

Assignment 5

Team 8

Sree Rama Raju Pericharla

Swetha Chandra Karroti

Pallavi Ramineni

GitHub URL:

<https://github.com/sreeram66/Parallel-Algorithm/tree/master/Assignment5>

Graph Algorithms:

PageRank

PageRank measures the importance of each vertex in a graph, assuming an edge from u to v represents an endorsement of v 's importance by u . For example, if a Twitter user is followed by many others, the user will be ranked highly.

GraphX comes with static and dynamic implementations of PageRank as methods on the [PageRank object](#). Static PageRank runs for a fixed number of iterations, while dynamic PageRank runs until the ranks converge (i.e., stop changing by more than a specified tolerance). [GraphOps](#) allows calling these algorithms directly as methods on Graph.

Input:

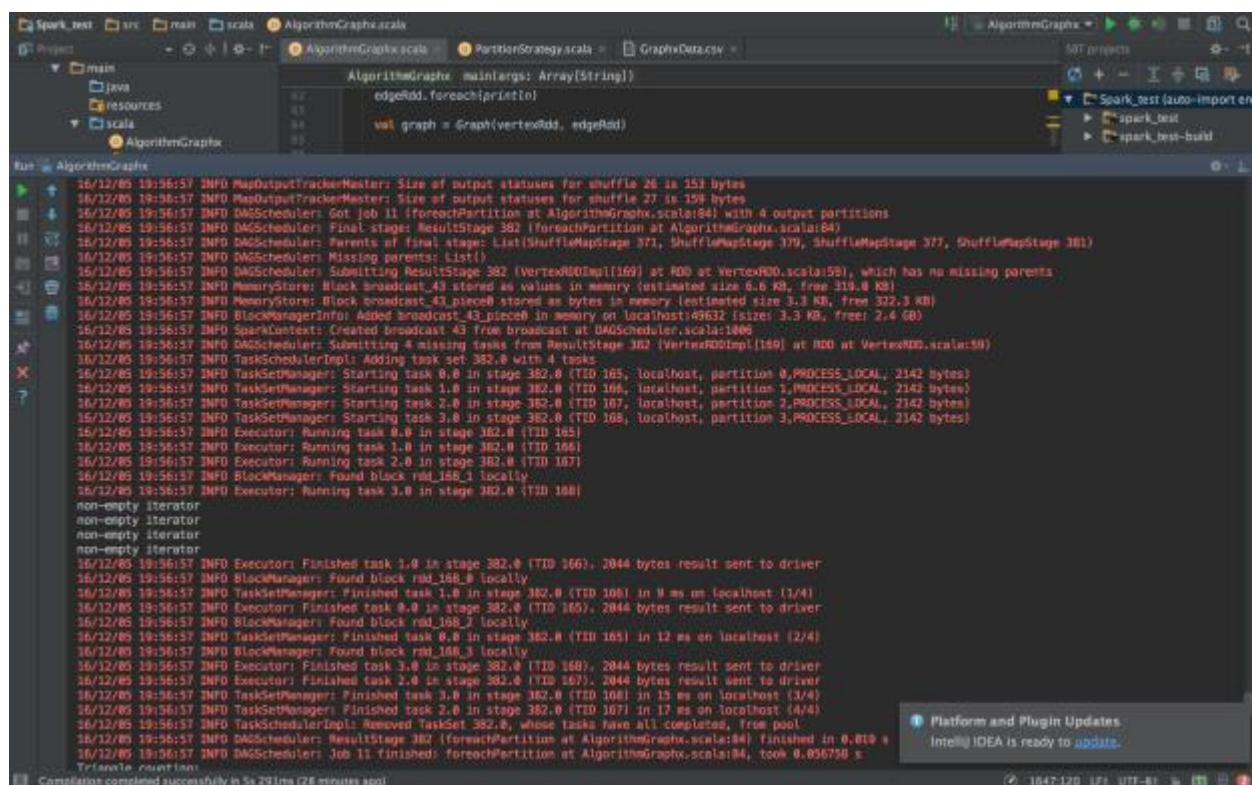
```
Dallas,to,Chicago
Chicago,from,Kansas City
Kansas City,from,New York
Las Vegas,from,Denver
```

```
Project | Spark_test | src | main | scala | AlgorithmGraphx.scala | PartitionStrategy.scala | GraphData.csv |
main |
  | java |
  | resources |
  | scala |
  | AlgorithmGraphx |
Run | AlgorithmGraphx |
16/12/05 19:36:56 INFO BlockManager: Found block rdd_105_1 locally
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Getting 3 non-empty blocks out of 4 blocks
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
(Chicago,0,3858749999999999)
16/12/05 19:36:56 INFO MemoryStore: Block rdd_122_1 stored as values in memory (estimated size 1976.0 B, free 316.0 KB)
16/12/05 19:36:56 INFO BlockManagerInfo: Added rdd_122_1 in memory on localhost:49632 (size: 1976.0 B, free: 2.4 GB)
16/12/05 19:36:56 INFO Executor: Finished task 3.0 in stage 251.0 (TID 116), 2759 bytes result sent to driver
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Getting 3 non-empty blocks out of 4 blocks
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 2 ms
16/12/05 19:36:56 INFO TaskSetManager: Finished task 3.0 in stage 251.0 (TID 116) in 15 ms on localhost (1/4)
16/12/05 19:36:56 INFO BlockManager: Raining task 2.0 in stage 251.0 (TID 115)
16/12/05 19:36:56 INFO CacheManager: Partition rdd_122.0 not found, computing it
16/12/05 19:36:56 INFO BlockManager: Found block rdd_105_0 locally
(Denver,0,15)
(New York,0,15)
16/12/05 19:36:56 INFO Executor: Finished task 1.0 in stage 251.0 (TID 114), 2759 bytes result sent to driver
16/12/05 19:36:56 INFO MemoryStore: Block rdd_122.0 stored as values in memory (estimated size 1960.0 B, free 337.9 KB)
16/12/05 19:36:56 INFO BlockManagerInfo: Added rdd_122.0 in memory on localhost:49632 (size: 1960.0 B, free: 2.4 GB)
