ex57b

August 18, 2022

[1]:

```python
from pyspark import SparkConf, SparkContext
from pyspark.sql import SparkSession
from graphframes import GraphFrame

conf = SparkConf().setAppName("ex57")
sc = SparkContext(conf=conf)
ssql = SparkSession.builder.getOrCreate()
```

```
22/08/18 18:46:17 WARN Utils: Your hostname, webbelle-XPS-15-7590 resolves to a
loopback address: 127.0.1.1; using 192.168.1.62 instead (on interface wlp58s0)
22/08/18 18:46:17 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
```

```
22/08/18 18:46:17 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
22/08/18 18:46:18 WARN Utils: Service 'SparkUI' could not bind on port 4040.
Attempting port 4041.
22/08/18 18:46:18 WARN Utils: Service 'SparkUI' could not bind on port 4041.
Attempting port 4042.
22/08/18 18:46:18 WARN Utils: Service 'SparkUI' could not bind on port 4042.
Attempting port 4043.
22/08/18 18:46:18 WARN Utils: Service 'SparkUI' could not bind on port 4043.
Attempting port 4044.
22/08/18 18:46:18 WARN Utils: Service 'SparkUI' could not bind on port 4044.
Attempting port 4045.
22/08/18 18:46:18 WARN Utils: Service 'SparkUI' could not bind on port 4045.
Attempting port 4046.
```

[2]:

```python
from pyspark.sql import types
from graphframes.lib import AggregateMessages
from pyspark.sql import functions as F
```

[3]:

```python
edgesPath = "data/Ex57b/data/edges.csv"
vertexesPath = "data/Ex57b/data/vertexes.csv"
outputPath = "out57b/"
```

[4]:

```
eDF = ssql.read.load(
    edgesPath,
    format="csv",
    header=True,
    inferSchema=True
)

vDF = ssql.read.load(
    vertexesPath,
    format="csv",
    header=True,
    inferSchema=True
)
```

[5]:

```
eDF.show(), eDF.printSchema()
vDF.show(), vDF.printSchema()
```

```
+---+---+--------+
|src|dst|linktype|
+---+---+--------+
| u1| u2|  friend|
| u1| u4|  friend|
| u1| u5|  friend|
| u2| u1|  friend|
| u2| u3|  follow|
| u3| u2|  follow|
| u4| u1|  friend|
| u4| u5|  friend|
| u5| u1|  friend|
| u5| u4|  friend|
| u5| u6|  follow|
| u6| u3|  follow|
+---+---+--------+

root
 |-- src: string (nullable = true)
 |-- dst: string (nullable = true)
 |-- linktype: string (nullable = true)

+---+-----+---+
| id| name|age|
+---+-----+---+
| u1|Alice| 34|
| u2|  Bob| 36|
| u3| John| 30|
| u4|David| 29|
| u5| Paul| 32|
| u6| Adel| 36|
| u7| Eddy| 60|
+---+-----+---+

root
 |-- id: string (nullable = true)
 |-- name: string (nullable = true)
 |-- age: integer (nullable = true)
```

[5]:
```
(None, None)
```

[6]:
```python
#ritornare un int piuttosto che un boolean ha molto più senso perchè posso␣
 ↪usare una sum come aggregazione dopo!
def checkAge(age):
    if age<35:
        return 1
    else:
        return 0

ssql.udf.register("ageCheck", checkAge, types.IntegerType())
```

[6]:
```
<function __main__.checkAge(age)>
```

[7]:
```python
filteredVDF = vDF.selectExpr("*", "ageCheck(age) AS AgeLess35")
```

[8]:
```python
filteredVDF.show()
```

```
+---+-----+---+---------+
| id| name|age|AgeLess35|
+---+-----+---+---------+
| u1|Alice| 34|        1|
| u2|  Bob| 36|        0|
| u3| John| 30|        1|
| u4|David| 29|        1|
| u5| Paul| 32|        1|
| u6| Adel| 36|        0|
| u7| Eddy| 60|        0|
+---+-----+---+---------+
```

[9]:
```python
g = GraphFrame(filteredVDF, eDF)
```

```
/home/webbelle/univenv/lib/python3.10/site-
packages/pyspark/sql/dataframe.py:148: UserWarning: DataFrame.sql_ctx is an
internal property, and will be removed in future releases. Use
```

```
DataFrame.sparkSession instead.
  warnings.warn(
```

[10]:

```
msgToDst = AggregateMessages.src["AgeLess35"]
```

[ ]:

```
#senza l'import e l'uso di F questo bit non funziona perchè va in conflitto con␣
  ↪python nativo!!!
ageLess35 = g.aggregateMessages(F.sum(AggregateMessages.msg),
    sendToSrc=None,
    sendToDst=msgToDst
).withColumnRenamed("sum(MSG)", "nn")
```

[12]:

```
ageLess35.show()
```

```
+---+---+
| id| nn|
+---+---+
| u3|  0|
| u4|  2|
| u5|  2|
| u1|  2|
| u6|  1|
| u2|  2|
+---+---+
```

[13]:

```
ageLess35.filter("nn>=2").write.csv(outputPath, header=True)
```