Statistics

## 1. Statistics

package sougou  
  
import org.apache.hadoop.io.{Text, LongWritable}  
import org.apache.hadoop.mapred.TextInputFormat  
import org.apache.spark.{SparkConf, SparkContext}  
  
/\*\*  
 \* Created by dishui on 2017/9/9.  
 \*/  
object Statistics {  
 def main(args: Array[String]) {  
 if (args.length < 2) {  
 println("args(0): 输入文件路径")  
 println("args(1): 输出文件路径")  
 return  
 }  
 val conf = new SparkConf()  
 .setAppName("sougouStatistics")  
// .setMaster("local[2]")  
 val sc = new SparkContext(conf)  
 // 乱码  
 val resourceRDD = sc.hadoopFile(args(0),classOf[TextInputFormat],classOf[LongWritable], classOf[Text], 1)  
 .map(p => new String(p.\_2.getBytes, 0, p.\_2.getLength, "GB2312").split("\\t"))  
  
 val cacheRDD = resourceRDD.cache()  
  
 val count = cacheRDD.count()  
 sc.makeRDD(Array(count)).saveAsTextFile(args(1) + "/count")  
  
 val timeResult = cacheRDD.map(x => (x(0), 1))  
 .reduceByKey(\_ + \_)  
 .sortBy(x => x.\_2, false)  
 .take(15)  
  
 sc.makeRDD(timeResult).saveAsTextFile(args(1) + "/timeResult")  
  
 val wordsResult = cacheRDD.map(x => (x(2), 1))  
 .reduceByKey(\_ + \_)  
 .sortBy(x => x.\_2, false)  
 .take(10)  
  
 sc.makeRDD(wordsResult).saveAsTextFile(args(1) + "/wordsResult")  
 sc.stop()  
 }  
  
}

结果:

