Hi,

I'm currently testing out Talend's new Natural Language Processing feature on a Spark job on a multiple node MapR cluster. I created a job with 3 components: tFileInputDelimited -> tNLPPreprocessing -> tFileOutputDelimited

The job always results in an error:

```
[statistics] connecting to socket on port 3354
[statistics] connected
[WARN ]: org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where appl
[WARN ]: org.apache.spark.SparkConf - In Spark 1.0 and later spark.local.dir will be overridden by the value set by the cluster manager (via SPARK

[Stage 0:>                                                          (0 + 2) / 2]
[Stage 0:>                                                          (0 + 2) / 2]
[Stage 0:>                                                          (0 + 2) / 2]
[Stage 0:>                                                          (0 + 2) / 2]
[WARN ]: org.apache.spark.scheduler.TaskSetManager - Lost task 1.0 in stage 0.0 (TID 1, 192.168.8.22): org.apache.avro.AvroRuntimeException: Bad i
 at testspark.test_nlp_0_1.row2Struct.put(row2Struct.java:28)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:843)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:1)
 at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.apply(JavaPairRDD.scala:1015)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply$mcV$sp(PairRDDFunctions.scala:1198)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1250)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1205)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1185)
 at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
 at org.apache.spark.scheduler.Task.run(Task.scala:89)
 at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:214)
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
 at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
 at java.lang.Thread.run(Thread.java:748)
[Stage 0:>                                                          (0 + 1) / 2]
[Stage 0:>                                                          (0 + 2) / 2]org.apache.spark.SparkException: Job aborted due to stage failure:
[ERROR]: org.apache.spark.scheduler.TaskSetManager - Task 1 in stage 0.0 failed 4 times; aborting job
 at testspark.test_nlp_0_1.row2Struct.put(row2Struct.java:28)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:843)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:1)
 at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.apply(JavaPairRDD.scala:1015)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply$mcV$sp(PairRDDFunctions.scala:1198)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1250)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1205)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1185)
 at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
 at org.apache.spark.scheduler.Task.run(Task.scala:89)
 at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:214)
```

```
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
 at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
 at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
 at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGScheduler$$failJobAndIndependentStages(DAGScheduler.scala:1431)
 at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1419)
 at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1418)
 at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
 at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
 at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala:1418)
 at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:799)
 at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:799)
 at scala.Option.foreach(Option.scala:236)
 at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed(DAGScheduler.scala:799)
 at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.doOnReceive(DAGScheduler.scala:1640)
 at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1599)
 at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1588)
 at org.apache.spark.util.EventLoop$$anon$1.run(EventLoop.scala:48)
 at org.apache.spark.scheduler.DAGScheduler.runJob(DAGScheduler.scala:620)
 at org.apache.spark.SparkContext.runJob(SparkContext.scala:1832)
 at org.apache.spark.SparkContext.runJob(SparkContext.scala:1845)
 at org.apache.spark.SparkContext.runJob(SparkContext.scala:1922)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1.apply$mcV$sp(PairRDDFunctions.scala:1213)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1.apply(PairRDDFunctions.scala:1156)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1.apply(PairRDDFunctions.scala:1156)
 at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:150)
 at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:111)
 at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
 at org.apache.spark.rdd.PairRDDFunctions.saveAsHadoopDataset(PairRDDFunctions.scala:1156)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopFile$4.apply$mcV$sp(PairRDDFunctions.scala:1060)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopFile$4.apply(PairRDDFunctions.scala:1026)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopFile$4.apply(PairRDDFunctions.scala:1026)
 at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:150)
 at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:111)
 at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
 at org.apache.spark.rdd.PairRDDFunctions.saveAsHadoopFile(PairRDDFunctions.scala:1026)
 at org.apache.spark.api.java.JavaPairRDD.saveAsHadoopFile(JavaPairRDD.scala:780)
 at testspark.test_nlp_0_1.Test_NLP.tFileInputDelimited_1_HDFSInputFormatProcess(Test_NLP.java:1215)
 at testspark.test_nlp_0_1.Test_NLP.run(Test_NLP.java:1651)
 at testspark.test_nlp_0_1.Test_NLP.runJobInTOS(Test_NLP.java:1466)
 at testspark.test_nlp_0_1.Test_NLP.main(Test_NLP.java:1351)
Caused by: org.apache.avro.AvroRuntimeException: Bad index
 at testspark.test_nlp_0_1.row2Struct.put(row2Struct.java:28)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:843)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:1)
