Assignment19.2

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.

scala> val rdd=sc.textFile("Sports_data.txt")
rdd: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at
<console>:27

scala> val head=rdd.first()
head: String = firstname,lastname,sports,medal_type,age,year,country

scala> val filter_set=rdd.filter(x=>x!=head)
filter_set: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at filter at
<console>:31

scala> val sports=filter_set.map(x=>x.split(",")).map(arrays
=>(arrays(0),arrays(1),arrays(2),arrays(3),arrays(4),arrays(5),arrays(6))).toDF("fir stname","lastname","sports","medal_type","age","year","country")
sports: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string,
```

sports: string, medal_type: string, age: string, year: string, country: string]

scala> sports.registerTempTable("sportstable")

```
scala> sports.show
   rstname|lastname|
                         sports|medal type|age|year|country
                                                             USAI
              cudrow|javellin|
                                        gold| 34|2015|
               louis|javellin|
    mathew|
                                               34 | 2015 |
                                                             RUS
   michael|
              phelps|swimming|
                                      silver
                                               32 | 2016 |
                                                             USA
                                      silver|
                                               30 | 2016 |
                                                             IND
      ushal
                   pt| running|
    serena|williams|
                       running
                                               31 | 2014 |
                                                             FRA
     roger| federer|
                                               32 | 2016 |
                                                             CHN |
   jeniferl
                  cox|swimming|
                                               32 | 2014 |
                                                             IND
  fernando| johnson|swimming|
                                               32 | 2016 |
                                      silver
                                                             CHN |
              cudrow|javellin|
                                               34 | 2017 |
                                                             USA|
      lisa|
                                        gold|
               louis|javellin|
                                                             RUS
    mathewl
                                               34 | 2015 |
   michael|
              phelps|swimming|
                                               32 | 2017 |
                                                             USA
      ushal
                   pt |
                       running
                                      silver
                                               30 | 2014 |
                                                             IND
                                                             FRA
    serena|williams|
                       running
                                               31 | 2016 |
     roger| federer|
                         tennis
                                               32 | 2017
                                                             CHN
   jeniferl
                  cox|swimming|
                                               32 | 2014 |
                                                             IND|
  fernando| johnson|swimming|
                                               32 | 2017 |
                                                             CHN |
              cudrow|javellin|
                                        gold
                                               34 | 2014 |
                                                             USA
      lisal
    mathew|
               louis|javellin|
                                               34 | 2014 |
                                                             RUS
                                                             USAI
   michaell
              phelps|swimming|
                                               32 | 2017
                                                             IND
```

probelem1

```
scala> val name=udf((fname:String,lname:String)=>{"Mr."+fname.slice(0,2)+"
"+lname+""})
name: org anache spark sql UserDefinedFunction =
```

name: org.apache.spark.sql.UserDefinedFunction =
UserDefinedFunction(<function2>,StringType,List(StringType, StringType))

scala> val first=sports.withColumn("Fullname",name(\$"firstname",\$"lastname"))
first: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string, sports:
string, medal_type: string, age: string, year: string, country: string, Fullname:
string]

```
scala> val result=first.drop("firstname").drop("lastname")
result: org.apache.spark.sql.DataFrame = [sports: string, medal_type: string, age:
```

string, year: string, country: string, Fullname: string]

