ex56

August 18, 2022

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

     conf = SparkConf().setAppName("ex56")
     sc = SparkContext(conf=conf)
     ssqdl = SparkSession.builder.getOrCreate()
```

```python
[2]: edgesPath = "data/Ex56/data/edges.csv"
     vertexesPath = "data/Ex56/data/vertexes.csv"
     outputPath = "out56/"
```

```python
[4]: eDF = ssqdl.read.load(
         edgesPath,
         format="csv",
         header=True,
         inferSchema=True
     )

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

```python
[5]: eDF.show(), eDF.printSchema()
     vDF.show(), vDF.printSchema()
```

```
+---+---+----------+
|src|dst|  linktype|
+---+---+----------+
| V1| V2|      like|
| V1| V3|    follow|
| V1| V4|    follow|
| V3| V2|    follow|
| V3| V4|    follow|
| V5| V2|  expertOf|
```

```
| V2| V4|correlated|
| V4| V2|correlated|
+---+---+----------+

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


+---+----------+--------+
| id|entityName|    name|
+---+----------+--------+
| V1|      user|   Paolo|
| V2|     topic|     SQL|
| V3|      user|   David|
| V4|     topic|Big Data|
| V5|      user|    John|
+---+----------+--------+

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

[5]: (None, None)

[6]:
```
filteredEDF = eDF.filter("linktype='follow' OR linktype='correlated'")
```

[ ]:
```
g = GraphFrame(vDF, filteredEDF)
```

[9]:
```
pathsDF = g.find("(v1)-[e1]->(v2);(v2)-[e2]->(v3)")
pathsDF.show(), pathsDF.printSchema()
```

```
/home/webbelle/univenv/lib/python3.10/site-
packages/pyspark/sql/dataframe.py:127: UserWarning: DataFrame constructor is
internal. Do not directly use it.
  warnings.warn("DataFrame constructor is internal. Do not directly use it.")

+------------------+------------------+------------------+----------------
----+------------------+
|                v1|                e1|                v2|
e2|                v3|
+------------------+------------------+------------------+----------------
----+------------------+
|  {V1, user, Paolo}|   {V1, V3, follow}|  {V3, user, David}|   {V3, V4,
follow}|{V4, topic, Big D…|
|  {V1, user, Paolo}|   {V1, V3, follow}|  {V3, user, David}|   {V3, V2,
```

```
follow}|    {V2, topic, SQL}|
|   {V1, user, Paolo}|    {V1, V4, follow}|{V4, topic, Big D…|{V4, V2,
correlated}|    {V2, topic, SQL}|
|   {V3, user, David}|    {V3, V2, follow}|    {V2, topic, SQL}|{V2, V4,
correlated}|{V4, topic, Big D…|
|   {V3, user, David}|    {V3, V4, follow}|{V4, topic, Big D…|{V4, V2,
correlated}|    {V2, topic, SQL}|
|    {V2, topic, SQL}|{V2, V4, correlated}|{V4, topic, Big D…|{V4, V2,
correlated}|    {V2, topic, SQL}|
|{V4, topic, Big D…|{V4, V2, correlated}|    {V2, topic, SQL}|{V2, V4,
correlated}|{V4, topic, Big D…|
+------------------+------------------+------------------+---------------
----+------------------+

root
 |-- v1: struct (nullable = false)
 |    |-- id: string (nullable = true)
 |    |-- entityName: string (nullable = true)
 |    |-- name: string (nullable = true)
 |-- e1: struct (nullable = false)
 |    |-- src: string (nullable = true)
 |    |-- dst: string (nullable = true)
 |    |-- linktype: string (nullable = true)
 |-- v2: struct (nullable = false)
 |    |-- id: string (nullable = true)
 |    |-- entityName: string (nullable = true)
 |    |-- name: string (nullable = true)
 |-- e2: struct (nullable = false)
 |    |-- src: string (nullable = true)
 |    |-- dst: string (nullable = true)
 |    |-- linktype: string (nullable = true)
 |-- v3: struct (nullable = false)
 |    |-- id: string (nullable = true)
 |    |-- entityName: string (nullable = true)
 |    |-- name: string (nullable = true)
```

[9]: (None, None)

```python
[10]: finalDF = pathsDF.filter("""
      v1.entityName='user'
      AND e1.linktype='follow'
      AND v2.entityName='topic'
      AND e2.linktype='correlated'
      AND v3.entityName='topic'
      AND v3.name='Big Data'
      """)
```

```
finalDF.show(), finalDF.printSchema()
```

```
+----------------+---------------+--------------+-------------------+------
--------------+
|              v1|             e1|            v2|                 e2|
v3|
+----------------+---------------+--------------+-------------------+------
--------------+
|{V3, user, David}|{V3, V2, follow}|{V2, topic, SQL}|{V2, V4, correlated}|{V4,
topic, Big D…|
+----------------+---------------+--------------+-------------------+------
--------------+

root
 |-- v1: struct (nullable = false)
 |    |-- id: string (nullable = true)
 |    |-- entityName: string (nullable = true)
 |    |-- name: string (nullable = true)
 |-- e1: struct (nullable = false)
 |    |-- src: string (nullable = true)
 |    |-- dst: string (nullable = true)
 |    |-- linktype: string (nullable = true)
 |-- v2: struct (nullable = false)
 |    |-- id: string (nullable = true)
 |    |-- entityName: string (nullable = true)
 |    |-- name: string (nullable = true)
 |-- e2: struct (nullable = false)
 |    |-- src: string (nullable = true)
 |    |-- dst: string (nullable = true)
 |    |-- linktype: string (nullable = true)
 |-- v3: struct (nullable = false)
 |    |-- id: string (nullable = true)
 |    |-- entityName: string (nullable = true)
 |    |-- name: string (nullable = true)
```

[10]: (None, None)

[13]:
```
finalResult = finalDF.selectExpr("v1.name AS USERNAME")
finalResult.write.csv(outputPath, header=True)
```