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Getting 3 non-empty blocks out of 4 blocks
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/12/05 19:36:56 INFO CacheManager: Partition rdd_122.2 not found, computing it
16/12/05 19:36:56 INFO BlockManager: Found block rdd_105_2 locally
16/12/05 19:36:56 INFO TaskSetManager: Finished task 1.0 in stage 251.0 (TID 114) in 23 ms on localhost (2/4)
16/12/05 19:36:56 INFO MemoryStore: Block rdd_122.2 stored as values in memory (estimated size 1976.0 B, free 339.9 KB)
16/12/05 19:36:56 INFO BlockManagerInfo: Added rdd_122.2 in memory on localhost:49632 (size: 1976.0 B, free: 2.4 GB)
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Getting 3 non-empty blocks out of 4 blocks
16/12/05 19:36:56 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 8 ms
16/12/05 19:36:56 INFO Executor: Finished task 0.0 in stage 251.0 (TID 113), 2759 bytes result sent to driver
16/12/05 19:36:56 INFO TaskSetManager: Finished task 0.0 in stage 251.0 (TID 113) in 29 ms on localhost (3/4)
16/12/05 19:36:56 INFO Executor: Finished task 2.0 in stage 251.0 (TID 115), 2759 bytes result sent to driver
(Dallas,0,47799375)
(Las Vegas,0,2774999999999999)
(Kansas City,0,2774999999999999)
16/12/05 19:36:56 INFO TaskSetManager: Finished task 2.0 in stage 251.0 (TID 115) in 27 ms on localhost (4/4)
16/12/05 19:36:56 INFO TaskSchedulerImpl: Removed TaskSet 251.0, whose tasks have all completed, from pool
16/12/05 19:36:56 INFO DAGScheduler: ResultStage 251 (foreach at AlgorithmGraphx.scala:74) finished in 0.031 s
16/12/05 19:36:56 INFO DAGScheduler: Job 7 finished: foreach at AlgorithmGraphx.scala:74, took 0.129122 s
16/12/05 19:36:56 INFO SparkContext: Starting job: reduce at VertexRDDImpl.scala:98
16/12/05 19:36:56 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 187 bytes
16/12/05 19:36:56 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 1 is 188 bytes
16/12/05 19:36:56 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 2 is 168 bytes
16/12/05 19:36:56 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 3 is 168 bytes
Compilation completed successfully in 5s 291ms (25 minutes ago) | 1647-120 LPI UTF-8 |
```

The connected components algorithm labels each connected component of the graph with the ID of its lowest-numbered vertex. For example, in a social network, connected components can approximate clusters. GraphX contains an implementation of the algorithm in the `ConnectedComponents` object, and we compute the connected components of the example social network dataset from the [PageRank section](#) as follows:

```
Dallas,to,Chicago
Chicago,from,Kansas City
Kansas City,from,New York
Las Vegas,from,Denver
```

