Session 19

Assignment 2

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 19 – Assignment 2 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 19 – Assignment 2** |
|  |  |
| **Start Date:** | 18/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc496186126)

[Change History 3](#_Toc496186127)

[1. Problem Statement 4](#_Toc496186128)

[2. Datasets 4](#_Toc496186129)

[3. Solution 5](#_Toc496186130)

[4. Results 7](#_Toc496186131)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 19/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

Using UDFs on DataFrame:

* Change firstname and lastname columns into:

Mr.first\_two\_letters\_of\_firstname<space>lastname, Example - michael, phelps becomes Mr.mi phelps

* 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

Using the datasets below:

# Datasets

**Sports\_data.txt**

firstname,lastname,sports,medal\_type,age,year,country

lisa,cudrow,javellin,gold,34,2015,USA

mathew,louis,javellin,gold,34,2015,RUS

michael,phelps,swimming,silver,32,2016,USA

usha,pt,running,silver,30,2016,IND

serena,williams,running,gold,31,2014,FRA

roger,federer,tennis,silver,32,2016,CHN

jenifer,cox,swimming,silver,32,2014,IND

fernando,johnson,swimming,silver,32,2016,CHN

lisa,cudrow,javellin,gold,34,2017,USA

mathew,louis,javellin,gold,34,2015,RUS

michael,phelps,swimming,silver,32,2017,USA

usha,pt,running,silver,30,2014,IND

serena,williams,running,gold,31,2016,FRA

roger,federer,tennis,silver,32,2017,CHN

jenifer,cox,swimming,silver,32,2014,IND

fernando,johnson,swimming,silver,32,2017,CHN

lisa,cudrow,javellin,gold,34,2014,USA

mathew,louis,javellin,gold,34,2014,RUS

michael,phelps,swimming,silver,32,2017,USA

usha,pt,running,silver,30,2014,IND

serena,williams,running,gold,31,2016,FRA

roger,federer,tennis,silver,32,2014,CHN

jenifer,cox,swimming,silver,32,2017,IND

fernando,johnson,swimming,silver,32,2017,CHN

# Solution

I have deliberately commented and printed out all steps for enhanced learning

**Code written**

**package** com.duncb.spark

**package** com.duncb.spark

**import** org.apache.spark.\_

**import** org.apache.spark.SparkContext.\_

**import** org.apache.spark.sql.\_

**import** org.apache.log4j.\_

**import** org.apache.spark.sql.functions.\_

**object** SparkSQL {

**case** **class** **Person**(firstname:*String*, lastname:*String*, sports:*String*,medal\_type:*String*,age:Int,year:Int,country:*String*)

**def** mapper(line:*String*): **Person** = {

**val** fields = line.split(',')

**val** person:**Person** = **Person**(fields(0).toString, fields(1).toString, fields(2).toString,fields(3).toString, fields(4).toInt,fields(5).toInt,fields(6).toString)

**return** person

}

/\*\* Our main function where the action happens \*/

**def** main(args: Array[*String*]) {

// Set the log level to only print errors

Logger.getLogger("org").setLevel(Level.ERROR)

// Use new SparkSession interface in Spark 2.0

**val** spark = SparkSession

.builder

.appName("SparkSQL")

.master("local[\*]")

.config("spark.sql.warehouse.dir", "file:///C:/temp") // Necessary to work around a Windows bug in Spark 2.0.0; omit if you're not on Windows.

.getOrCreate()

**val** lines = spark.sparkContext.textFile("file:///N:/Datasets/Sports\_data.txt")

**val** people = lines.map(mapper)

// Infer the schema, and register the DataSet as a table.

**import** spark.implicits.\_

**val** schemaPeople = people.toDS()

schemaPeople.printSchema

schemaPeople.createOrReplaceTempView("people")

// SQL can be run over DataFrames that have been registered as a table

**val** olmp = spark.sql("SELECT \* FROM people" )

**def** cn = udf((firstname:*String*,lastname:*String*) => {

"Mr."+firstname + " " + lastname

})

**val** result = olmp.withColumn("fullname",cn(olmp("firstname").substr(1,2),olmp("lastname"))).drop(olmp("firstname")).drop(olmp("lastname"))

println(" assignment 19.2 problem 1")

result.show()

println("Table with names changed")

println("------------------------------------------------------------")

**def** rank1 =udf((medal\_type: *String*, age:Int)=> {

**if** (medal\_type =="gold" && age>=32) "Pro"

**else** **if** (medal\_type =="gold" && age<=32) "Amateur"

**else** **if** (medal\_type =="silver" && age>=32) "Expert"

**else** **if** (medal\_type =="silver" && age<=32) "Rookie"

**else** ""

})

**val** ranking= olmp.withColumn("Ranking",rank1(olmp("medal\_type"),olmp("age")))

println(" assignment 19.2 problem 2")

ranking.show()

println("Table with Ranking added")

println("------------------------------------------------------------")

spark.stop()

}

}

# Results

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

root

|-- firstname: string (nullable = true)

|-- lastname: string (nullable = true)

|-- sports: string (nullable = true)

|-- medal\_type: string (nullable = true)

|-- age: integer (nullable = false)

|-- year: integer (nullable = false)

|-- country: string (nullable = true)

