Topgear - JSON

Consider the following json file

{"id": 1, "fname": "Jeanette", "lname": "Penddreth", "gender": "Female" }
{"id": 2, "fname": "Giavani", "lname": "Frediani", "gender": "Male"}
{"id": 3, "fname": "Noell", "lname": "Bea", "gender": "Female"}
{"id": 4, "fname": "Willard", "lname": "Valek", "gender": "Male"}

1. Create RDD for above json file.

```
scala> import spark.implicits._
scala> val rdd = spark.read.json("text.json")
```

2. display the contents in tabular format.

```
scala> rdd.show()
+--------+------+---+---------+
|   fname|gender| id|    lname|
+--------+------+---+---------+
|Jeanette|Female|  1|Penddreth|
| Giavani|  Male|  2| Frediani|
|   Noell|Female|  3|      Bea|
| Willard|  Male|  4|    Valek|
+--------+------+---+---------+
```

3. Apply filter on id , gender to restrict output.

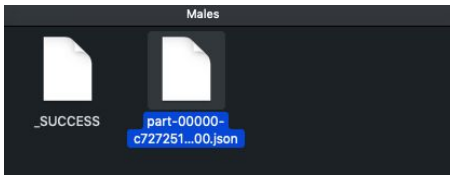
```
scala> rdd.filter(x => x(2) != 2 && x(1)=="Female").show
+--------+------+---+---------+
|   fname|gender| id|    lname|
+--------+------+---+---------+
|Jeanette|Female|  1|Penddreth|
|   Noell|Female|  3|      Bea|
+--------+------+---+---------+
```

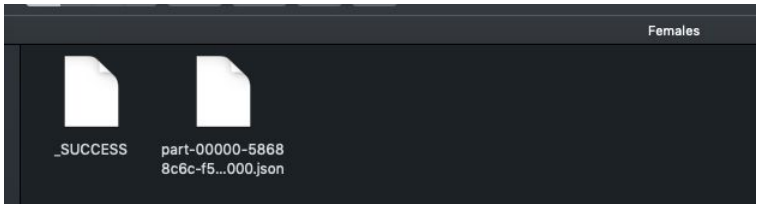
4. display contents in array format.

```
scala> rdd.collect.foreach(println)
[Jeanette,Female,1,Penddreth]
[Giavani,Male,2,Frediani]
[Noell,Female,3,Bea]
[Willard,Male,4,Valek]
```

5. classify whole data into 2 files based on gender and save them as json files.

```
scala> rdd.filter(x => x(1)=="Female").write.json("Females")
scala> rdd.filter(x => x(1)=="Male").write.json("Males")
```





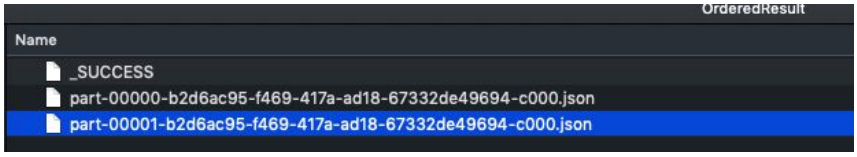
6. differentiate between show and collect when used with data frames.

SNo.	show()	collect()
1	Displays the top 20 rows of DataFrame in a tabular form.	Returns an array that contains all of Rows in the DataFrame.
2	Used for debugging purpose or looking at the data or result.	Used to collect the result into another DataFrame or RDD
3	can be used with DataFrame but not on RDD.	can be used with both DataFrame and RDD.

7. sort the given data based out of gender and store in a file.

scala> rdd.sort($"gender").write.json("OrderedResult")

Saved in the OrderedResult folder:



8. convert given json file into a text file.

scala> val jsonRDD = spark.read.json("text.json")

scala> jsonRDD.rdd.saveAsTextFile("output")

Saved in the output folder: