu  数据库实验室网站：http://dblab.xmu.edu.cn |

**附录2：课程教材介绍**



林子雨、郑海山、赖永炫编著《Spark编程基础（Python版）》

人民邮电出版社 ISBN:978-7-115-52439-3 定价：49.80元

厦门大学林子雨、郑海山、赖永炫老师编著《Spark编程基础（Python版）》，以Python作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Structured Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

本书可以作为高等院校计算机、软件工程、数据科学与大数据技术等专业的进阶级大数据课程教材，用于指导Spark编程实践，也可供相关技术人员参考。

欢迎访问《Spark编程基础(Python版)》教材官方网站：http://dblab.xmu.edu.cn/post/spark-python/



扫一扫访问教材官网

**附录3：高校大数据课程公共服务平台介绍**



高校大数据课程公共服务平台，由中国高校首个“数字教师”的提出者和建设者——林子雨老师发起，由厦门大学数据库实验室全力打造，由厦门大学云计算与大数据研究中心、海峡云计算与大数据应用研究中心携手共建。这是国内第一个服务于高校大数据课程建设的公共服务平台，旨在促进国内高校大数据课程体系建设，提高大数据课程教学水平，降低大数据课程学习门槛，提升学生课程学习效果。平台服务对象涵盖高校、教师和学生。平台为高校开设大数据课程提供全流程辅助，为教师开展教学工作提供一站式服务，为学生学习大数据课程提供全方位辅导。平台重点打造“11个1工程”，即1本教材（含官网）、1个教师服务站、1个学生服务站、1个公益项目、1堂巡讲公开课、1个示范班级、1门在线课程、1个交流群（QQ群、微信群）、1个保障团队、1个培训基地和1个实验平台。目前平台每年访问量已经超过200万次，累计访问量超过1000万次，成为国内高校大数据教学知名品牌。

平台主页：http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/



扫一扫访问平台主页