 at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.apply(JavaPairRDD.scala:1015)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply$mcV$sp(PairRDDFunctions.scala:1198)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1250)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1205)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1185)
 at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
```

```
 at org.apache.spark.scheduler.Task.run(Task.scala:89)
 at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:214)
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
 at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
 at java.lang.Thread.run(Thread.java:748)
org.apache.spark.SparkException: Job aborted due to stage failure: Task 1 in stage 0.0 failed 4 times, most recent failure: Lost task 1.3 in stage
 at testspark.test_nlp_0_1.row2Struct.put(row2Struct.java:28)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:843)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:1)
 at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.apply(JavaPairRDD.scala:1015)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply$mcV$sp(PairRDDFunctions.scala:1198)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
[statistics] disconnected
 at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1250)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1205)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1185)
 at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
 at org.apache.spark.scheduler.Task.run(Task.scala:89)
 at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:214)
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
 at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
 at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
 at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGScheduler$$failJobAndIndependentStages(DAGScheduler.scala:1431)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1419)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1418)
at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala:1418)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:799)
at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:799)
at scala.Option.foreach(Option.scala:236)
at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed(DAGScheduler.scala:799)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.doOnReceive(DAGScheduler.scala:1640)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1599)
at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onReceive(DAGScheduler.scala:1588)
at org.apache.spark.util.EventLoop$$anon$1.run(EventLoop.scala:48)
at org.apache.spark.scheduler.DAGScheduler.runJob(DAGScheduler.scala:620)
at org.apache.spark.SparkContext.runJob(SparkContext.scala:1832)
at org.apache.spark.SparkContext.runJob(SparkContext.scala:1845)
at org.apache.spark.SparkContext.runJob(SparkContext.scala:1922)
at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1.apply$mcV$sp(PairRDDFunctions.scala:1213)
at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1.apply(PairRDDFunctions.scala:1156)
at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1.apply(PairRDDFunctions.scala:1156)
at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:150)
at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:111)
at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
at org.apache.spark.rdd.PairRDDFunctions.saveAsHadoopDataset(PairRDDFunctions.scala:1156)
at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopFile$4.apply$mcV$sp(PairRDDFunctions.scala:1060)
at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopFile$4.apply(PairRDDFunctions.scala:1026)
at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopFile$4.apply(PairRDDFunctions.scala:1026)
at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:150)
```

```
 at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:111)
 at org.apache.spark.rdd.RDD.withScope(RDD.scala:316)
 at org.apache.spark.rdd.PairRDDFunctions.saveAsHadoopFile(PairRDDFunctions.scala:1026)
 at org.apache.spark.api.java.JavaPairRDD.saveAsHadoopFile(JavaPairRDD.scala:780)
 at testspark.test_nlp_0_1.Test_NLP.tFileInputDelimited_1_HDFSInputFormatProcess(Test_NLP.java:1215)
 at testspark.test_nlp_0_1.Test_NLP.run(Test_NLP.java:1651)
 at testspark.test_nlp_0_1.Test_NLP.runJobInTOS(Test_NLP.java:1466)
 at testspark.test_nlp_0_1.Test_NLP.main(Test_NLP.java:1351)
Caused by: org.apache.avro.AvroRuntimeException: Bad index
 at testspark.test_nlp_0_1.row2Struct.put(row2Struct.java:28)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:843)
 at testspark.test_nlp_0_1.Test_NLP$tNLPPreprocessing_1Preprocessing.call(Test_NLP.java:1)
 at org.apache.spark.api.java.JavaPairRDD$$anonfun$toScalaFunction$1.apply(JavaPairRDD.scala:1015)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at scala.collection.Iterator$$anon$11.next(Iterator.scala:328)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply$mcV$sp(PairRDDFunctions.scala:1198)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13$$anonfun$apply$6.apply(PairRDDFunctions.scala:1197)
 at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:1250)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1205)
 at org.apache.spark.rdd.PairRDDFunctions$$anonfun$saveAsHadoopDataset$1$$anonfun$13.apply(PairRDDFunctions.scala:1185)
 at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:66)
 at org.apache.spark.scheduler.Task.run(Task.scala:89)
 at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:214)
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
 at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
 at java.lang.Thread.run(Thread.java:748)
Exception in thread "main" java.lang.RuntimeException: TalendJob: 'Test_NLP' - Failed with exit code: 1.
 at testspark.test_nlp_0_1.Test_NLP.main(Test_NLP.java:1361)
[ERROR]: testspark.test_nlp_0_1.Test_NLP - TalendJob: 'Test_NLP' - Failed with exit code: 1.
Job Test_NLP ended at 16:18 05/10/2017. [exit code=1]
```

I've tried running other Spark jobs with non-NLP components, so Spark configuration shouldn't be an issue.

Are there any rules/formats which I need to adjust in order to use tNLPPreprocessing? Or does the problem lie somewhere else?

Thank you.