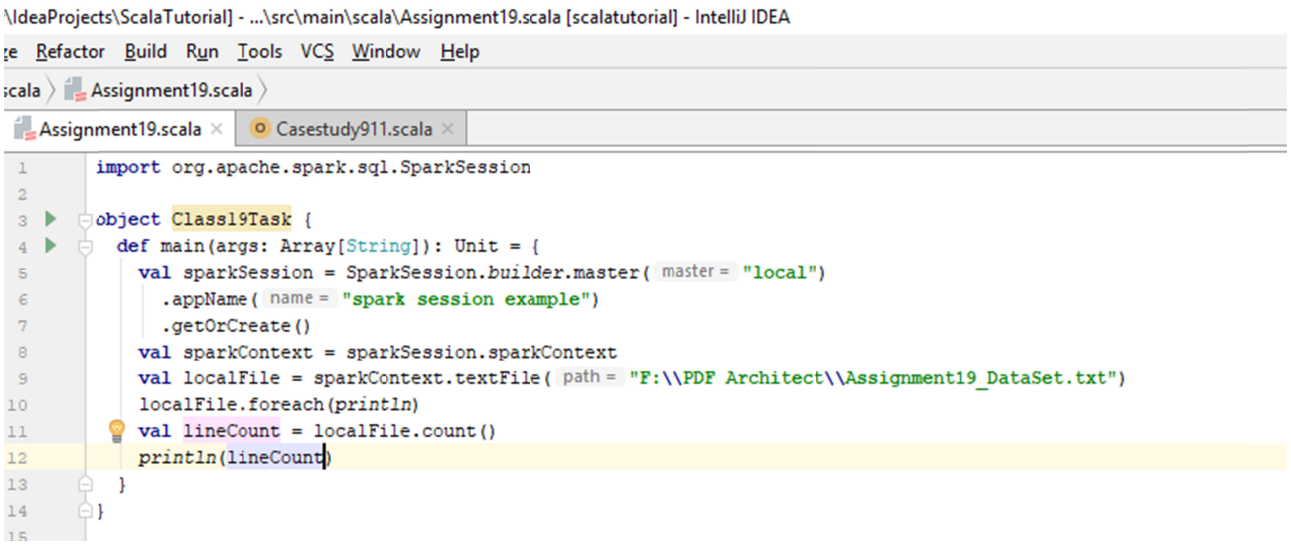
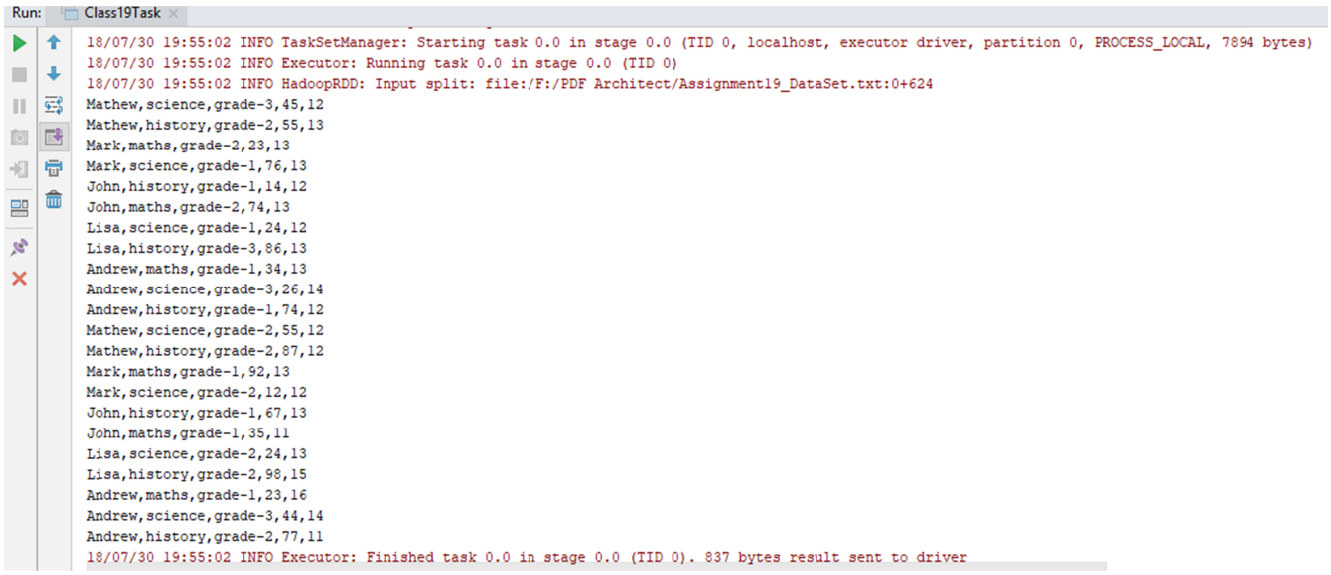
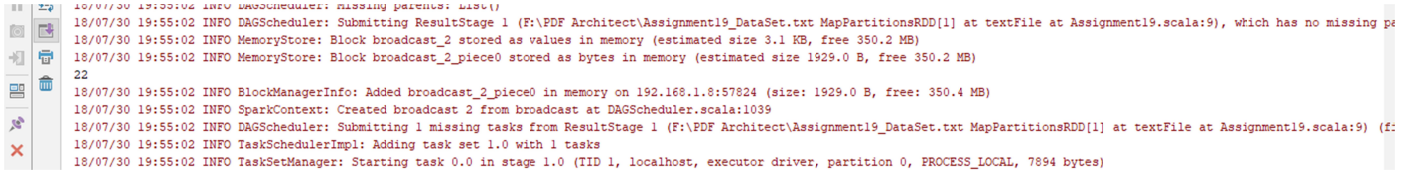
Task 1

