Error Saving Pipeline Containing DrugNormalizer()

4/21/2021

Spark Connect

```
config <- spark_config()
config$sparklyr.log.console <- FALSE
config$spark.sql.shuffle.partitions <- 800
sc <- spark_connect(master = "local", config = config, version = "3.0.1")
```

```
## Warning in normalizePath(unlist(unique(all_jars))): path[2]="https://
## pypi.johnsnowlabs.com/3.0.1-55ff3c9b188fcf797bd517100a553b6cf23d0fd9/spark-nlp-
## jsl-3.0.1.jar": No such file or directory
```

Empty Dataframe

```
empty_df <- data.frame(matrix(ncol = 1, nrow = 0))
colnames(empty_df) <- c("text_cleaner")
empty_spark_df <- sparklyr::sdf_copy_to(sc, empty_df, "empty_spark_df", memory = TRUE, overwrite
= TRUE)
```

Pipeline

```
document_assembler <-
  sparknlp::nlp_document_assembler(
    sc,
    input_col = "text_cleaner",
    output_col = "document")

drug_normalizer <-
  sparknlp::nlp_drug_normalizer(
    sc,
    input_cols = "document",
    output_col = "document_normalized",
    policy = "all")

entity_pipeline <-
  sparklyr::ml_pipeline(
    document_assembler,
    drug_normalizer)

entity_detection_pipeline <-
  sparklyr::ml_fit(entity_pipeline, empty_spark_df)
```

