CSEE5590 Big Data Programming

In Class Programming –13 Report
(Jongkook Son)

Project Overview:

Lesson Title: Graph Frames and GraphX Algorithms
Lesson Description: Distributed Collection of Data

Requirements/Task(s):

Dataset:
https://umkc.box.com/s/vaji06fsdw8bmajg3ndh62v2o1hqpqvv
Graph Frames in Pyspark / Scala
Part – 1:
1. Import the dataset as a csv file and create data frames directly on import than create graph out of the data frame created.
2. Triangle Count
3. Find Shortest Paths w.r.t. Landmarks
4. Apply Page Rank algorithm on the dataset.
5. Save graphs generated to a file.

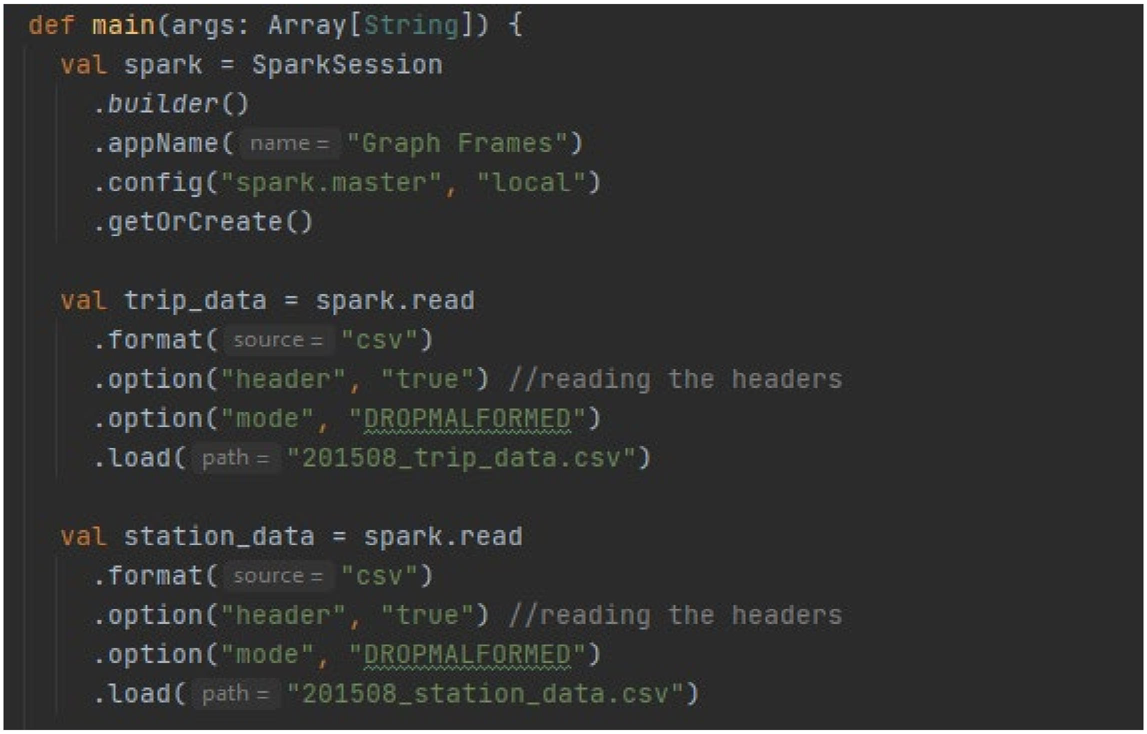
Bonus:
1. Apply Label Propagation Algorithm
2. Apply BFS algorithm

What I learned in ICP:

I learned How to implement GraphFrame using spark. GraphFrames is a package for Apache Spark which provides DataFrame-based Graphs. It provides high-level APIs in Scala, Java, and Python. It aims to provide both the functionality of GraphX and extended functionality taking advantage of Spark DataFrames. By using trip data and station data we make that data as a graph using graph frame. I could get the shortest path between two vertices and implemented pagerank algorithm and bfs algorithm

1. Import the dataset as a csv file and create data frames directly on import than create graph out of the data frame created.

