

Big Data Programming Assignment 4

Varaprasad Kurra

Panther ID: 002430487

Source code:

```
import java.util.ArrayList;
import java.util.Iterator;
import java.util.List;
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaPairRDD;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;
import org.apache.spark.api.java.function.Function;
import org.apache.spark.api.java.function.Function2;
import org.apache.spark.api.java.function.PairFlatMapFunction;
import org.apache.spark.api.java.function.PairFunction;
import org.apache.spark.broadcast.Broadcast;
import org.spark_project.guava.collect.Iterables;
import scala.Tuple2;
import shapeless.newtype;

public class SparkPageRank
{

    private static final String LINK_URI =
"file:///C:/Users/VaraPrasad/Desktop/Summer_Semester/pageRankTinyWeb.txt";
    private static final int totalIteration = 10;
    private static final double beta = 0.85;

    public static void main(String[] args)
    {
        // initializing spark
        SparkConf conf = new
SparkConf().setAppName("SparkAverage").setMaster("local[2]");
        JavaSparkContext sc = new JavaSparkContext(conf);
        sc.setLogLevel("WARN");

        // identify all the neighbors for each page
        class PageNeighbor implements
PairFunction<String,String,String>
        {
            @Override
            public Tuple2<String,String> call(String line) throws
Exception {
                String[] link = line.split(" ");
                return new Tuple2<String,String>(link[0],link[1]);
            }
        }

        // prepare functions needed by aggregateByKey
        ArrayList<String> neiList = new ArrayList<>();
        Function2<ArrayList<String>, String, ArrayList<String>>
accumulator = new Function2<ArrayList<String>, String, ArrayList<String>>() {
            private static final long serialVersionUID = 2323;
```

```

        @Override
        public ArrayList<String> call(ArrayList<String> nei, String
n) throws Exception {
            nei.add(n);
            return nei;
        }
    };

    Function2<ArrayList<String>, ArrayList<String>,
ArrayList<String>> merger = new Function2<ArrayList<String>,
ArrayList<String>, ArrayList<String>>() {
        private static final long serialVersionUID = 9898;
        @Override
        public ArrayList<String> call(ArrayList<String> nei1,
ArrayList<String> nei2) throws Exception {
            nei1.addAll(nei2);
            return nei1;
        }
    };

    JavaRDD<String> linkFile = sc.textFile(LINK_URI);
    JavaPairRDD<String, ArrayList<String>> links =
linkFile.mapToPair(new PageNeighbor()).aggregateByKey(neiList, accumulator,
merger).cache();

    System.out.println("links has [" + links.count() + "] elements");
    System.out.println(links.take((int)links.count()).toString());

    // get the total number of pages in the network
    final Broadcast<Long> numOfPages = sc.broadcast(links.count());

    // initialize the pageRank vector, each component has
1/numOfPages at its initial rank
    JavaPairRDD<String, Double> pageRank = links.mapValues( new
Function<ArrayList<String>, Double>() {
        @Override
        public Double call(ArrayList<String> neighborURL) {
            return 1.0/numOfPages.value();
            // return 1.0;
        }
    } );
    System.out.println("pageRank has [" + pageRank.count() + "]
elements");

    System.out.println(pageRank.take((int)pageRank.count()).toString());

    // define helper classes
    class RankContribution implements
PairFlatMapFunction<Tuple2<ArrayList<String>, Double>, String, Double> {

        @Override
        public Iterator<Tuple2<String, Double>>
call(Tuple2<ArrayList<String>, Double> linkConfig) throws Exception {
            List<Tuple2<String, Double>> results = new
ArrayList<Tuple2<String, Double>>();
            int neighborCount = Iterables.size(linkConfig._1);
            for (String neighborURL : linkConfig._1) {

```

```

        results.add(new
Tuple2<String,Double>(neighborURL,linkConfig._2/neighborCount));
    }
    return results.iterator();
}

}

class RankAdjust implements Function2<Double,Double,Double> {
    @Override
    public Double call(final Double value1,final Double value2)
{
        return value1 + value2;
    }
}

class FinalRankAdjust implements Function<Double,
Double>
{
    public Double call(final Double rank)
    {
        return beta*rank +(1-beta)*0.25;
    }
}

for ( int i = 0 ; i < totalIteration; i ++ ) {

    JavaPairRDD<String, Tuple2<ArrayList<String>, Double>>
joinedRDD = links.join(pageRank);
    //System.out.println("joinedRDD has [" + joinedRDD.count()
+ "]" elements");

    //System.out.println(joinedRDD.take((int)joinedRDD.count()).toString())
;

    JavaRDD<Tuple2<ArrayList<String>, Double>> weightRDD =
joinedRDD.values();
    //System.out.println("weightRDD has [" + weightRDD.count()
+ "]" elements");
    //
    System.out.println(weightRDD.take((int)weightRDD.count()).toString());

    // calculate contribution
    JavaPairRDD<String, Double> contribs =
weightRDD.flatMapToPair(new RankContribution());
    //System.out.println("contribs has [" + contribs.count() +
"] elements");

    //System.out.println(contribs.take((int)contribs.count()).toString());

    // adjust current rank
    pageRank = contribs.reduceByKey(new RankAdjust());
    pageRank = pageRank.mapValues(new
FinalRankAdjust());

```

```

        System.out.println("pageRank has [" + pageRank.count() + "]"
elements");

        System.out.println(pageRank.take((int)pageRank.count()).toString());
    }
    System.out.println("pageRank has altogether [" + pageRank.count()
+ "]" elements");
    System.out.println(pageRank.take((int)pageRank.count()).toString());
    // pageRank.saveAsTextFile("pageRankResult.txt");
    numOfPages.unpersist();
    sc.close();
}
}

```

Class & Function call added to avoid the Spider-Trap:

```

class FinalRankAdjust implements Function<Double, Double>
{
    public Double call(final Double rank)
    {
        return beta*rank + (1-beta)*0.25;
    }
}

```

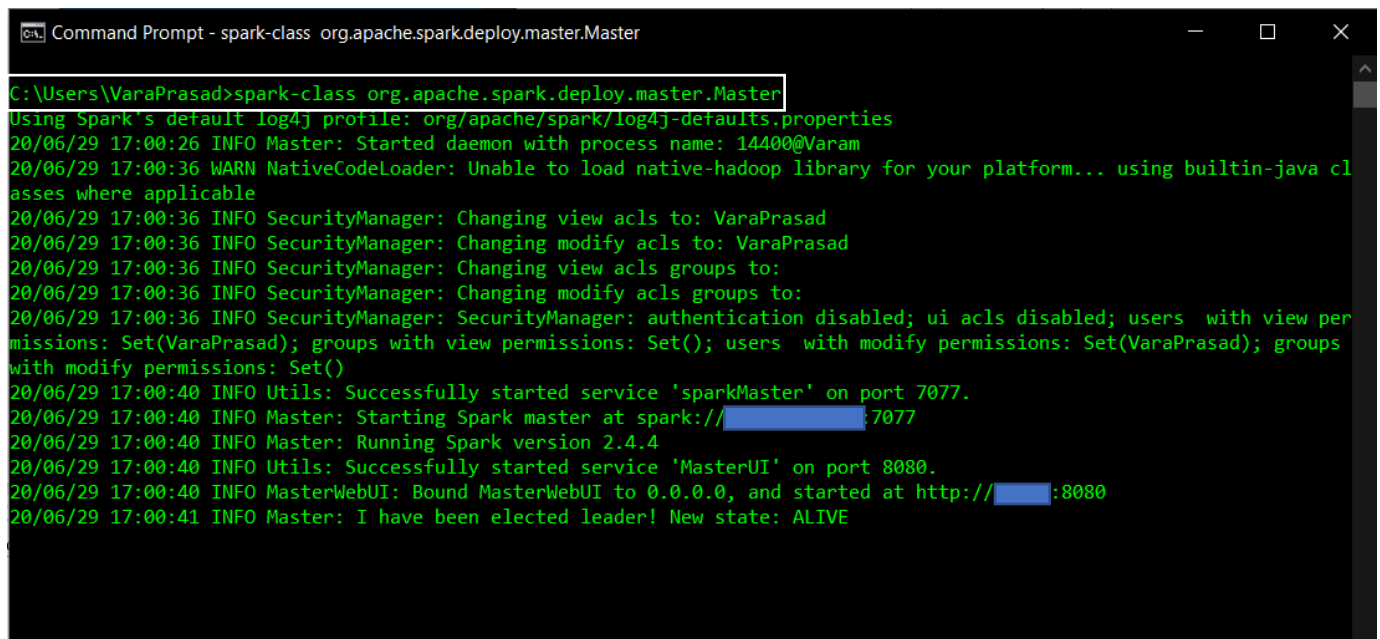
Implementing the new class:

```

pageRank = pageRank.mapValues(new FinalRankAdjust());

```

Command to Start a Spark Master Node:



```

C:\Users\VaraPrasad>spark-class org.apache.spark.deploy.master.Master
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
20/06/29 17:00:26 INFO Master: Started daemon with process name: 14400@Varam
20/06/29 17:00:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
20/06/29 17:00:36 INFO SecurityManager: Changing view acls to: VaraPrasad
20/06/29 17:00:36 INFO SecurityManager: Changing modify acls to: VaraPrasad
20/06/29 17:00:36 INFO SecurityManager: Changing view acls groups to:
20/06/29 17:00:36 INFO SecurityManager: Changing modify acls groups to:
20/06/29 17:00:36 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view per
missions: Set(VaraPrasad); groups with view permissions: Set(); users with modify permissions: Set(VaraPrasad); groups
with modify permissions: Set()
20/06/29 17:00:40 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
20/06/29 17:00:40 INFO Master: Starting Spark master at spark://[redacted]:7077
20/06/29 17:00:40 INFO Master: Running Spark version 2.4.4
20/06/29 17:00:40 INFO Utils: Successfully started service 'MasterUI' on port 8080.
20/06/29 17:00:40 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://[redacted]:8080
20/06/29 17:00:41 INFO Master: I have been elected leader! New state: ALIVE

```



Stop loading this page

Spark Master at spark://[redacted]:7077

URL: spark://[redacted]:7077

Alive Workers: 1

Cores in use: 4 Total, 0 Used

Memory in use: 6.9 GB Total, 0.0 B Used

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory
worker-20200630102131-[redacted]0209	[redacted]60209	ALIVE	4 (0 Used)	6.9 GB (0.0 B Used)

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	----------------	------	-------	----------

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	----------------	------	-------	----------

Command to Start a Spark Worker Node:

```
C:\Users\VaraPrasad>spark-class org.apache.spark.deploy.worker.Worker spark://[redacted]:7077
using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
20/06/30 10:21:16 INFO Worker: Started daemon with process name: 11276@Varam
20/06/30 10:21:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
20/06/30 10:21:26 INFO SecurityManager: Changing view acls to: VaraPrasad
20/06/30 10:21:26 INFO SecurityManager: Changing modify acls to: VaraPrasad
20/06/30 10:21:26 INFO SecurityManager: Changing view acls groups to:
20/06/30 10:21:26 INFO SecurityManager: Changing modify acls groups to:
20/06/30 10:21:26 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view per
missions: Set(VaraPrasad); groups with view permissions: Set(); users with modify permissions: Set(VaraPrasad); groups
with modify permissions: Set()
20/06/30 10:21:31 INFO Utils: Successfully started service 'sparkWorker' on port 60209.
20/06/30 10:21:31 INFO Worker: Starting Spark worker [redacted]:60209 with 4 cores, 6.9 GB RAM
20/06/30 10:21:31 INFO Worker: Running Spark version 2.4.4
20/06/30 10:21:31 INFO Worker: Spark home: C:\spark-2.4.4-bin-hadoop2.7
20/06/30 10:21:32 INFO Utils: Successfully started service 'WorkerUI' on port 8081.
20/06/30 10:21:32 INFO WorkerWebUI: Bound WorkerWebUI to 0.0.0.0, and started at http://[redacted]:8081
20/06/30 10:21:32 INFO Worker: Connecting to master [redacted]:7077...
20/06/30 10:21:32 INFO TransportClientFactory: Successfully created connection to [redacted]:7077 after 88 ms (0 ms s
pent in bootstraps)
20/06/30 10:21:33 INFO Worker: Successfully registered with master spark://[redacted]:7077
```