1. Write a program to read a text file and print the number of rows of data in the document.







2. Write a program to read a text file and print the number of words in the document.

scala 〉 Assignment19.scala 〉

Assignment19.scala × Casestudy911.scala ×

```scala
import org.apache.spark.sql.SparkSession

object Class19Task {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master( master = "local")
      .appName( name = "spark session example")
      .getOrCreate()
    val sparkContext = sparkSession.sparkContext
    val localFile = sparkContext.textFile( path = "F:\\PDF Architect\\Assignment19_DataSet.txt")
    val words = localFile.flatMap(_.split( regex = ","))
    val wordCount = words.map( x=> (x,1)).reduceByKey(_+_)
    wordCount.foreach( println )
  }
}
```

Run: Class19Task ×

```
18/07/30 20:21:30 INFO ShutdownHookManager: Deleting directory C:\Users\Anupama Stanley\AppData\Local\Temp\spark-a94b698e-3a42-4570-85fe-cf00d0284062
(maths,6)
(14,3)
(67,1)
(98,1)
(15,1)
(Mark,4)
(grade-3,4)
(grade-1,9)
(35,1)
(history,8)
(Andrew,6)
(45,1)
(55,2)
(Mathew,4)
(16,1)
(86,1)
(92,1)
(34,1)
(science,8)
(26,1)
(87,1)
(grade-2,9)
(44,1)
(12,8)
(13,9)
(24,2)
(76,1)
(John,4)
(77,1)
(74,2)
(11,2)
(Lisa,4)
(23,2)
```

Compilation completed successfully in 1 m 28 s 455 ms (4 minutes ago)

3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.

Assignment19.scala | Casestudy911.scala

```scala
import org.apache.spark.sql.SparkSession

object Class19Task {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master( master = "local")
      .appName( name = "spark session example")
      .getOrCreate()
    val sparkContext = sparkSession.sparkContext
    val localFile = sparkContext.textFile( path = "F:\\PDF Architect\\Assignment19_DataSet1.txt")
    localFile.foreach(println)
    val words = localFile.flatMap(_.split( regex = "-"))
    val wordCount = words.map( x=> (x,1)).reduceByKey(_+_)
    wordCount.foreach( println )
  }
}
```

Input File:

```
18/07/30 20:32:29 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor drive
18/07/30 20:32:29 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
18/07/30 20:32:30 INFO HadoopRDD: Input split: file:/F:/PDF Architect/Assignment19_DataSet1.txt:0+140
Mathew-science-grade_3-45-12
Mathew-history-grade_2-55-13
Mark-maths-grade_2-23-13
Mark-science-grade_1-76-13
John-history-grade_1-14-12
18/07/30 20:32:30 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 837 bytes result sent to drive
18/07/30 20:32:30 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 277 ms on localhost (e
18/07/30 20:32:30 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from poo
```