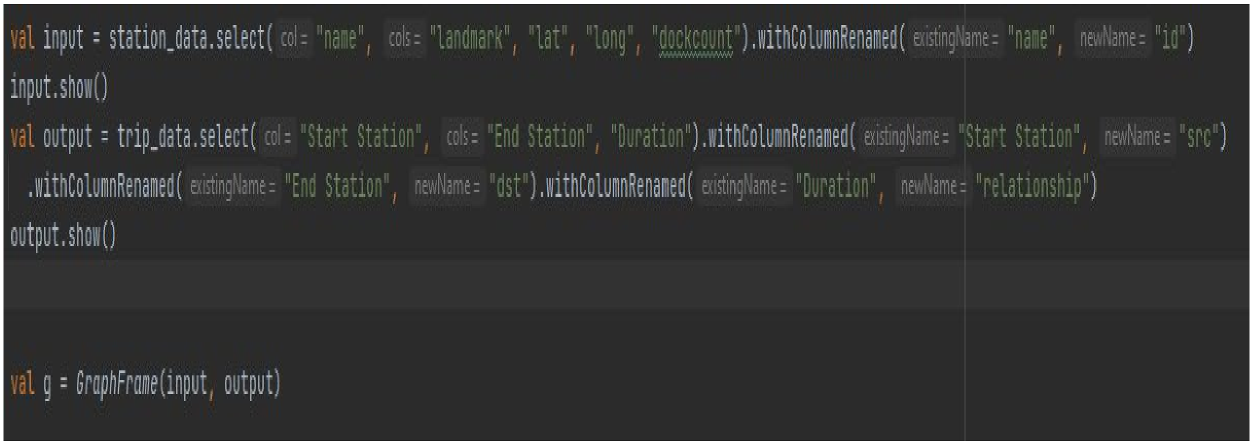
<Code>



```
def main(args: Array[String]) {
  val spark = SparkSession
    .builder()
    .appName( name = "Graph Frames")
    .config("spark.master", "local")
    .getOrCreate()

  val trip_data = spark.read
    .format( source = "csv")
    .option("header", "true") //reading the headers
    .option("mode", "DROPMALFORMED")
    .load( path = "201508_trip_data.csv")

  val station_data = spark.read
    .format( source = "csv")
    .option("header", "true") //reading the headers
    .option("mode", "DROPMALFORMED")
    .load( path = "201508_station_data.csv")
```

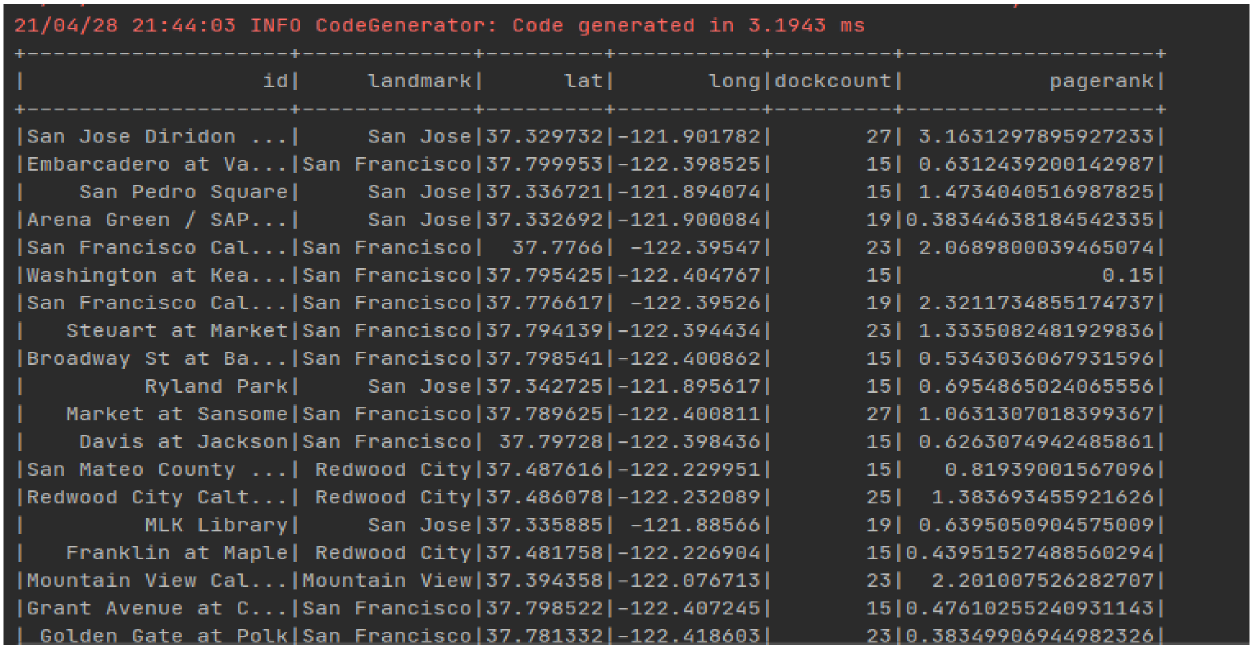


```
val input = station_data.select( col = "name", cols = "landmark", "lat", "long", "dockcount").withColumnRenamed( existingName = "name", newName = "id")
input.show()
val output = trip_data.select( col = "Start Station", cols = "End Station", "Duration").withColumnRenamed( existingName = "Start Station", newName = "src")
  .withColumnRenamed( existingName = "End Station", newName = "dst").withColumnRenamed( existingName = "Duration", newName = "relationship")
output.show()


val g = GraphFrame(input, output)
```

