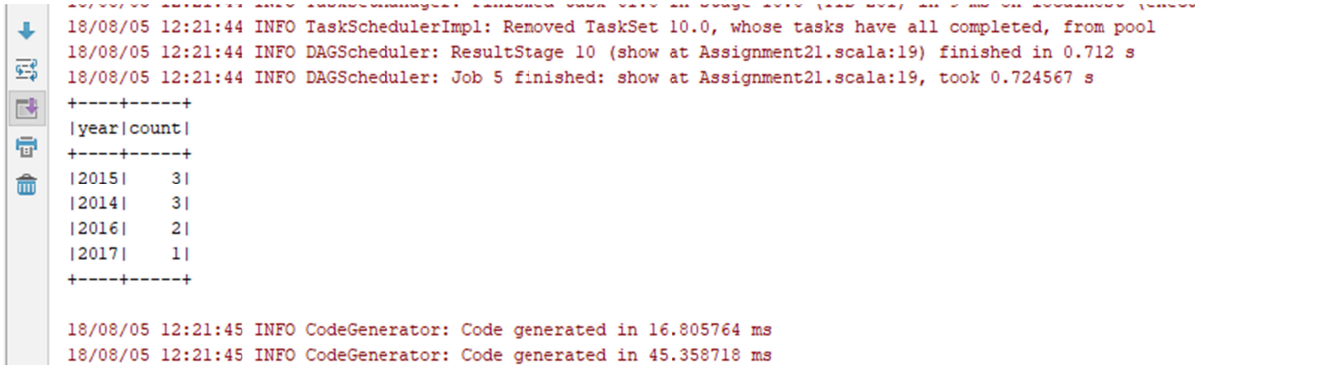
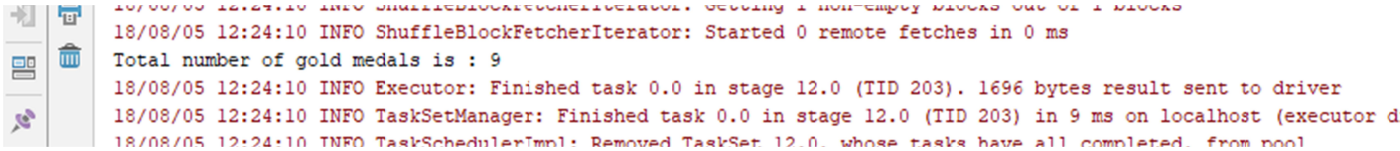
Task 1

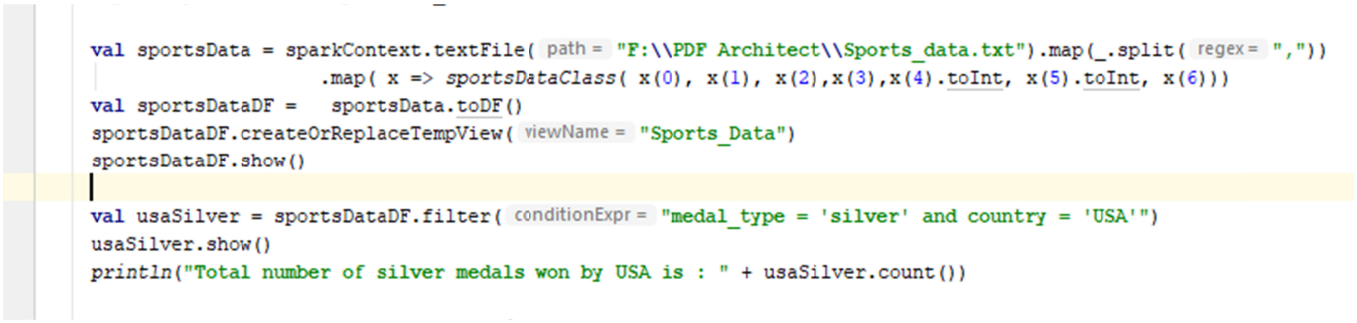
Using spark-sql, Find:

1. What are the total numbers of gold medal winners every year?





2. How many silver medals have been won by USA in each sport?



```
18/08/05 12:35:52 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/08/05 12:35:52 INFO DAGScheduler: ResultStage 1 (show at Assignment21.scala:23) finished in 0.095 s
18/08/05 12:35:52 INFO DAGScheduler: Job 1 finished: show at Assignment21.scala:23, took 0.103350 s
+---------+--------+--------+----------+---+----+-------+
|firstname|lastname|  sports|medal_type|age|year|country|
+---------+--------+--------+----------+---+----+-------+
|  michael|  phelps|swimming|    silver| 32|2016|    USA|
|  michael|  phelps|swimming|    silver| 32|2017|    USA|
|  michael|  phelps|swimming|    silver| 32|2017|    USA|
+---------+--------+--------+----------+---+----+-------+

18/08/05 12:35:53 INFO CodeGenerator: Code generated in 30.0532 ms
18/08/05 12:35:53 INFO CodeGenerator: Code generated in 55.493976 ms
```

```
18/08/05 15:00:00 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 18 ms
18/08/05 15:00:00 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 1825 bytes result sent to driver
Total number of silver medals won by USA is : 3
18/08/05 15:00:00 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 74 ms on localhost (executor drive
18/08/05 15:00:00 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
18/08/05 15:00:00 INFO DAGScheduler: ResultStage 3 (count at Assignment21.scala:19) finished in 0.125 s
```