```
scala> first.show
+----+
|firstname|lastname| sports|medal_type|age|year|country|
                                                 Fullname
+----+
  lisa| cudrow|javellin|
                        gold 34 2015 USA Mr.li cudrow
 mathew louis javellin gold 34 2015 RUS Mr.ma louis
 michael phelps|swimming|
                           silver 32 | 2016 | USA | Mr.mi phelps |
          pt| running|
                      silver| 30|2016| IND|
  usha
                                            Mr.us pt
 serena|williams| running|
                          gold 31 2014 FRA Mr. se williams
  roger| federer| tennis|
                       silver | 32 | 2016 | CHN | Mr.ro federer |
 jenifer|
         cox|swimming|
                         silver| 32|2014| IND|
                                               Mr.je cox
                            silver 32 | 2016 | CHN | Mr.fe johnson |
fernando| johnson|swimming|
  lisa| cudrow|javellin|
                        gold | 34|2017 | USA | Mr.li cudrow |
 mathew louis|javellin|
                         gold 34 2015 RUS Mr.ma louis
 michael | phelps|swimming|
                           silver | 32 | 2017 | USA | Mr.mi phelps |
          pt| running|
                      silver| 30|2014| IND|
   usha
                                            Mr.us pt
 serena|williams| running|
                          gold | 31 | 2016 | FRA | Mr.se williams |
  roger | federer | tennis |
                                      CHN Mr.ro federer
                       silver | 32 | 2017 |
                         silver| 32|2014| IND|
 jenifer
          cox|swimming|
                                               Mr.je cox
 fernando| johnson|swimming|
                            silver 32 | 2017 | CHN | Mr.fe johnson |
  lisa | cudrow|javellin|
                        gold 34 2014 USA Mr.li cudrow
 mathew | louis|javellin|
                         gold | 34 | 2014 | RUS | Mr.ma louis |
 michael | phelps|swimming | silver | 32 | 2017 | USA | Mr.mi phelps |
         pt| running| silver| 30|2014| IND|
                                            Mr.us pt
+----+---+----+----+
only showing top 20 rows
```

scala> fir	st.snow +		+	·	+		
firstname	lastname	sports	medal_type	age	year	country	Fullname
+	+	+	+		+	+	
lisa	cudrow	javellin	gold	34	2015	USA	Mr.li cudrow
mathew	louis	javellin	gold	34	2015	RUS	Mr.ma louis
michael	phelps	swimming	silver	32	2016	USA	Mr.mi phelps
usha	pt	running	silver	30	2016	IND	Mr.us pt
serena	williams	running	gold	31	2014	FRA	Mr.se williams
roger	federer	tennis	silver	32	2016	CHN	Mr.ro federer
jenifer	cox	swimming	silver	32	2014	IND	Mr.je cox
fernando	johnson	swimming	silver	32	2016	CHN	Mr.fe johnsor
lisa	cudrow	javellin	j gold	34	2017	USA	Mr.li cudrow
mathew	louis	javellin	j gold	34	2015	RUS	Mr.ma louis
michael	phelps	swimming	silver	32	2017	USA	Mr.mi phelps
usha	pt	running	silver	30	2014	IND	Mr.us pt
serena	williams	running	j gold	31	2016	FRA	Mr.se williams
roger	federer	tennis	silver	32	2017	CHN	Mr.ro federer
jenifer	cox	swimming	silver	32	2014	IND	Mr.je cox
fernando	johnson	swimming	silver	32	2017	CHN	Mr.fe johnson
lisa	cudrow	javellin	j gold	34	2014	USA	Mr.li cudrow
mathew	louis	javellin	gold	34	2014	RUS	Mr.ma louis
michael	phelps	swimming	silver	32	2017	USA	Mr.mi phelps
j usha	j pt	running	silver	30	2014	IND	Mr.us pt
+	+		+	+	+		

scalar rea	cult chow				
scala> res	sutt.Snow +	++			
sports	medal_type	age	year	country	Fullname
javellin			2015		
javellin swimming		: :	2015 2016		
running running			2016 2014		Mr.us pt Mr.se williams
tennis	silver	j 32 j:	2016 2014	CHN	Mr.ro federer
swimming	silver	j 32 j <i>:</i>	2016	CHN	Mr.fe johnson
javellin javellin	gold	j 34 j <i>i</i>	2017 2015	RUS	Mr.ma louis
swimming running			2017 2014	,	
running tennis	, -		2016 2017	,	Mr.se williams Mr.ro federer
İswimming	silver	j 32 j:	2014	IND	Mr.je cox
swimming javellin	gold	j 34 j <i>i</i>	2017 2014	USA	Mr.li cudrow
javellin swimming			2014 2017	'	
running +	silver	30 : +	2014	IND	Mr.us pt

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 medallists, with age <= 31 are ranked rookie. Add new column into the created UDF.

scala> val

```
player=udf((medal:String,age:String)=>{if(age.toInt<=31&&medal=="silver")
{"rookie";}else{if(age.toInt>=32&&medal=="silver")
{"expert";}else{if(age.toInt<=31&&medal=="gold")
{"amateur";}else{if(age.toInt>=32&&medal=="gold"){"pro";}else{"could not
determine class";}}}})
player: org.apache.spark.sql.UserDefinedFunction =
UserDefinedFunction(<function2>,StringType,List(StringType, StringType))
scala> val result_set=sports.withColumn("class",player($"medal_type",$"age"))
result_set: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string,
sports: string, medal type: string, age: string, year: string, country: string, class:
string]
scala> result set.show
+----++---++----+
|firstname|lastname| sports|medal_type|age|year|country| class|
+----+
   lisa| cudrow|javellin|
                          gold | 34|2015 | USA|
  mathew | louis|javellin|
                           gold | 34|2015 | RUS | pro |
 michael| phelps|swimming|
                             silver 32 | 2016 | USA | expert |
   ushal
           pt| running|
                        silver 30 2016 IND rookie
  serena|williams| running|
                            gold 31 2014 FRA amateur
  roger | federer | tennis |
                         silver 32 | 2016 | CHN | expert |
          cox|swimming|
                           silver 32 | 2014 | IND | expert |
 jenifer
 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 |
```

```
silver | 30 | 2014 | IND | rookie |
   usha
          pt | running |
 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|
   ushal
         pt| running| silver| 30|2014| IND| rookie|
+----+
only showing top 20 rows
```

	+	+	+	+	+	++
firstname	lastname	sports	medal_type	age year	country	class
	+			+	+	++
lisa		javellin	, - ,			pro
mathew		javellin	, - ,			
michael		swimming		•		
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			[CHN	expert
jenifer	cox	swimming	silver	32 2014	IND	expert
fernando	johnson			32 2017	j CHN	expert
	cudrow			34 2014	j USA	
	louis			•		
	phelps	, -	, ,	•		
usha		running				