Save

```
sparklyr::ml_save(entity_detection_pipeline,
                  glue("{Home}/drug-norm-pipeline"),
                  overwrite = TRUE)
```

```
## Error: org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory file:/home/katie
g/drug-norm-pipeline/stages/1_drug_normalizer__9cc4a44a_459f_4d80_a2aa_786ce3563444/fields/patte
rns already exists
##  at org.apache.hadoop.mapred.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:131)
##  at org.apache.spark.internal.io.HadoopMapRedWriteConfigUtil.assertConf(SparkHadoopWriter.sca
la:289)
##  at org.apache.spark.internal.io.SparkHadoopWriter$.write(SparkHadoopWriter.scala:71)
##  at org.apache.spark.rdd.PairRDDFunctions.$anonfun$saveAsHadoopDataset$1(PairRDDFunctions.sca
la:1090)
##  at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:151)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:112)
##  at org.apache.spark.rdd.RDD.withScope(RDD.scala:388)
##  at org.apache.spark.rdd.PairRDDFunctions.saveAsHadoopDataset(PairRDDFunctions.scala:1088)
##  at org.apache.spark.rdd.PairRDDFunctions.$anonfun$saveAsHadoopFile$4(PairRDDFunctions.scala:
1061)
##  at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:151)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:112)
##  at org.apache.spark.rdd.RDD.withScope(RDD.scala:388)
##  at org.apache.spark.rdd.PairRDDFunctions.saveAsHadoopFile(PairRDDFunctions.scala:1026)
##  at org.apache.spark.rdd.SequenceFileRDDFunctions.$anonfun$saveAsSequenceFile$1(SequenceFileR
DDFunctions.scala:69)
##  at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:151)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:112)
##  at org.apache.spark.rdd.RDD.withScope(RDD.scala:388)
##  at org.apache.spark.rdd.SequenceFileRDDFunctions.saveAsSequenceFile(SequenceFileRDDFunction
s.scala:54)
##  at org.apache.spark.rdd.RDD.$anonfun$saveAsObjectFile$1(RDD.scala:1561)
##  at scala.runtime.java8.JFunction0$mcV$sp.apply(JFunction0$mcV$sp.java:23)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:151)
##  at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:112)
##  at org.apache.spark.rdd.RDD.withScope(RDD.scala:388)
##  at org.apache.spark.rdd.RDD.saveAsObjectFile(RDD.scala:1561)
##  at com.johnsnowlabs.nlp.serialization.MapFeature.serializeObject(Feature.scala:151)
##  at com.johnsnowlabs.nlp.serialization.MapFeature.serializeObject(Feature.scala:144)
##  at com.johnsnowlabs.nlp.serialization.Feature.serialize(Feature.scala:34)
##  at com.johnsnowlabs.nlp.serialization.Feature.serializeInfer(Feature.scala:40)
##  at com.johnsnowlabs.nlp.FeaturesWriter.$anonfun$saveImpl$1(ParamsAndFeaturesWritable.scala:1
5)
##  at com.johnsnowlabs.nlp.FeaturesWriter.$anonfun$saveImpl$1$adapted(ParamsAndFeaturesWritabl
e.scala:13)
##  at scala.collection.mutable.ResizableArray.foreach(ResizableArray.scala:62)
##  at scala.collection.mutable.ResizableArray.foreach$(ResizableArray.scala:55)
##  at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:49)
##  at com.johnsnowlabs.nlp.FeaturesWriter.saveImpl(ParamsAndFeaturesWritable.scala:13)
##  at org.apache.spark.ml.util.MLWriter.save(ReadWrite.scala:168)
##  at org.apache.spark.ml.Pipeline$SharedReadWrite$.$anonfun$saveImpl$5(Pipeline.scala:257)
##  at org.apache.spark.ml.MLEvents.withSaveInstanceEvent(events.scala:176)
##  at org.apache.spark.ml.MLEvents.withSaveInstanceEvent$(events.scala:171)
##  at org.apache.spark.ml.util.Instrumentation.withSaveInstanceEvent(Instrumentation.scala:42)
##  at org.apache.spark.ml.Pipeline$SharedReadWrite$.$anonfun$saveImpl$4(Pipeline.scala:257)
```

```
##  at org.apache.spark.ml.Pipeline$SharedReadWrite$.$anonfun$saveImpl$4$adapted(Pipeline.scala:
254)
##  at scala.collection.IndexedSeqOptimized.foreach(IndexedSeqOptimized.scala:36)
##  at scala.collection.IndexedSeqOptimized.foreach$(IndexedSeqOptimized.scala:33)
##  at scala.collection.mutable.ArrayOps$ofRef.foreach(ArrayOps.scala:198)
##  at org.apache.spark.ml.Pipeline$SharedReadWrite$.$anonfun$saveImpl$1(Pipeline.scala:254)
##  at org.apache.spark.ml.Pipeline$SharedReadWrite$.$anonfun$saveImpl$1$adapted(Pipeline.scala:
247)
##  at org.apache.spark.ml.util.Instrumentation$.$anonfun$instrumented$1(Instrumentation.scala:1
91)
##  at scala.util.Try$.apply(Try.scala:213)
##  at org.apache.spark.ml.util.Instrumentation$.instrumented(Instrumentation.scala:191)
##  at org.apache.spark.ml.Pipeline$SharedReadWrite$.saveImpl(Pipeline.scala:247)
##  at org.apache.spark.ml.PipelineModel$PipelineModelWriter.saveImpl(Pipeline.scala:346)
##  at org.apache.spark.ml.util.MLWriter.save(ReadWrite.scala:168)
##  at org.apache.spark.ml.PipelineModel$PipelineModelWriter.super$save(Pipeline.scala:344)
##  at org.apache.spark.ml.PipelineModel$PipelineModelWriter.$anonfun$save$4(Pipeline.scala:344)
##  at org.apache.spark.ml.MLEvents.withSaveInstanceEvent(events.scala:176)
##  at org.apache.spark.ml.MLEvents.withSaveInstanceEvent$(events.scala:171)
##  at org.apache.spark.ml.util.Instrumentation.withSaveInstanceEvent(Instrumentation.scala:42)
##  at org.apache.spark.ml.PipelineModel$PipelineModelWriter.$anonfun$save$3(Pipeline.scala:344)
##  at org.apache.spark.ml.PipelineModel$PipelineModelWriter.$anonfun$save$3$adapted(Pipeline.sc
ala:344)
##  at org.apache.spark.ml.util.Instrumentation$.$anonfun$instrumented$1(Instrumentation.scala:1
91)
##  at scala.util.Try$.apply(Try.scala:213)
##  at org.apache.spark.ml.util.Instrumentation$.instrumented(Instrumentation.scala:191)
##  at org.apache.spark.ml.PipelineModel$PipelineModelWriter.save(Pipeline.scala:344)
##  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
##  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
##  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
##  at java.lang.reflect.Method.invoke(Method.java:498)
##  at sparklyr.Invoke.invoke(invoke.scala:147)
##  at sparklyr.StreamHandler.handleMethodCall(stream.scala:136)
##  at sparklyr.StreamHandler.read(stream.scala:61)
##  at sparklyr.BackendHandler.$anonfun$channelRead0$1(handler.scala:58)
##  at scala.util.control.Breaks.breakable(Breaks.scala:42)
##  at sparklyr.BackendHandler.channelRead0(handler.scala:39)
##  at sparklyr.BackendHandler.channelRead0(handler.scala:14)
##  at io.netty.channel.SimpleChannelInboundHandler.channelRead(SimpleChannelInboundHandler.jav
a:99)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:379)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:365)
##  at io.netty.channel.AbstractChannelHandlerContext.fireChannelRead(AbstractChannelHandlerCont
ext.java:357)
##  at io.netty.handler.codec.MessageToMessageDecoder.channelRead(MessageToMessageDecoder.java:1
02)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:379)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:365)
##  at io.netty.channel.AbstractChannelHandlerContext.fireChannelRead(AbstractChannelHandlerCont
```

```
ext.java:357)
##  at io.netty.handler.codec.ByteToMessageDecoder.fireChannelRead(ByteToMessageDecoder.java:32
1)
##  at io.netty.handler.codec.ByteToMessageDecoder.channelRead(ByteToMessageDecoder.java:295)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:379)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:365)
##  at io.netty.channel.AbstractChannelHandlerContext.fireChannelRead(AbstractChannelHandlerCont
ext.java:357)
##  at io.netty.channel.DefaultChannelPipeline$HeadContext.channelRead(DefaultChannelPipeline.ja
va:1410)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:379)
##  at io.netty.channel.AbstractChannelHandlerContext.invokeChannelRead(AbstractChannelHandlerCo
ntext.java:365)
##  at io.netty.channel.DefaultChannelPipeline.fireChannelRead(DefaultChannelPipeline.java:919)
##  at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.jav
a:163)
##  at io.netty.channel.n
```