+--------+----------+---+----+-------+--------------+

| sports|medal\_type|age|year|country| fullname|

+--------+----------+---+----+-------+--------------+

|javellin| gold| 34|2015| USA| Mr.li cudrow|

|javellin| gold| 34|2015| RUS| Mr.ma louis|

|swimming| silver| 32|2016| USA| Mr.mi phelps|

| running| silver| 30|2016| IND| Mr.us pt|

| running| gold| 31|2014| FRA|Mr.se williams|

| tennis| silver| 32|2016| CHN| Mr.ro federer|

|swimming| silver| 32|2014| IND| Mr.je cox|

|swimming| silver| 32|2016| CHN| Mr.fe johnson|

|javellin| gold| 34|2017| USA| Mr.li cudrow|

|javellin| gold| 34|2015| RUS| Mr.ma louis|

|swimming| silver| 32|2017| USA| Mr.mi phelps|

| running| silver| 30|2014| IND| Mr.us pt|

| running| gold| 31|2016| FRA|Mr.se williams|

| tennis| silver| 32|2017| CHN| Mr.ro federer|

|swimming| silver| 32|2014| IND| Mr.je cox|

|swimming| silver| 32|2017| CHN| Mr.fe johnson|

|javellin| gold| 34|2014| USA| Mr.li cudrow|

|javellin| gold| 34|2014| RUS| Mr.ma louis|

|swimming| silver| 32|2017| USA| Mr.mi phelps|

| running| silver| 30|2014| IND| Mr.us pt|

+--------+----------+---+----+-------+--------------+

only showing top 20 rows

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

root

|-- firstname: string (nullable = true)

|-- lastname: string (nullable = true)

|-- sports: string (nullable = true)

|-- medal\_type: string (nullable = true)

|-- age: integer (nullable = false)

|-- year: integer (nullable = false)

|-- country: string (nullable = true)

assignment 19.2 problem 1

+--------+----------+---+----+-------+--------------+

| sports|medal\_type|age|year|country| fullname|

+--------+----------+---+----+-------+--------------+

|javellin| gold| 34|2015| USA| Mr.li cudrow|

|javellin| gold| 34|2015| RUS| Mr.ma louis|

|swimming| silver| 32|2016| USA| Mr.mi phelps|

| running| silver| 30|2016| IND| Mr.us pt|

| running| gold| 31|2014| FRA|Mr.se williams|

| tennis| silver| 32|2016| CHN| Mr.ro federer|

|swimming| silver| 32|2014| IND| Mr.je cox|

|swimming| silver| 32|2016| CHN| Mr.fe johnson|

|javellin| gold| 34|2017| USA| Mr.li cudrow|

|javellin| gold| 34|2015| RUS| Mr.ma louis|

|swimming| silver| 32|2017| USA| Mr.mi phelps|

| running| silver| 30|2014| IND| Mr.us pt|

| running| gold| 31|2016| FRA|Mr.se williams|

| tennis| silver| 32|2017| CHN| Mr.ro federer|

|swimming| silver| 32|2014| IND| Mr.je cox|

|swimming| silver| 32|2017| CHN| Mr.fe johnson|

|javellin| gold| 34|2014| USA| Mr.li cudrow|

|javellin| gold| 34|2014| RUS| Mr.ma louis|

|swimming| silver| 32|2017| USA| Mr.mi phelps|

| running| silver| 30|2014| IND| Mr.us pt|

+--------+----------+---+----+-------+--------------+

only showing top 20 rows

Table with names changed

------------------------------------------------------------

assignment 19.2 problem 2

+---------+--------+--------+----------+---+----+-------+-------+

|firstname|lastname| sports|medal\_type|age|year|country|Ranking|

+---------+--------+--------+----------+---+----+-------+-------+

| lisa| cudrow|javellin| gold| 34|2015| USA| Pro|

| mathew| louis|javellin| gold| 34|2015| RUS| Pro|

| michael| phelps|swimming| silver| 32|2016| USA| Expert|

| usha| pt| running| silver| 30|2016| IND| Rookie|

| serena|williams| running| gold| 31|2014| FRA|Amateur|

| roger| federer| tennis| silver| 32|2016| CHN| Expert|

| jenifer| cox|swimming| silver| 32|2014| IND| Expert|

| fernando| johnson|swimming| silver| 32|2016| CHN| Expert|

| lisa| cudrow|javellin| gold| 34|2017| USA| Pro|

| mathew| louis|javellin| gold| 34|2015| RUS| Pro|

| michael| phelps|swimming| silver| 32|2017| USA| Expert|

| usha| pt| running| silver| 30|2014| IND| Rookie|

| serena|williams| running| gold| 31|2016| FRA|Amateur|

| roger| federer| tennis| silver| 32|2017| CHN| Expert|

| jenifer| cox|swimming| silver| 32|2014| IND| Expert|

| fernando| johnson|swimming| silver| 32|2017| CHN| Expert|

| lisa| cudrow|javellin| gold| 34|2014| USA| Pro|

| mathew| louis|javellin| gold| 34|2014| RUS| Pro|

| michael| phelps|swimming| silver| 32|2017| USA| Expert|

| usha| pt| running| silver| 30|2014| IND| Rookie|

+---------+--------+--------+----------+---+----+-------+-------+

only showing top 20 rows

Table with Ranking added

------------------------------------------------------------