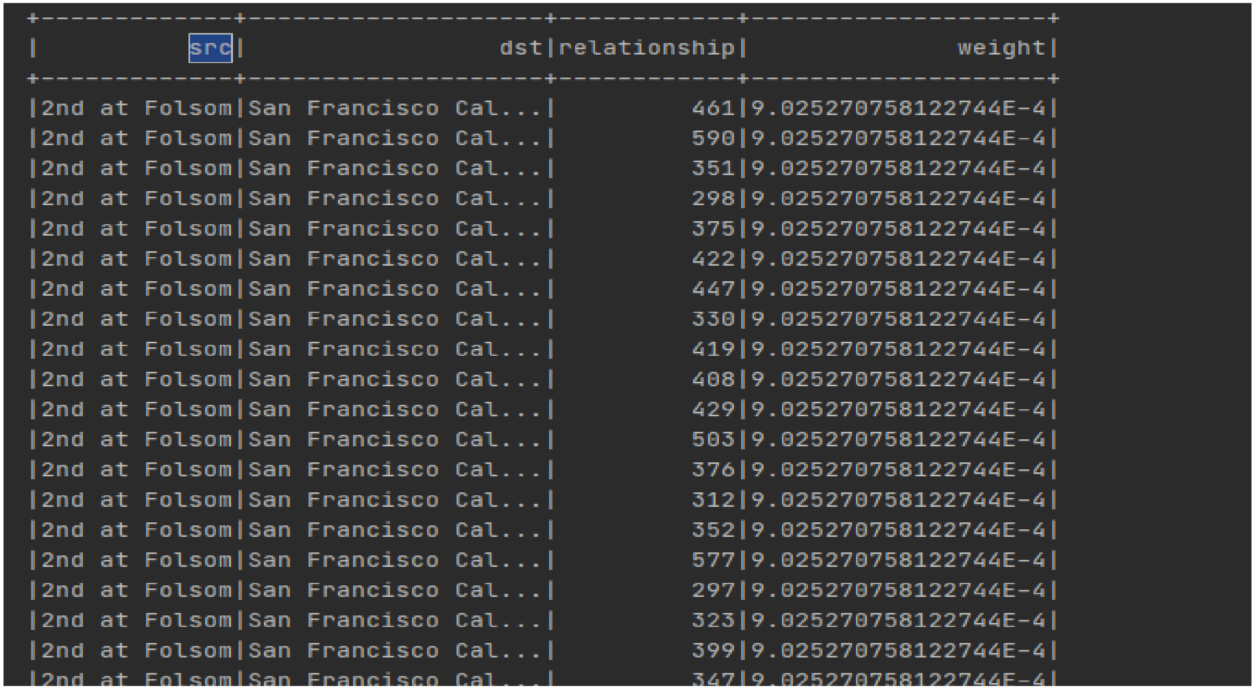
<Output>

Station data



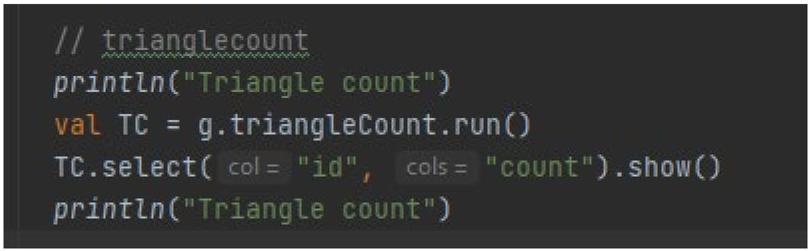
```
21/04/28 21:44:03 INFO CodeGenerator: Code generated in 3.1943 ms
+--------------------+-------------+---------+-----------+---------+------------------+
|                  id|     landmark|      lat|       long|dockcount|          pagerank|
+--------------------+-------------+---------+-----------+---------+------------------+
|San Jose Diridon ...|     San Jose|37.329732|-121.901782|       27| 3.1631297895927233|
|Embarcadero at Va...|San Francisco|37.799953|-122.398525|       15| 0.6312439200142987|
|    San Pedro Square|     San Jose|37.336721|-121.894074|       15| 1.4734040516987825|
|Arena Green / SAP...|     San Jose|37.332692|-121.900084|       19|0.38344638184542335|
|San Francisco Cal...|San Francisco|  37.7766| -122.39547|       23| 2.0689800039465074|
|Washington at Kea...|San Francisco|37.795425|-122.404767|       15|               0.15|
|San Francisco Cal...|San Francisco|37.776617| -122.39526|       19| 2.3211734855174737|
|   Steuart at Market|San Francisco|37.794139|-122.394434|       23| 1.3335082481929836|
|Broadway St at Ba...|San Francisco|37.798541|-122.400862|       15| 0.5343036067931596|
|         Ryland Park|     San Jose|37.342725|-121.895617|       15| 0.6954865024065556|
|   Market at Sansome|San Francisco|37.789625|-122.400811|       27| 1.0631307018399367|
|    Davis at Jackson|San Francisco| 37.79728|-122.398436|       15| 0.6263074942485861|
|San Mateo County ...| Redwood City|37.487616|-122.229951|       15|   0.81939001567096|
|Redwood City Calt...| Redwood City|37.486078|-122.232089|       25|  1.383693455921626|
|         MLK Library|     San Jose|37.335885| -121.88566|       19| 0.6395050904575009|
|   Franklin at Maple| Redwood City|37.481758|-122.226904|       15|0.43951527488560294|
|Mountain View Cal...|Mountain View|37.394358|-122.076713|       23|  2.201007526282707|
|Grant Avenue at C...|San Francisco|37.798522|-122.407245|       15|0.47610255240931143|
| Golden Gate at Polk|San Francisco|37.781332|-122.418603|       23|0.38349906944982326|
```

