1、[ERROR] Failed to execute goal org.apache.maven.plugins:maven-compiler-plugin:3.5:compile (default-compile) on project easyml-manage: Compilation failure: Compilation failure:

[ERROR] /D:/gyl/gitee/EasyMl/easyml-manage/src/main/java/com/trusfort/easyml/controller/interceptor/StartupListener.java:[3,40] 找不到符号

[ERROR] 符号: 类 JobService

[ERROR] 位置: 程序包 com.trusfort.easyml.service.task

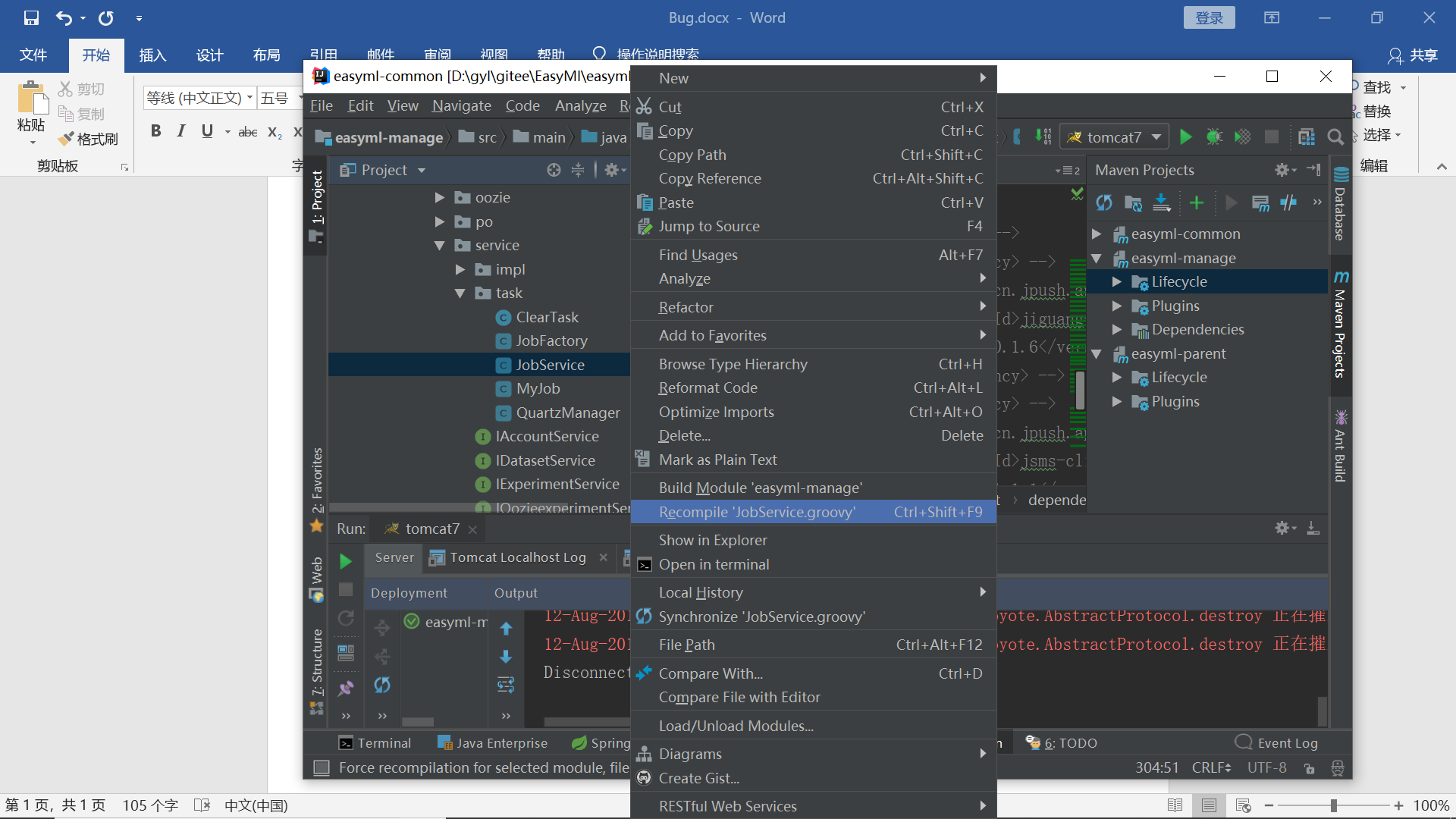
[ERROR] /D:/gyl/gitee/EasyMl/easyml-manage/src/main/java/com/trusfort/easyml/controller/interceptor/StartupListener.java:[22,5] 找不到符号