Command to submit the Job:

```
Command Prompt
C:\VB_share>spark-submit --class SparkPageRank --master spark://[redacted]:7077 SparkPageRank.jar
20/06/30 10:24:20 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
20/06/30 10:24:20 INFO SparkContext: Running Spark version 2.4.4
20/06/30 10:24:21 INFO SparkContext: Submitted application: SparkAverage
20/06/30 10:24:21 INFO SecurityManager: Changing view acls to: VaraPrasad
20/06/30 10:24:21 INFO SecurityManager: Changing modify acls to: VaraPrasad
20/06/30 10:24:21 INFO SecurityManager: Changing view acls groups to:
20/06/30 10:24:21 INFO SecurityManager: Changing modify acls groups to:
20/06/30 10:24:21 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view per
missions: Set(VaraPrasad); groups with view permissions: Set(); users with modify permissions: Set(VaraPrasad); groups
with modify permissions: Set()
20/06/30 10:24:23 INFO Utils: Successfully started service 'sparkDriver' on port 60230.
20/06/30 10:24:24 INFO SparkEnv: Registering MapOutputTracker
20/06/30 10:24:24 INFO SparkEnv: Registering BlockManagerMaster
20/06/30 10:24:24 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topo
logy information
20/06/30 10:24:24 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
20/06/30 10:24:24 INFO DiskBlockManager: Created local directory at C:\Users\VaraPrasad\AppData\Local\Temp\blockmgr-6198
de80-9e67-41b5-bd0b-2355f90e4b85
20/06/30 10:24:24 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
20/06/30 10:24:24 INFO SparkEnv: Registering OutputCommitCoordinator
20/06/30 10:24:25 INFO Utils: Successfully started service 'SparkUI' on port 4040.
20/06/30 10:24:25 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://[redacted]:4040
20/06/30 10:24:26 INFO SparkContext: Added JAR file:/C:/VB_share/SparkPageRank.jar at spark://[redacted]:60230/jars/SparkPage
Rank.jar with timestamp 1593527066018
```

Results Screen Shot:

```
Command Prompt
20/06/30 10:24:26 INFO BlockManagerMasterEndpoint: Registering block manager Varam:60243 with 366.3 MB RAM, BlockManager
Id(driver, Varam, 60243, None)
20/06/30 10:24:26 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, Varam, 60243, None)
20/06/30 10:24:26 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, Varam, 60243, None)
links has [4] elements
[(B,[A, D]), (D,[B, C]), (A,[B, C, D]), (C,[C])]
pageRank has [4] elements
[(B,0.25), (D,0.25), (A,0.25), (C,0.25)]
pageRank has [4] elements
[(B,0.2145833333333332), (D,0.2145833333333332), (A,0.14375), (C,0.42708333333333326)]
pageRank has [4] elements
[(B,0.16942708333333334), (D,0.16942708333333334), (A,0.12869791666666666), (C,0.53244791666666666)]
pageRank has [4] elements
[(B,0.1459709201388889), (D,0.1459709201388889), (A,0.10950651041666667), (C,0.5985516493055555)]
pageRank has [4] elements
[(B,0.13056448567708334), (D,0.13056448567708334), (A,0.09953764105902779), (C,0.6393333875868055)]
pageRank has [4] elements
[(B,0.12119223804615165), (D,0.12119223804615165), (A,0.09298990641276042), (C,0.6646256174949362)]
pageRank has [4] elements
[(B,0.11535384131989658), (D,0.11535384131989658), (A,0.08900670116961445), (C,0.6802856161905922)]
pageRank has [4] elements
[(B,0.11174394789234682), (D,0.11174394789234682), (A,0.08652538256095604), (C,0.6899867216543502)]
pageRank has [4] elements
[(B,0.10950670291318494), (D,0.10950670291318494), (A,0.0849911778542474), (C,0.6959954163193824)]
pageRank has [4] elements
[(B,0.1081211824634737), (D,0.1081211824634737), (A,0.0840403487381036), (C,0.6997172863349487)]
pageRank has [4] elements
[(B,0.10726293468943902), (D,0.10726293468943902), (A,0.08345150254697634), (C,0.7020226280741454)]
pageRank has altogether [4] elements
[(B,0.10726293468943902), (D,0.10726293468943902), (A,0.08345150254697634), (C,0.7020226280741454)]
```