Trip data

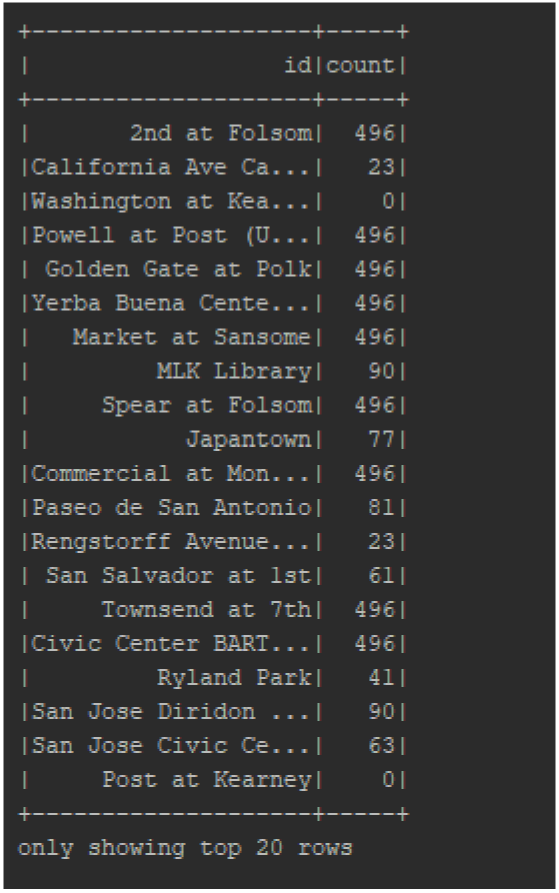


```
+-------------+--------------------+------------+--------------------+
|          src|                 dst|relationship|              weight|
+-------------+--------------------+------------+--------------------+
|2nd at Folsom|San Francisco Cal...|         461|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         590|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         351|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         298|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         375|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         422|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         447|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         330|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         419|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         408|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         429|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         503|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         376|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         312|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         352|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         577|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         297|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         323|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         399|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         347|9.025270758122744E-4|
```

2. Triangle Count



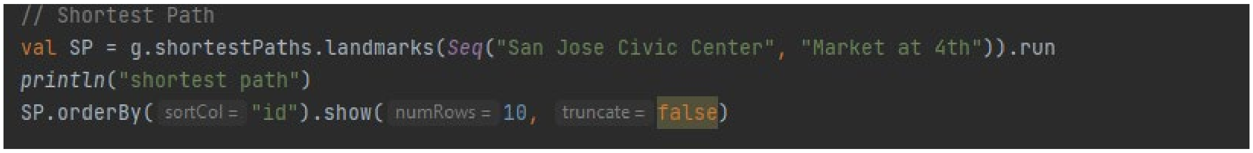
```
// trianglecount
println("Triangle count")
val TC = g.triangleCount.run()
TC.select( col = "id",  cols = "count").show()
println("Triangle count")
```



```
+--------------------+-----+
|                  id|count|
+--------------------+-----+
|       2nd at Folsom|  496|
|California Ave Ca...|   23|
|Washington at Kea...|    0|
|Powell at Post (U...|  496|
| Golden Gate at Polk|  496|
|Yerba Buena Cente...|  496|
|   Market at Sansome|  496|
|         MLK Library|   90|
|     Spear at Folsom|  496|
|           Japantown|   77|
|Commercial at Mon...|  496|
|Paseo de San Antonio|   81|
|Rengstorff Avenue...|   23|
| San Salvador at 1st|   61|
|     Townsend at 7th|  496|
|Civic Center BART...|  496|
|         Ryland Park|   41|
|San Jose Diridon ...|   90|
|San Jose Civic Ce...|   63|
|     Post at Kearney|    0|
+--------------------+-----+
only showing top 20 rows
```

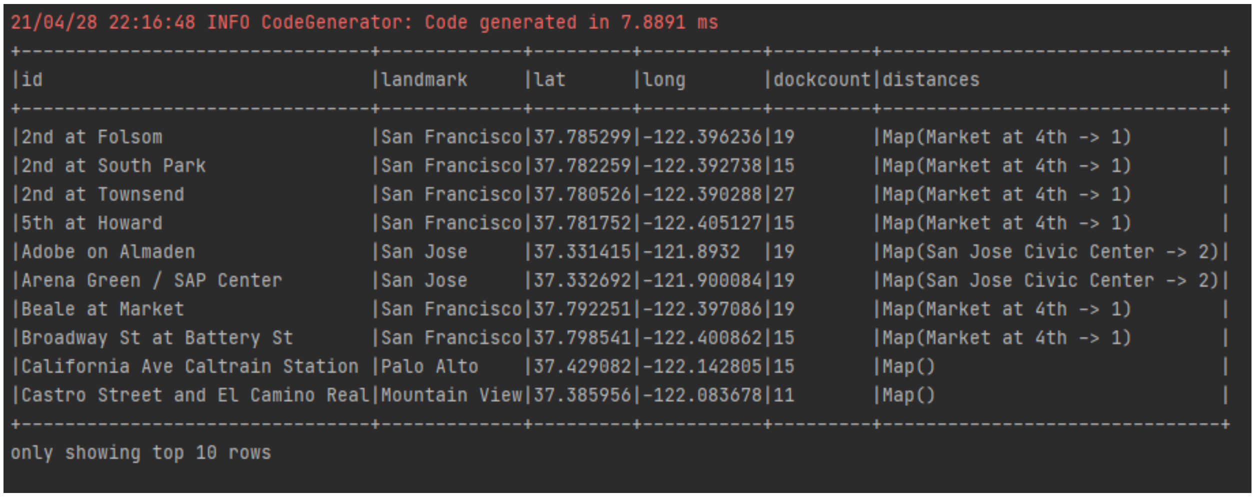
3. Find Shortest Paths w.r.t. Landmarks

<code>



```
// Shortest Path
val SP = g.shortestPaths.landmarks(Seq("San Jose Civic Center", "Market at 4th")).run
println("shortest path")
SP.orderBy( sortCol = "id").show( numRows = 10,  truncate = false)
```

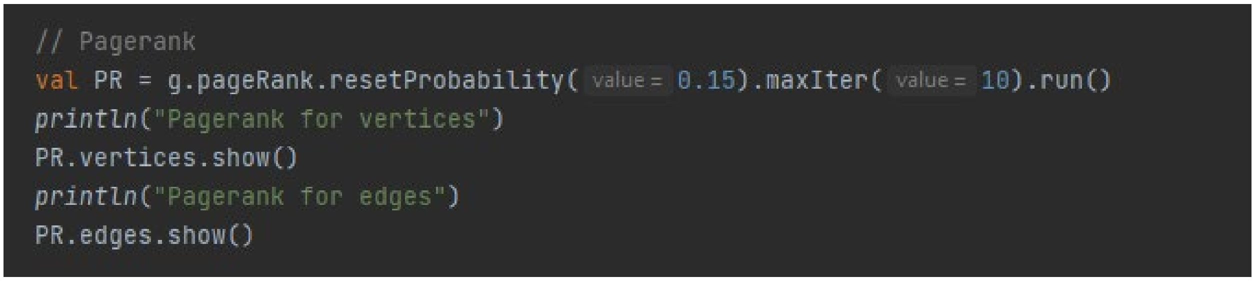
<Output>



```
21/04/28 22:16:48 INFO CodeGenerator: Code generated in 7.8891 ms
+--------------------------------+-------------+---------+-----------+---------+-------------------------------+
|id                              |landmark     |lat      |long       |dockcount|distances                      |
+--------------------------------+-------------+---------+-----------+---------+-------------------------------+
|2nd at Folsom                   |San Francisco|37.785299|-122.396236|19       |Map(Market at 4th -> 1)        |
|2nd at South Park               |San Francisco|37.782259|-122.392738|15       |Map(Market at 4th -> 1)        |
|2nd at Townsend                 |San Francisco|37.780526|-122.390288|27       |Map(Market at 4th -> 1)        |
|5th at Howard                   |San Francisco|37.781752|-122.405127|15       |Map(Market at 4th -> 1)        |
|Adobe on Almaden                |San Jose     |37.331415|-121.8932  |19       |Map(San Jose Civic Center -> 2)|
|Arena Green / SAP Center        |San Jose     |37.332692|-121.900084|19       |Map(San Jose Civic Center -> 2)|
|Beale at Market                 |San Francisco|37.792251|-122.397086|19       |Map(Market at 4th -> 1)        |
|Broadway St at Battery St       |San Francisco|37.798541|-122.400862|15       |Map(Market at 4th -> 1)        |
|California Ave Caltrain Station |Palo Alto    |37.429082|-122.142805|15       |Map()                          |
|Castro Street and El Camino Real|Mountain View|37.385956|-122.083678|11       |Map()                          |
+--------------------------------+-------------+---------+-----------+---------+-------------------------------+
only showing top 10 rows
```