Output:



```
AlgorithmGraphx.scala
main(args: Array[String])
edgeRdd.foreach(println)
val graph = Graph(verticesRdd, edgeRdd)

Run AlgorithmGraphx
16/12/85 19:56:57 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 26 is 153 bytes
16/12/85 19:56:57 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 27 is 159 bytes
16/12/85 19:56:57 INFO DAGScheduler: Got job 11 (foreachPartition at AlgorithmGraphx.scala:84) with 4 output partitions
16/12/85 19:56:57 INFO DAGScheduler: Final stage: ResultStage 382 (foreachPartition at AlgorithmGraphx.scala:84)
16/12/85 19:56:57 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 371, ShuffleMapStage 370, ShuffleMapStage 361)
16/12/85 19:56:57 INFO DAGScheduler: Missing parents: List()
16/12/85 19:56:57 INFO DAGScheduler: Submitting ResultStage 382 (VertexRDDImpl[169] at RDD at VertexRDD.scala:59), which has no missing parents
16/12/85 19:56:57 INFO MemoryStore: Block broadcast_43 stored as values in memory (estimated size 6.0 KB, free 319.0 KB)
16/12/85 19:56:57 INFO MemoryStore: Block broadcast_43_piece0 stored as bytes in memory (estimated size 3.3 KB, free 322.3 KB)
16/12/85 19:56:57 INFO SparkContext: Created broadcast 43 from broadcast at DAGScheduler.scala:1896
16/12/85 19:56:57 INFO DAGScheduler: Submitting 4 missing tasks from ResultStage 382 (VertexRDDImpl[169] at RDD at VertexRDD.scala:59)
16/12/85 19:56:57 INFO TaskSchedulerImpl: Adding task set 382.0 with 4 tasks
16/12/85 19:56:57 INFO TaskSetManager: Starting task 0.0 in stage 382.0 (TID 165, localhost, partition 0, PROCESS_LOCAL, 2142 bytes)
16/12/85 19:56:57 INFO TaskSetManager: Starting task 1.0 in stage 382.0 (TID 166, localhost, partition 1, PROCESS_LOCAL, 2142 bytes)
16/12/85 19:56:57 INFO TaskSetManager: Starting task 2.0 in stage 382.0 (TID 167, localhost, partition 2, PROCESS_LOCAL, 2142 bytes)
16/12/85 19:56:57 INFO TaskSetManager: Starting task 3.0 in stage 382.0 (TID 168, localhost, partition 3, PROCESS_LOCAL, 2142 bytes)
16/12/85 19:56:57 INFO Executor: Running task 0.0 in stage 382.0 (TID 165)
16/12/85 19:56:57 INFO Executor: Running task 1.0 in stage 382.0 (TID 166)
16/12/85 19:56:57 INFO Executor: Running task 2.0 in stage 382.0 (TID 167)
16/12/85 19:56:57 INFO BlockManager: Found block rdd_168_1 locally
16/12/85 19:56:57 INFO Executor: Running task 3.0 in stage 382.0 (TID 168)
non-empty iterator
non-empty iterator
non-empty iterator
non-empty iterator
16/12/85 19:56:57 INFO Executor: Finished task 1.0 in stage 382.0 (TID 166), 2044 bytes result sent to driver
16/12/85 19:56:57 INFO BlockManager: Found block rdd_168_0 locally
16/12/85 19:56:57 INFO TaskSetManager: Finished task 1.0 in stage 382.0 (TID 166) in 9 ms on localhost (1/4)
16/12/85 19:56:57 INFO Executor: Finished task 0.0 in stage 382.0 (TID 165), 2044 bytes result sent to driver
16/12/85 19:56:57 INFO BlockManager: Found block rdd_168_2 locally
16/12/85 19:56:57 INFO TaskSetManager: Finished task 0.0 in stage 382.0 (TID 165) in 12 ms on localhost (2/4)
16/12/85 19:56:57 INFO BlockManager: Found block rdd_168_3 locally
16/12/85 19:56:57 INFO Executor: Finished task 3.0 in stage 382.0 (TID 168), 2044 bytes result sent to driver
16/12/85 19:56:57 INFO Executor: Finished task 2.0 in stage 382.0 (TID 167), 2044 bytes result sent to driver
16/12/85 19:56:57 INFO TaskSetManager: Finished task 3.0 in stage 382.0 (TID 168) in 15 ms on localhost (3/4)
16/12/85 19:56:57 INFO TaskSetManager: Finished task 2.0 in stage 382.0 (TID 167) in 17 ms on localhost (4/4)
16/12/85 19:56:57 INFO TaskSchedulerImpl: Removed TaskSet 382.0, whose tasks have all completed, from pool
16/12/85 19:56:57 INFO DAGScheduler: ResultStage 382 (foreachPartition at AlgorithmGraphx.scala:84) finished in 0.019 s
16/12/85 19:56:57 INFO DAGScheduler: Job 11 finished: foreachPartition at AlgorithmGraphx.scala:84, took 0.056759 s
Triaxite: execution
Compilation completed successfully in 5s 291ms (28 minutes ago)
```

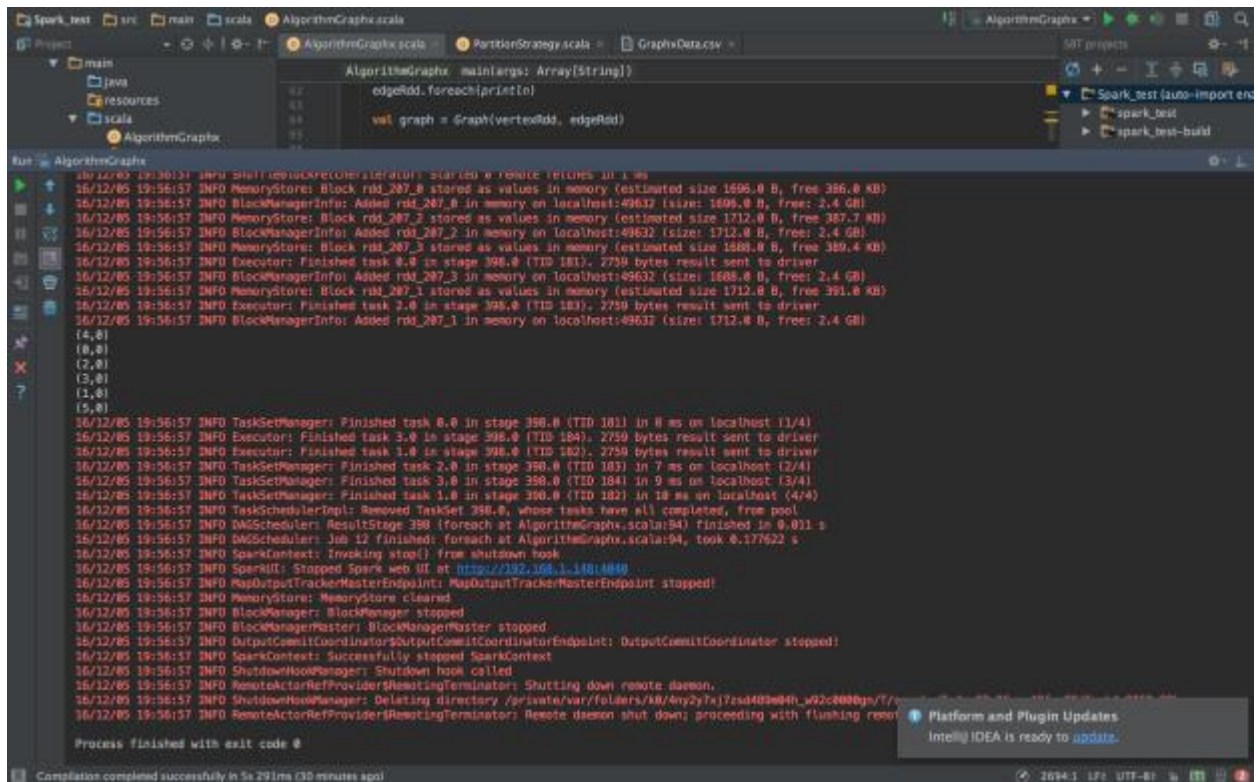
Triangle Counting

A vertex is part of a triangle when it has two adjacent vertices with an edge between them. GraphX implements a triangle counting algorithm in the [TriangleCount](#) object that determines the number of triangles passing through each vertex, providing a measure of clustering. We compute the triangle count of the social network dataset from the [PageRank](#) section. Note that *TriangleCount* requires the edges to be in canonical orientation ($srcId < dstId$) and the graph to be partitioned using [Graph.partitionBy](#).