Task 2

Using udfs on dataframe

1. Change firstname, lastname columns into Mr.first_two_letters_of_firstname<space>lastname for example - michael, phelps becomes Mr.mi phelps

```
val sportsData = sparkContext.textFile( path = "F:\\PDF Architect\\Sports_data.txt").map(_.split( regex = ","))
                    .map( x => sportsDataClass( x(0), x(1), x(2),x(3),x(4).toInt, x(5).toInt, x(6)))
val sportsDataDF =   sportsData.toDF()
sportsDataDF.createOrReplaceTempView( viewName = "Sports_Data")
sportsDataDF.show()

val Name = udf((First_Name: String, Last_Name: String)=> "Mr.".concat(First_Name.substring(0,2).concat( str = " ").concat(Last_Name)))
sportsDataDF.withColumn( colName = "Concat_First_Last",Name( $"firstname",$"lastname"))
        .select( col = "Concat_First_Last", cols = "sports","medal_type", "age","year","country").show()
```

```
18/08/05 13:25:53 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/08/05 13:25:53 INFO DAGScheduler: Job 1 finished: show at Assignment21.scala:20, took 0.170772 s
+-----------------+--------+----------+---+----+-------+
|Concat_First_Last|  sports|medal_type|age|year|country|
+-----------------+--------+----------+---+----+-------+
|     Mr.li cudrow|javellin|      gold| 34|2015|    USA|
|      Mr.ma louis|javellin|      gold| 34|2015|    RUS|
|     Mr.mi phelps|swimming|    silver| 32|2016|    USA|
|         Mr.us pt| running|    silver| 30|2016|    IND|
|   Mr.se williams| running|      gold| 31|2014|    FRA|
|    Mr.ro federer|  tennis|    silver| 32|2016|    CHN|
|        Mr.je cox|swimming|    silver| 32|2014|    IND|
|    Mr.fe johnson|swimming|    silver| 32|2016|    CHN|
|     Mr.li cudrow|javellin|      gold| 34|2017|    USA|
|      Mr.ma louis|javellin|      gold| 34|2015|    RUS|
|     Mr.mi phelps|swimming|    silver| 32|2017|    USA|
|         Mr.us pt| running|    silver| 30|2014|    IND|
|   Mr.se williams| running|      gold| 31|2016|    FRA|
|    Mr.ro federer|  tennis|    silver| 32|2017|    CHN|
|        Mr.je cox|swimming|    silver| 32|2014|    IND|
|    Mr.fe johnson|swimming|    silver| 32|2017|    CHN|
|     Mr.li cudrow|javellin|      gold| 34|2014|    USA|
|      Mr.ma louis|javellin|      gold| 34|2014|    RUS|
|     Mr.mi phelps|swimming|    silver| 32|2017|    USA|
|         Mr.us pt| running|    silver| 30|2014|    IND|
+-----------------+--------+----------+---+----+-------+
only showing top 20 rows

18/08/05 13:25:53 INFO SparkContext: Invoking stop() from shutdown hook
18/08/05 13:25:53 INFO SparkUI: Stopped Spark web UI at http://192.168.1.8:4040
```