4. Apply Page Rank algorithm on the dataset

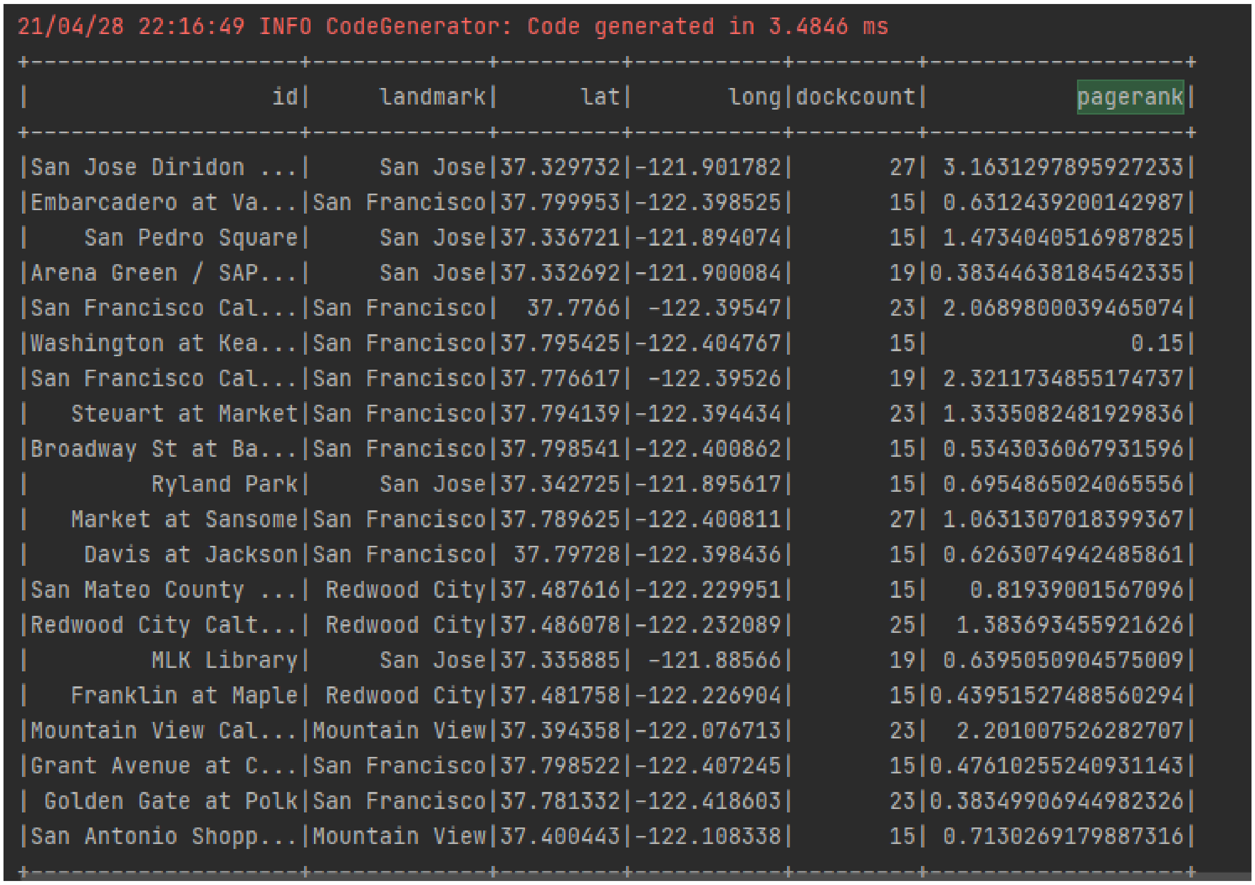
<Code>



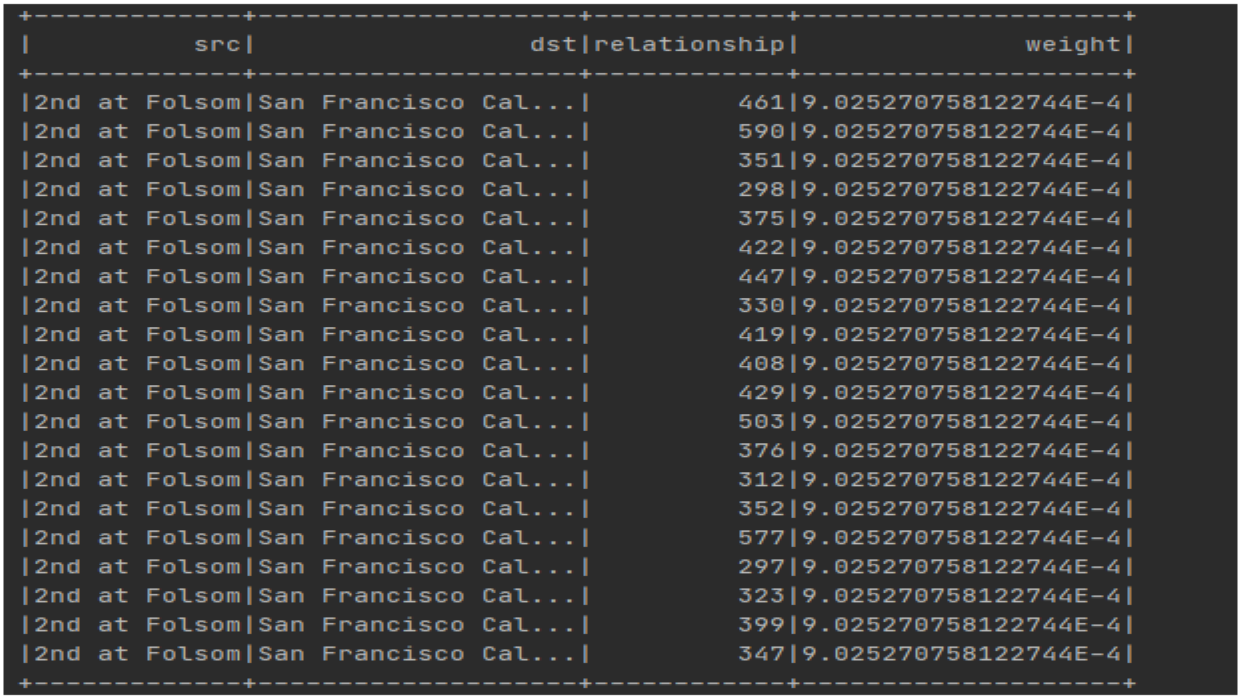
```
// Pagerank
val PR = g.pageRank.resetProbability( value = 0.15).maxIter( value = 10).run()
println("Pagerank for vertices")
PR.vertices.show()
println("Pagerank for edges")
PR.edges.show()
```

<Output>

Pagerank for vertices

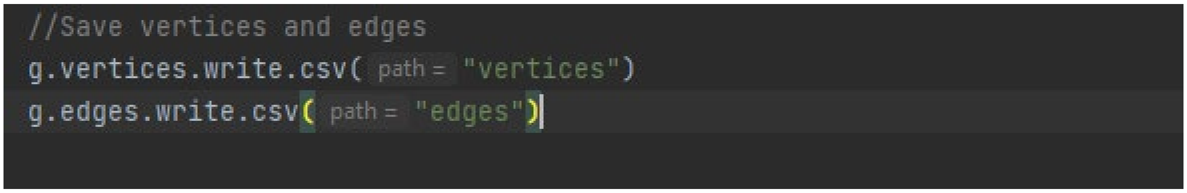


```
21/04/28 22:16:49 INFO CodeGenerator: Code generated in 3.4846 ms
+--------------------+-------------+---------+-----------+---------+-------------------+
|                  id|     landmark|      lat|       long|dockcount|           pagerank|
+--------------------+-------------+---------+-----------+---------+-------------------+
|San Jose Diridon ...|     San Jose|37.329732|-121.901782|       27| 3.1631297895927233|
|Embarcadero at Va...|San Francisco|37.799953|-122.398525|       15| 0.6312439200142987|
|    San Pedro Square|     San Jose|37.336721|-121.894074|       15| 1.4734040516987825|
|Arena Green / SAP...|     San Jose|37.332692|-121.900084|       19|0.38344638184542335|
|San Francisco Cal...|San Francisco|  37.7766| -122.39547|       23| 2.0689800039465074|
|Washington at Kea...|San Francisco|37.795425|-122.404767|       15|               0.15|
|San Francisco Cal...|San Francisco|37.776617| -122.39526|       19| 2.3211734855174737|
|   Steuart at Market|San Francisco|37.794139|-122.394434|       23| 1.3335082481929836|
|Broadway St at Ba...|San Francisco|37.798541|-122.400862|       15| 0.5343036067931596|
|         Ryland Park|     San Jose|37.342725|-121.895617|       15| 0.6954865024065556|
|   Market at Sansome|San Francisco|37.789625|-122.400811|       27| 1.0631307018399367|
|    Davis at Jackson|San Francisco| 37.79728|-122.398436|       15| 0.6263074942485861|
|San Mateo County ...| Redwood City|37.487616|-122.229951|       15|   0.81939001567096|
|Redwood City Calt...| Redwood City|37.486078|-122.232089|       25|  1.383693455921626|
|         MLK Library|     San Jose|37.335885| -121.88566|       19| 0.6395050904575009|
|   Franklin at Maple| Redwood City|37.481758|-122.226904|       15|0.43951527488560294|
|Mountain View Cal...|Mountain View|37.394358|-122.076713|       23|  2.201007526282707|
|Grant Avenue at C...|San Francisco|37.798522|-122.407245|       15|0.47610255240931143|
| Golden Gate at Polk|San Francisco|37.781332|-122.418603|       23|0.38349906944982326|
|San Antonio Shopp...|Mountain View|37.400443|-122.108338|       15| 0.7130269179887316|
+--------------------+-------------+---------+-----------+---------+-------------------+
```

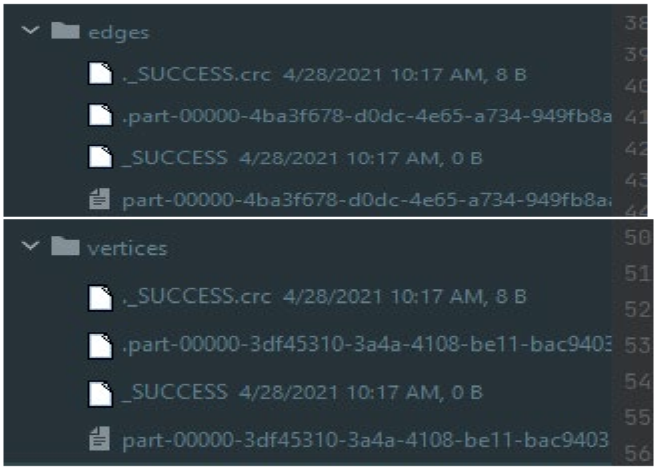


```
+-------------+--------------------+------------+--------------------+
|          src|                 dst|relationship|              weight|
+-------------+--------------------+------------+--------------------+
|2nd at Folsom|San Francisco Cal...|         461|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         590|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         351|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         298|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         375|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         422|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         447|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         330|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         419|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         408|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         429|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         503|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         376|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         312|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         352|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         577|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         297|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         323|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         399|9.025270758122744E-4|
|2nd at Folsom|San Francisco Cal...|         347|9.025270758122744E-4|
+-------------+--------------------+------------+--------------------+
```