Input:

```
Dallas,to,Chicago
Chicago,from,Kansas City
Kansas City,from,New York
Las Vegas,from,Denver
```

Output:



```
AlgorithmGraphx.scala
mainArgs: Array[String]
edgeRdd.foreach(print)
val graph = Graph(vertexRdd, edgeRdd)

Run AlgorithmGraphx
16/12/85 19:56:57 INFO ShellUtils: createCoordinator started 0 remote replicas in 1 ms
16/12/85 19:56:57 INFO MemoryStore: Block rdd_287_0 stored as values in memory (estimated size 1696.0 B, free 386.0 KB)
16/12/85 19:56:57 INFO BlockManagerInfo: Added rdd_287_0 in memory on localhost:49632 (size: 1696.0 B, free: 2.4 GB)
16/12/85 19:56:57 INFO MemoryStore: Block rdd_287_2 stored as values in memory (estimated size 1712.0 B, free 387.7 KB)
16/12/85 19:56:57 INFO BlockManagerInfo: Added rdd_287_2 in memory on localhost:49632 (size: 1712.0 B, free: 2.4 GB)
16/12/85 19:56:57 INFO MemoryStore: Block rdd_287_3 stored as values in memory (estimated size 1888.0 B, free 389.4 KB)
16/12/85 19:56:57 INFO Executor: Finished task 0.0 in stage 396.0 (TID 181). 2759 bytes result sent to driver
16/12/85 19:56:57 INFO BlockManagerInfo: Added rdd_287_3 in memory on localhost:49632 (size: 1888.0 B, free: 2.4 GB)
16/12/85 19:56:57 INFO MemoryStore: Block rdd_287_1 stored as values in memory (estimated size 1712.0 B, free 391.0 KB)
16/12/85 19:56:57 INFO Executor: Finished task 2.0 in stage 396.0 (TID 183). 2759 bytes result sent to driver
16/12/85 19:56:57 INFO BlockManagerInfo: Added rdd_287_1 in memory on localhost:49632 (size: 1712.0 B, free: 2.4 GB)
(4,0)
(0,0)
(2,0)
(3,0)
(1,0)
(5,0)
16/12/85 19:56:57 INFO TaskSetManager: Finished task 0.0 in stage 396.0 (TID 181) in 8 ms on localhost (1/4)
16/12/85 19:56:57 INFO Executor: Finished task 3.0 in stage 396.0 (TID 184). 2759 bytes result sent to driver
16/12/85 19:56:57 INFO Executor: Finished task 1.0 in stage 396.0 (TID 182). 2759 bytes result sent to driver
16/12/85 19:56:57 INFO TaskSetManager: Finished task 2.0 in stage 396.0 (TID 183) in 7 ms on localhost (2/4)
16/12/85 19:56:57 INFO TaskSetManager: Finished task 3.0 in stage 396.0 (TID 184) in 9 ms on localhost (3/4)
16/12/85 19:56:57 INFO TaskSetManager: Finished task 1.0 in stage 396.0 (TID 182) in 10 ms on localhost (4/4)
16/12/85 19:56:57 INFO TaskSchedulerImpl: Removed TaskSet 396.0, whose tasks have all completed, from pool
16/12/85 19:56:57 INFO DAGScheduler: ResultStage 396 (foreach at AlgorithmGraphx.scala:94) finished in 0.031 s
16/12/85 19:56:57 INFO DAGScheduler: Job 12 finished: foreach at AlgorithmGraphx.scala:94, took 0.177622 s
16/12/85 19:56:57 INFO SparkContext: Invoking stop() from shutdown hook
16/12/85 19:56:57 INFO SparkUI: Stopped Spark web UI at http://132.164.1.148:4040
16/12/85 19:56:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/12/85 19:56:57 INFO MemoryStore: MemoryStore cleared
16/12/85 19:56:57 INFO BlockManager: BlockManager stopped
16/12/85 19:56:57 INFO BlockManagerMaster: BlockManagerMaster stopped
16/12/85 19:56:57 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
16/12/85 19:56:57 INFO SparkContext: Successfully stopped SparkContext
16/12/85 19:56:57 INFO ShutdownHookManager: Shutdown hook called
16/12/85 19:56:57 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
16/12/85 19:56:57 INFO ShutdownHookManager: Deleting directory /private/var/folders/X8/4ny2y7xj7zsd489w4h_w02c888bgn/T/...
16/12/85 19:56:57 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote...

Platform and Plugin Updates
IntelliJ IDEA is ready to update.

Compiler completed successfully in 5s 291ms (30 minutes ago)
```