大数据分析第一次作业

201250203 陈张熠

1. 源代码及运行截图（项目结构在运行截图中）

import org.apache.spark.{SparkConf, SparkContext}

object WordCount {

def main(args: Array[String]): Unit ={

val conf = new SparkConf().setAppName("wordCount").setMaster("local")

val sc=new SparkContext(conf)

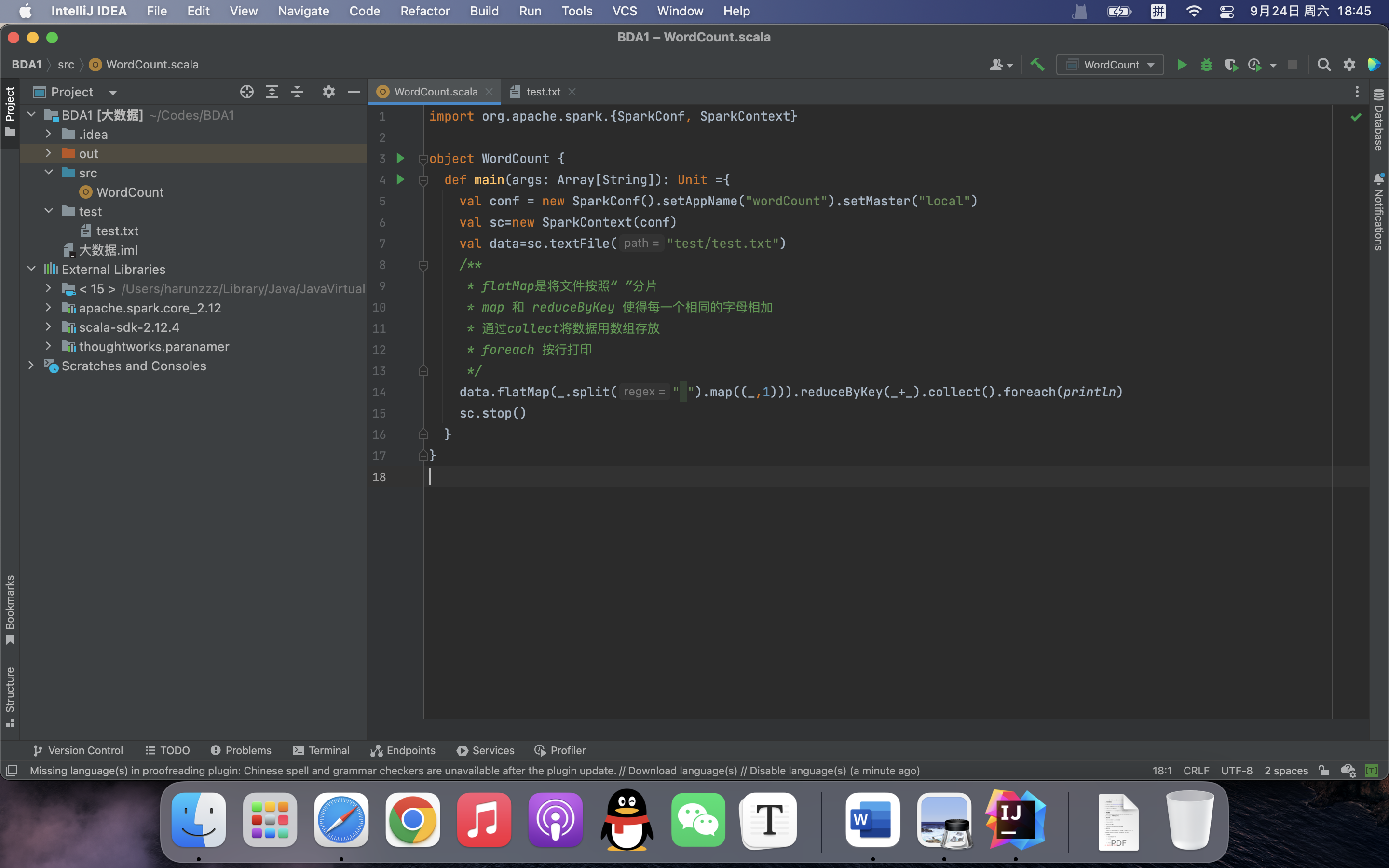
val data=sc.textFile("test/test.txt")

data.flatMap(\_.split(" ").map((\_,1))).reduceByKey(\_+\_).collect().foreach(println)

sc.stop()

}

}



1. 实现思路

新建sparkConf对象，再通过此对象新建SparkContext对象

然后通过map 和 reduceByKey使得每一个相同的字母相加

通过collect将数据用数组存放，foreach 按行打印

1. 实验心得

Scala和spark的似乎向下兼容性不好，导致很容易报错，狂报错导致心态爆炸。

需要对应好版本号。

Scala是完全面向对象的语言，包括所有基本数据类型如int，布尔类型。此外，scala可以与Java进行无缝的连接。

Hadoop采用分布式数据存储架构，而Spark需要采用hdfs的数据存储。

Spark的处理速度更快，但Hadoop的容错性更高。