5. Save graphs generated to a file

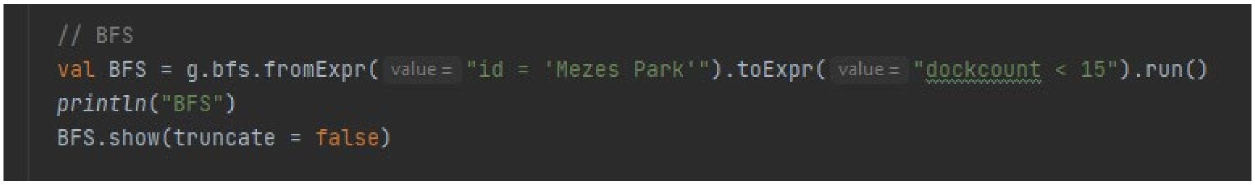


```
//Save vertices and edges
g.vertices.write.csv( path = "vertices")
g.edges.write.csv( path = "edges")
```

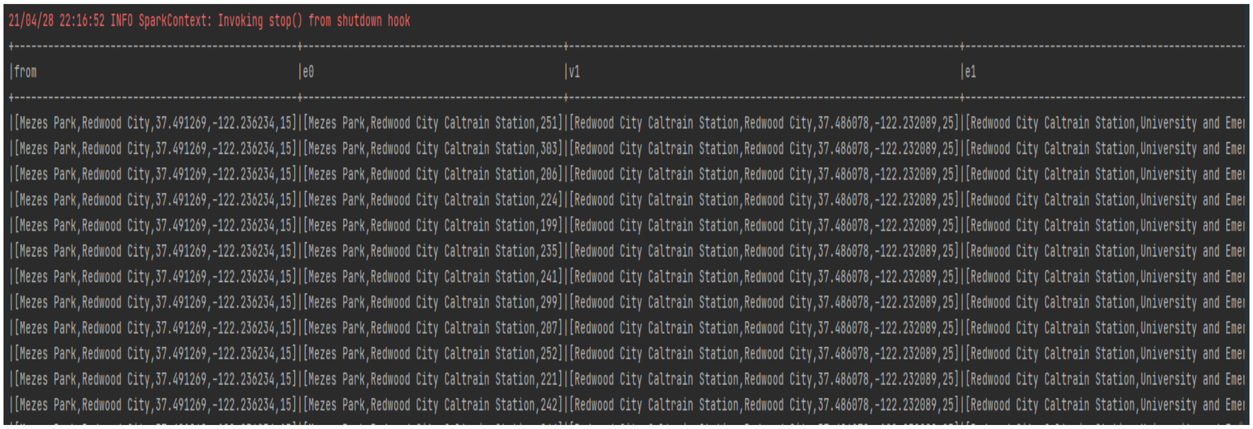


```
edges
    ._SUCCESS.crc  4/28/2021 10:17 AM, 8 B
    .part-00000-4ba3f678-d0dc-4e65-a734-949fb8a
    _SUCCESS  4/28/2021 10:17 AM, 0 B
    part-00000-4ba3f678-d0dc-4e65-a734-949fb8a
vertices
    ._SUCCESS.crc  4/28/2021 10:17 AM, 8 B
    .part-00000-3df45310-3a4a-4108-be11-bac9403
    _SUCCESS  4/28/2021 10:17 AM, 0 B
    part-00000-3df45310-3a4a-4108-be11-bac9403
```

Bonus. Apply BFS algorithm



```
// BFS
val BFS = g.bfs.fromExpr( value = "id = 'Mezes Park'").toExpr( value = "dockcount < 15").run()
println("BFS")
BFS.show(truncate = false)
```



```
21/04/28 22:16:52 INFO SparkContext: Invoking stop() from shutdown hook
+--------------------------------------------------+----------------------------------------------+---------------------------------------------------------------------+--------------------------------------------------
|from                                              |e0                                            |v1                                                                   |e1
+--------------------------------------------------+----------------------------------------------+---------------------------------------------------------------------+--------------------------------------------------
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,251]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,303]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,206]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,224]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,199]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,235]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,241]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,299]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,207]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,252]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,221]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
|[Mezes Park,Redwood City,37.491269,-122.236234,15]|[Mezes Park,Redwood City Caltrain Station,242]|[Redwood City Caltrain Station,Redwood City,37.486078,-122.232089,25]|[Redwood City Caltrain Station,University and Emer
```