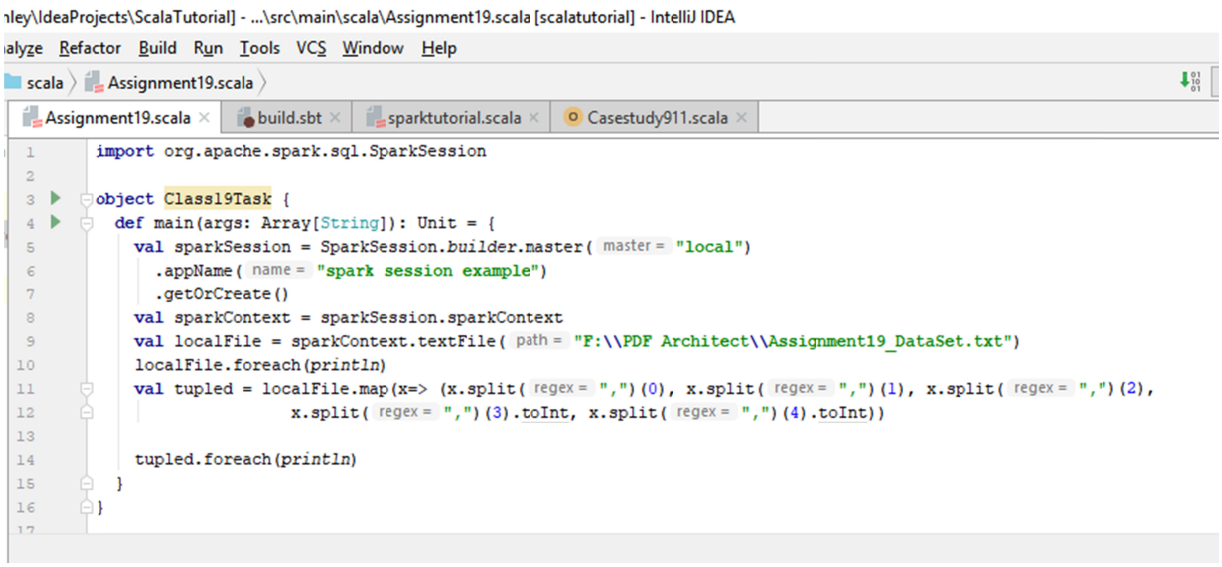
Run: Class19Task

```
18/07/30 20:32:31 INFO BlockManagerInfo: Removed broadcast_2_piece0 on 192.168.1.8:58306 in memory (size: 2.8 KB, free: 350.4 MB)
18/07/30 20:32:31 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 139 ms on localhost (executor driver) (1/1)
18/07/30 20:32:31 INFO DAGScheduler: ResultStage 2 (foreach at Assignment19.scala:13) finished in 0.192 s
(maths,1)
(14,1)
(Mark,2)
(history,2)
(45,1)
(55,1)
(Mathew,2)
(grade_1,2)
(science,2)
(John,1)
(12,2)
(13,3)
(76,1)
(grade_2,2)
(grade_3,1)
(23,1)
18/07/30 20:32:31 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/07/30 20:32:31 INFO DAGScheduler: Job 1 finished: foreach at Assignment19.scala:13, took 0.845802 s
```

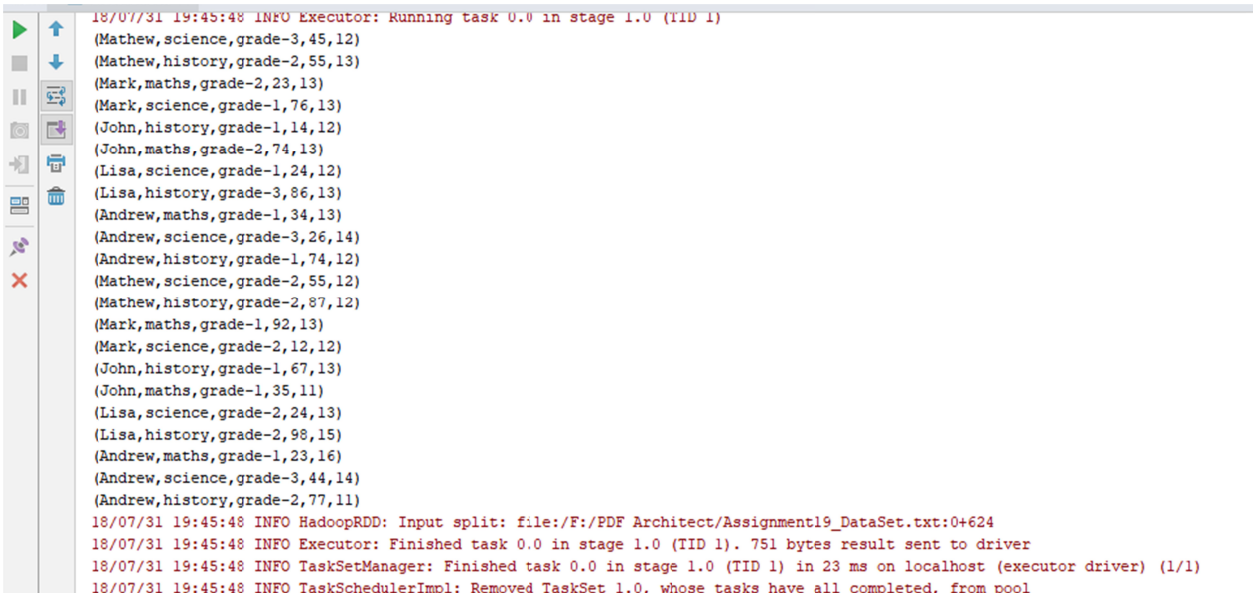
Task 2

Problem Statement 1:

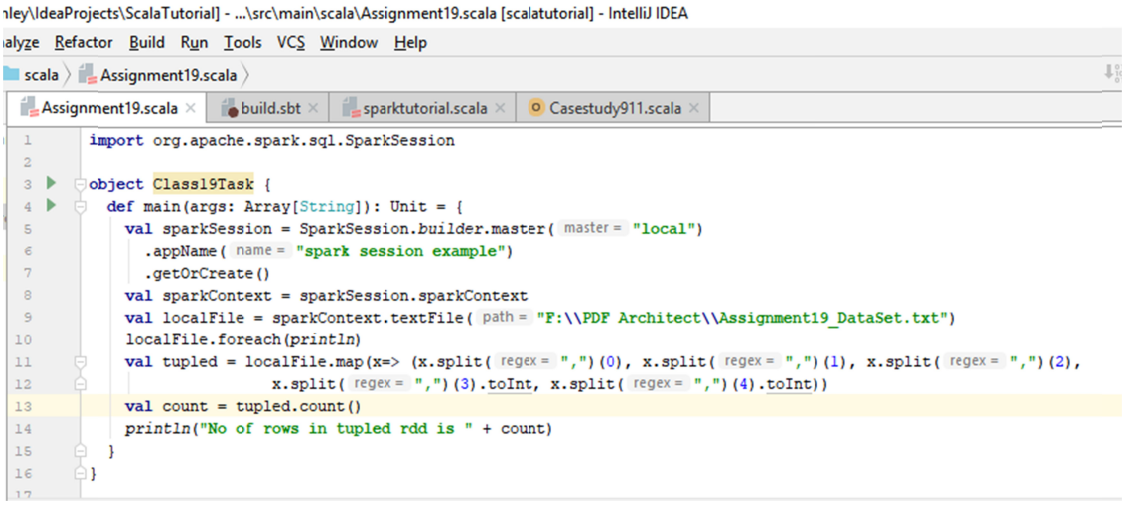
1. Read the text file, and create a tupled rdd.

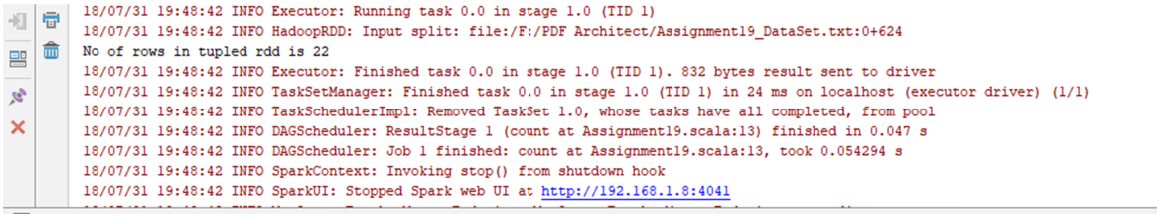


Tupled rdd output

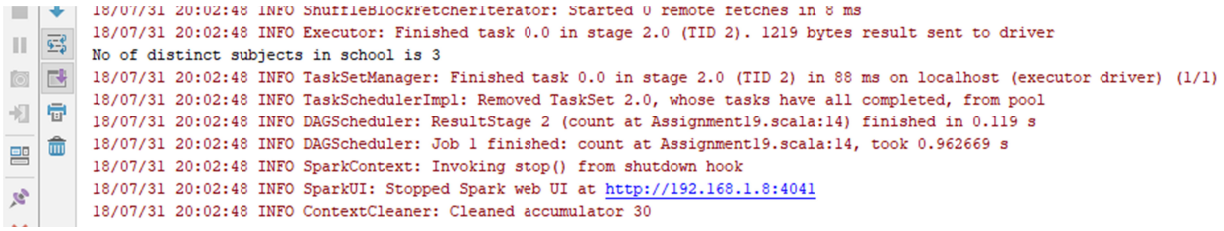


2. Find the count of total number of rows present.





3. What is the distinct number of subjects present in the entire school.



4. What is the count of the number of students in the school, whose name is Mathew and marks, is 55

Assignment19.scala | build.sbt | sparktutorial.scala | Casestudy911.scala

```
import org.apache.spark.sql.SparkSession

object Class19Task {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master( master = "local")
      .appName( name = "spark session example")
      .getOrCreate()
    val sparkContext = sparkSession.sparkContext
    val localFile = sparkContext.textFile( path = "F:\\PDF Architect\\Assignment19_DataSet.txt")

    val students = localFile.map(x=> ((x.split( regex = ",")(0),x.split( regex = ",")(3).toInt ),1))
    val student = students.filter(x => x._1._1 == "Mathew" && x._1._2 == 55)
    val studentagg = student.reduceByKey((x,y)=> x+y).foreach(println)
  }
}
```

```
18/07/31 20:25:43 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/31 20:25:43 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 10 ms
((Mathew,55),2)
18/07/31 20:25:43 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1138 bytes result sent to driver
18/07/31 20:25:43 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 116 ms on localhost (executor driver) (1/1)
18/07/31 20:25:43 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/07/31 20:25:43 INFO DAGScheduler: ResultStage 1 (foreach at Assignment19.scala:13) finished in 0.193 s
18/07/31 20:25:43 INFO DAGScheduler: Job 0 finished: foreach at Assignment19.scala:13, took 1.209210 s
18/07/31 20:25:43 INFO SparkContext: Invoking stop() from shutdown hook
18/07/31 20:25:43 INFO SparkUI: Stopped Spark web UI at http://192.168.1.8:4041
18/07/31 20:25:43 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/07/31 20:25:43 INFO MemoryStore: MemoryStore cleared
```

Problem Statement 2:

1. What is the count of students per grade in the school?

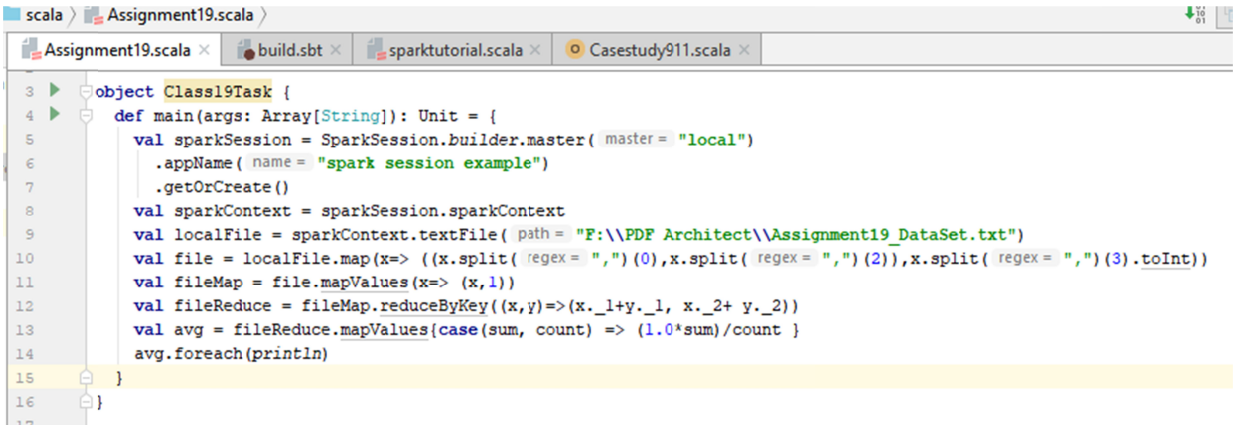
scala > Assignment19.scala

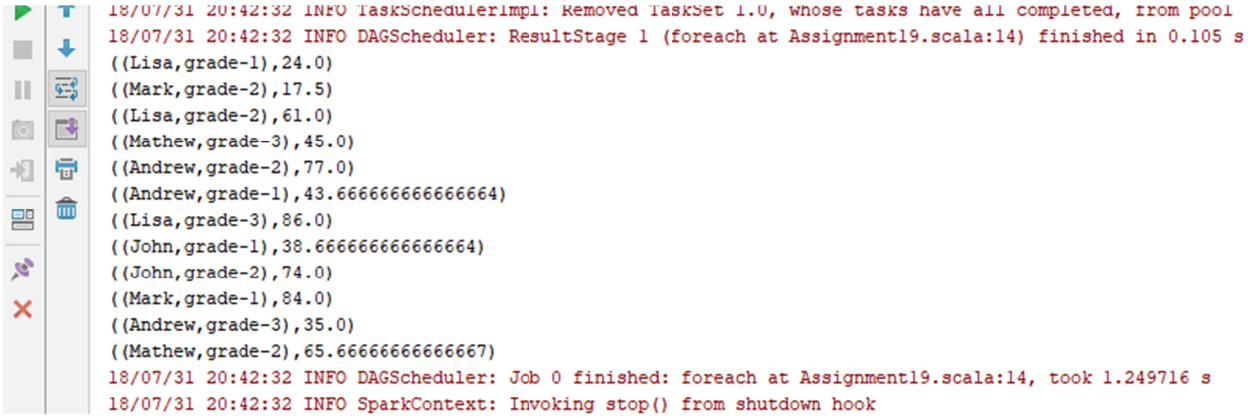
Assignment19.scala | build.sbt | sparktutorial.scala | Casestudy911.scala

```
object Class19Task {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master( master = "local")
      .appName( name = "spark session example")
      .getOrCreate()
    val sparkContext = sparkSession.sparkContext
    val localFile = sparkContext.textFile( path = "F:\\PDF Architect\\Assignment19_DataSet.txt")
    val grade = localFile.map(x=> (x.split( regex = ",")(2),1)).reduceByKey((x,y)=> x+y)
    grade.foreach(println)
  }
}
```

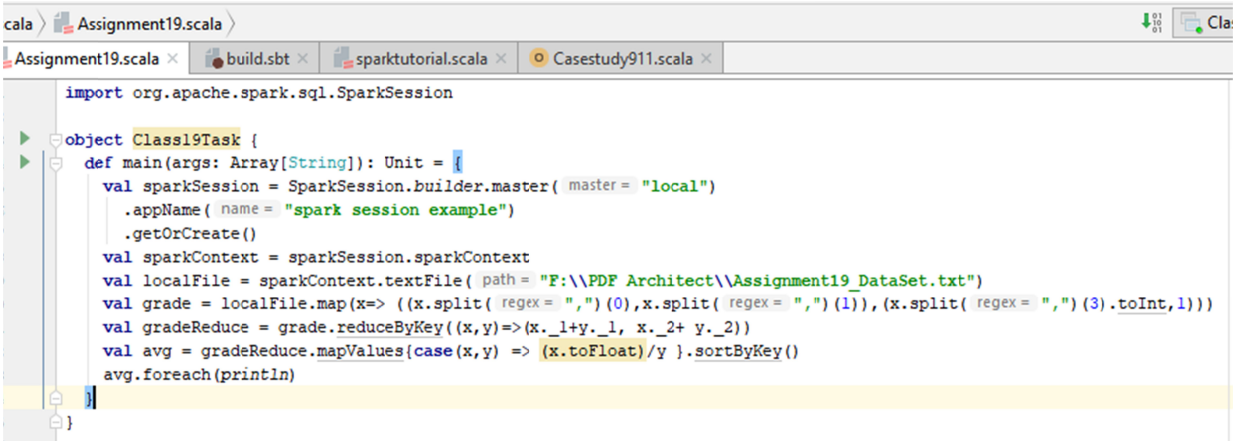
```
18/07/31 20:32:56 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/31 20:32:56 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 8 ms
(grade-3,4)
(grade-1,9)
(grade-2,9)
18/07/31 20:32:56 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1138 bytes result sent to driver
18/07/31 20:32:56 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 76 ms on localhost (exec
18/07/31 20:32:56 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/07/31 20:32:56 INFO DAGScheduler: ResultStage 1 (foreach at Assignment19.scala:11) finished in 0.117 s
```

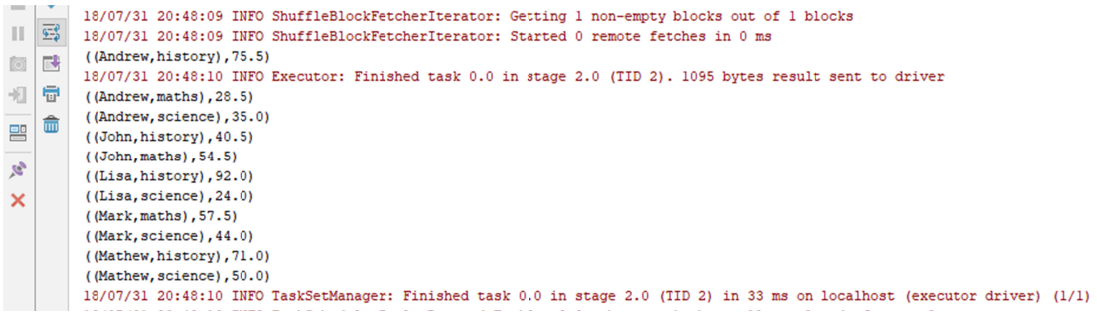
2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)



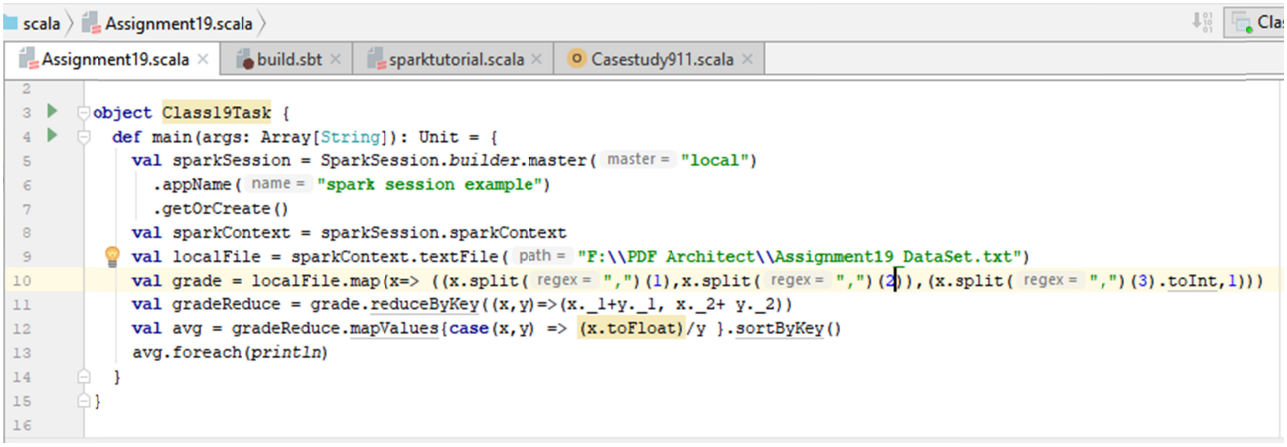


3. What is the average score of students in each subject across all grades?

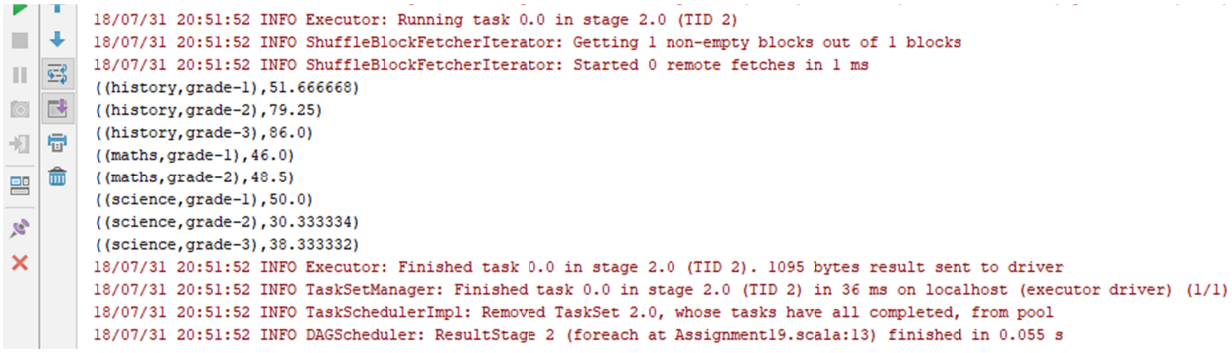




4. What is the average score of students in each subject per grade?

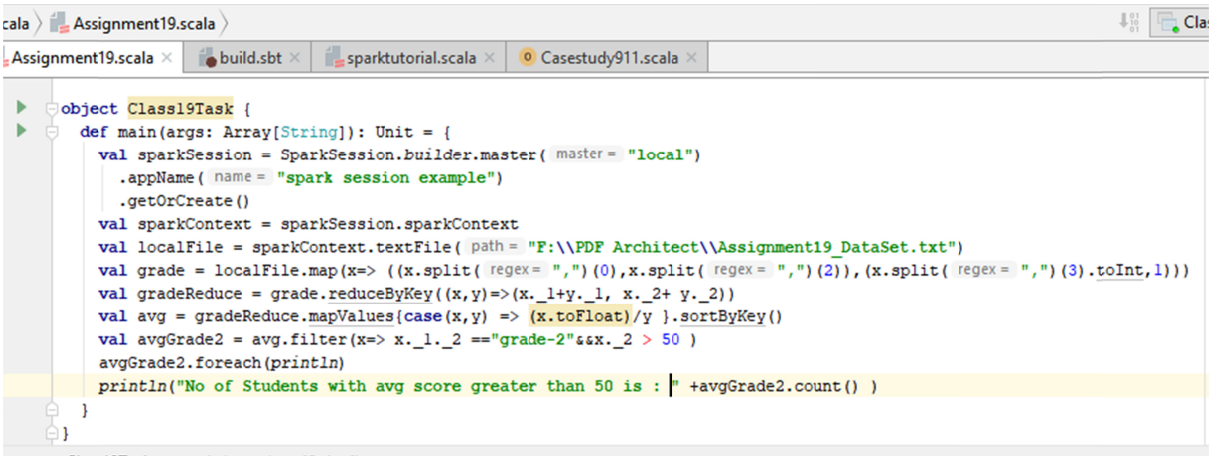


```
object Class19Task {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master( master = "local")
      .appName( name = "spark session example")
      .getOrCreate()
    val sparkContext = sparkSession.sparkContext
    val localFile = sparkContext.textFile( path = "F:\\PDF Architect\\Assignment19_DataSet.txt")
    val grade = localFile.map(x=> ((x.split( regex = ",")(1),x.split( regex = ",")(2)),(x.split( regex = ",")(3).toInt,1)))
    val gradeReduce = grade.reduceByKey((x,y)=>(x._1+y._1, x._2+ y._2))
    val avg = gradeReduce.mapValues{case(x,y) => (x.toFloat)/y }.sortByKey()
    avg.foreach(println)
  }
}
```

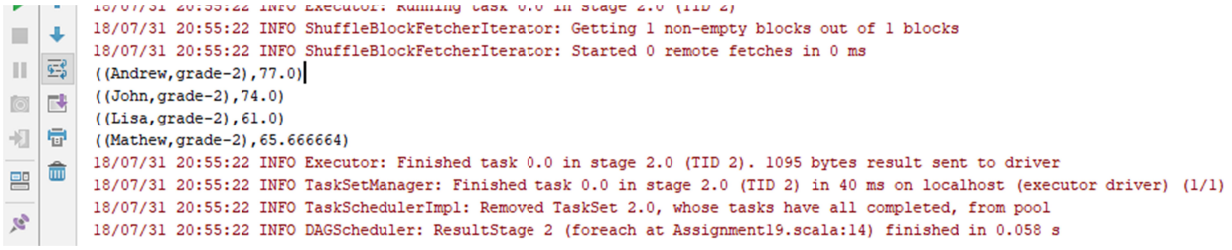


```
18/07/31 20:51:52 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
18/07/31 20:51:52 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/31 20:51:52 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
((history,grade-1),51.666668)
((history,grade-2),79.25)
((history,grade-3),86.0)
((maths,grade-1),46.0)
((maths,grade-2),48.5)
((science,grade-1),50.0)
((science,grade-2),30.333334)
((science,grade-3),38.333332)
18/07/31 20:51:52 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 1095 bytes result sent to driver
18/07/31 20:51:52 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 36 ms on localhost (executor driver) (1/1)
18/07/31 20:51:52 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/07/31 20:51:52 INFO DAGScheduler: ResultStage 2 (foreach at Assignment19.scala:13) finished in 0.055 s
```

5. For all students in grade-2, how many have average score greater than 50?



```
object Class19Task {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master( master = "local")
      .appName( name = "spark session example")
      .getOrCreate()
    val sparkContext = sparkSession.sparkContext
    val localFile = sparkContext.textFile( path = "F:\\PDF Architect\\Assignment19_DataSet.txt")
    val grade = localFile.map(x=> ((x.split( regex = ",")(0),x.split( regex = ",")(2)),(x.split( regex = ",")(3).toInt,1)))
    val gradeReduce = grade.reduceByKey((x,y)=>(x._1+y._1, x._2+ y._2))
    val avg = gradeReduce.mapValues{case(x,y) => (x.toFloat)/y }.sortByKey()
    val avgGrade2 = avg.filter(x=> x._1._2 =="grade-2"&&x._2 > 50 )
    avgGrade2.foreach(println)
    println("No of Students with avg score greater than 50 is : " +avgGrade2.count() )
  }
}
```



```
18/07/31 20:55:22 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/31 20:55:22 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
((Andrew,grade-2),77.0)
((John,grade-2),74.0)
((Lisa,grade-2),61.0)
((Mathew,grade-2),65.666664)
18/07/31 20:55:22 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 1095 bytes result sent to driver
18/07/31 20:55:22 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 40 ms on localhost (executor driver) (1/1)
18/07/31 20:55:22 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/07/31 20:55:22 INFO DAGScheduler: ResultStage 2 (foreach at Assignment19.scala:14) finished in 0.058 s
```

```
18/07/31 21:00:00 INFO Executor: Running task 0.0 in stage 5.0 (TID 3)
18/07/31 21:00:00 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/31 21:00:00 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
No of Students with avg score greater than 50 is : 4
18/07/31 21:00:00 INFO Executor: Finished task 0.0 in stage 5.0 (TID 3). 1176 bytes result sent to driver
18/07/31 21:00:00 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 3) in 29 ms on localhost (executor driver) (1/1)
```

Problem Statement 3:

Are there any students in the college that satisfy the below criteria:

1. Average score per student_name across all grades is same as average score per student_name per grade

Hint - Use Intersection Property

Assignment19.scala | build.sbt | sparktutorial.scala | Casestudy911.scala

```scala
object Class19Task {
  def main(args: Array[String]): Unit = {
    val sparkSession = SparkSession.builder.master( master = "local")
      .appName( name = "spark session example")
      .getOrCreate()
    val sparkContext = sparkSession.sparkContext
    val localFile = sparkContext.textFile( path = "F:\\PDF Architect\\Assignment19_DataSet.txt")
    val fileRDD = localFile.map( x =>x.split( regex = ","))
    val students1 = fileRDD.map( x => ((x(2),x(0)),x(3).toInt))
    val sum1 = students1.distinct.groupByKey().mapValues(x => x.sum)
    sum1.foreach(x => println(x))
    val students2 = fileRDD.map(x=> ((x(2), x(0)),1))
    val count1 = students2.groupByKey().mapValues( x => x.sum )
    count1.foreach( x => println(x))
    val joinedRDD = sum1.join(count1)
    joinedRDD.foreach( x => println(x))
    val avgStd = joinedRDD.map( x => ((x._1),(x._2._1/x._2._2)))
    val avgGrade = avgStd.map( x => (x._1._2, x._2 ))
    val sumPerStd = fileRDD.map( x => ( x(0), x(3).toInt)).groupByKey().mapValues( y => y.sum)
    sumPerStd.foreach( x => println(x))
    val countStd = fileRDD.map( x => (x(0), 1)).groupByKey().mapValues( y => y.sum )
    countStd.foreach( x => println(x))
    val avgStdName = sumPerStd.join(countStd).map( x => (x._1, x._2._1/x._2._2))
    avgStdName.foreach( x => println(x))
    val commonStd = avgStdName.intersection(avgGrade)
    println( "count of Students with same avg per name and grade " + commonStd.count())
    commonStd.foreach( x => println("Students with same avg per name and grade " +x))
  }
}
```

```
18/07/31 23:07:23 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/07/31 23:07:23 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/07/31 23:07:23 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/07/31 23:07:23 INFO Executor: Finished task 0.0 in stage 23.0 (TID 13). 1176 bytes result sent to driver
18/07/31 23:07:23 INFO TaskSetManager: Finished task 0.0 in stage 23.0 (TID 13) in 18 ms on localhost (executor driver) (1/1)
18/07/31 23:07:23 INFO TaskSchedulerImpl: Removed TaskSet 23.0, whose tasks have all completed, from pool
count of Students with same avg per name and grade 0
18/07/31 23:07:23 INFO DAGScheduler: ResultStage 23 (count at Assignment19.scala:28) finished in 0.030 s
18/07/31 23:07:23 INFO DAGScheduler: Job 6 finished: count at Assignment19.scala:28, took 0.202995 s
18/07/31 23:07:23 INFO SparkContext: Starting job: foreach at Assignment19.scala:29
```