2. Add a new column called ranking using udfs on dataframe, where:

gold medalist, with age >= 32 are ranked as pro

gold medalists, with age <= 31 are ranked amateur

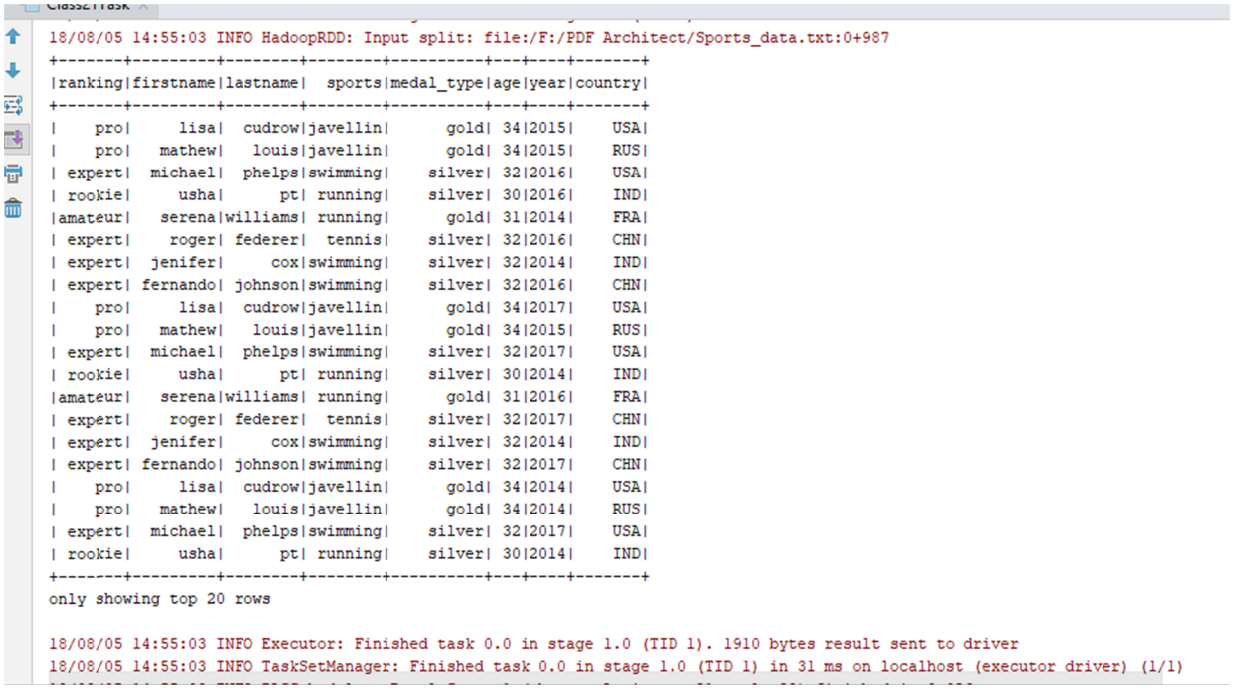
silver medalist, with age >= 32 are ranked as expert

silver medalists, with age <= 31 are ranked rookie

```
val sportsData = sparkContext.textFile( path = "F:\\PDF Architect\\Sports_data.txt").map(_.split( regex = ","))
                  .map( x => sportsDataClass( x(0), x(1), x(2),x(3),x(4).toInt, x(5).toInt, x(6)))
val sportsDataDF =   sportsData.toDF()
sportsDataDF.createOrReplaceTempView( viewName = "Sports_Data")
sportsDataDF.show()

def Rank( Age:Int, Medal: String):String= (Age, Medal) match{
    case( Age, Medal) if Medal == "gold" && Age >= 32  => "pro"
    case( Age, Medal) if Medal  == "gold" && Age <= 31 => "amateur"
    case( Age, Medal) if Medal == "silver" &&  Age >= 32 => "expert"
    case( Age, Medal) if Medal == "silver" && Age <= 31 => "rookie"
    case _ => " "
}

val Ranks = udf( Rank(_:Int,_:String))
sportsDataDF.withColumn( colName = "Ranking",Ranks($"age",$"medal_type"))
      .select( col = "ranking", cols = "firstname", "lastname", "sports", "medal_type", "age", "year", "country").show()
```



```
18/08/05 14:55:03 INFO HadoopRDD: Input split: file:/F:/PDF Architect/Sports_data.txt:0+987
+-------+---------+--------+--------+----------+---+----+-------+
|ranking|firstname|lastname|  sports|medal_type|age|year|country|
+-------+---------+--------+--------+----------+---+----+-------+
|    pro|     lisa|  cudrow|javellin|      gold| 34|2015|    USA|
|    pro|   mathew|   louis|javellin|      gold| 34|2015|    RUS|
| expert|  michael|  phelps|swimming|    silver| 32|2016|    USA|
| rookie|     usha|      pt| running|    silver| 30|2016|    IND|
|amateur|   serena|williams| running|      gold| 31|2014|    FRA|
| expert|    roger| federer|  tennis|    silver| 32|2016|    CHN|
| expert|  jenifer|     cox|swimming|    silver| 32|2014|    IND|
| expert| fernando| johnson|swimming|    silver| 32|2016|    CHN|
|    pro|     lisa|  cudrow|javellin|      gold| 34|2017|    USA|
|    pro|   mathew|   louis|javellin|      gold| 34|2015|    RUS|
| expert|  michael|  phelps|swimming|    silver| 32|2017|    USA|
| rookie|     usha|      pt| running|    silver| 30|2014|    IND|
|amateur|   serena|williams| running|      gold| 31|2016|    FRA|
| expert|    roger| federer|  tennis|    silver| 32|2017|    CHN|
| expert|  jenifer|     cox|swimming|    silver| 32|2014|    IND|
| expert| fernando| johnson|swimming|    silver| 32|2017|    CHN|
|    pro|     lisa|  cudrow|javellin|      gold| 34|2014|    USA|
|    pro|   mathew|   louis|javellin|      gold| 34|2014|    RUS|
| expert|  michael|  phelps|swimming|    silver| 32|2017|    USA|
| rookie|     usha|      pt| running|    silver| 30|2014|    IND|
+-------+---------+--------+--------+----------+---+----+-------+
only showing top 20 rows

18/08/05 14:55:03 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1910 bytes result sent to driver
18/08/05 14:55:03 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 31 ms on localhost (executor driver) (1/1)
```