

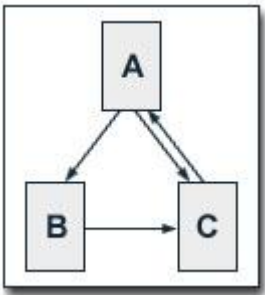
## Technical documentation

This example demonstrates a program written in Scala that implements Web Page Ranking algorithm, which will read web page relations from text file in HDFS and calculate their page ranks.

1 – step:

The following picture a) shows 3 pseudo web pages marked as **A**, **B** and **C** and their relations among each other denoted with arrows. Arrows goes from web page **A** to **B**, this relation means that web page **A** contains link reference to web page **B** and so on. I saved web page relations in the following picture as **WebPageRelations.txt** file in the format as shown in the **b)**.

a)



b)

A B  
A C  
B C  
C A

In the following picture you can see that my Hadoop is up, my **WebPageRelations.txt** file is in HDFS and the contents of **WebPageRelations.txt** file.

```
Terminal
hduser_1@altay-pc ~ $ jps
4570 NameNode
4681 DataNode
5170 NodeManager
5059 ResourceManager
5518 Jps
4891 SecondaryNameNode
hduser_1@altay-pc ~ $ hadoop fs -ls /user/hdfsHomeDir
15/11/03 19:05:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r-- 1 hduser_1 supergroup      15 2015-11-02 21:30 /user/hdfsHomeDir/WebPageRelations.txt
-rw-r--r-- 1 hduser_1 supergroup    2744 2015-10-19 14:20 /user/hdfsHomeDir/hw2_npu.txt
-rw-r--r-- 1 hduser_1 supergroup     54 2015-10-19 14:11 /user/hdfsHomeDir/textInput.txt
hduser_1@altay-pc ~ $ hadoop fs -cat /user/hdfsHomeDir/WebPageRelations.txt
15/11/03 19:05:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
A B
A C
B C
C A
hduser_1@altay-pc ~ $
```

2 – step:

I just run my Scala program in Scala IDE and the result of the program is printed to output console which you can see in the following picture. Each pseudo web page has it's own calculated page rank.



The screenshot shows the Scala IDE interface. The top pane displays the `SparkPageRank.scala` file with the following code:

```
object SparkPageRank {  
  def showWarning() {  
    System.err.println(  
      """WARN: This is a naive implementation of PageRank and is given as an example!  
      |Please use the PageRank implementation found in org.apache.spark.graphx.lib.PageRank  
      |for more conventional use.  
      """).stripMargin)  
  }  
  
  def main(args: Array[String]) {  
    val sparkConf = new SparkConf().setAppName("PageRank")  
      .setMaster("local")  
    val iters = 10  
    val ctx = new SparkContext(sparkConf)  
    val lines = ctx.textFile("hdfs://localhost:54310/user/hdfsHomeDir/WebPageRelations.txt")  
    val links = lines.map { s =>  
      val parts = s.split("\\s+")  
      (parts(0), parts(1))  
    }.distinct().groupByKey().cache()  
    // Display the result of links
```

The bottom pane shows the console output:

```
<terminated> SparkPageRank_config [Scala Application] /usr/lib/jvm/java-7-openjdk-amd64/bin/java (Nov 3, 2015, 7:06:54 PM)  
2015-11-03 19:07:08,750 [main] INFO org.apache.spark.scheduler.DAGScheduler - Job 0 finished: collect at SparkPageRank  
B has rank: 0.6432494117885129.  
A has rank: 1.1667391764027368.  
C has rank: 1.1900114118087488.
```