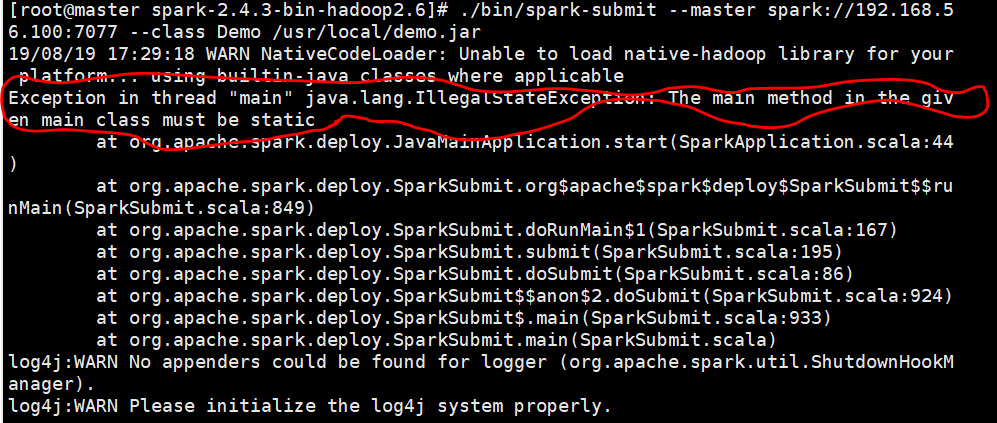
[ERROR] 符号: 类 JobService

[ERROR] 位置: 类 com.trusfort.easyml.controller.interceptor.StartupListener

[ERROR] -> [Help 1]

解决方案：找到项目中JobService类的位置，右键该类，点击Recompile即可。



2、

这是因为scala程序的主类是这样声明的：

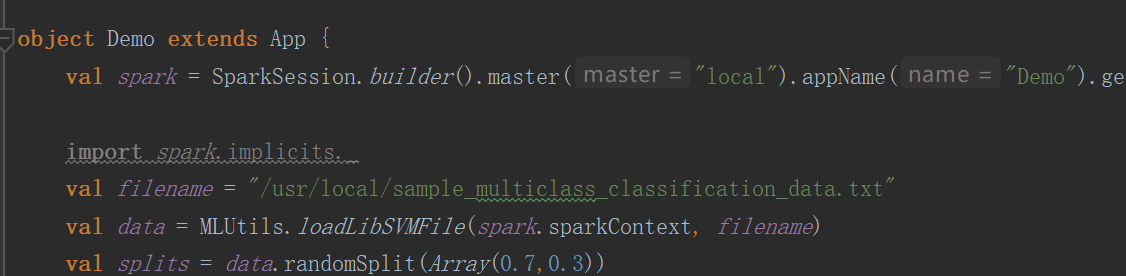
Class Demo {

def main(args : Array[String]) {

}

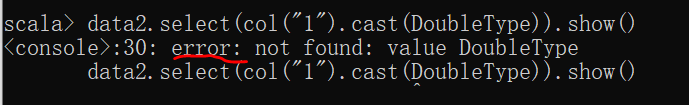
}

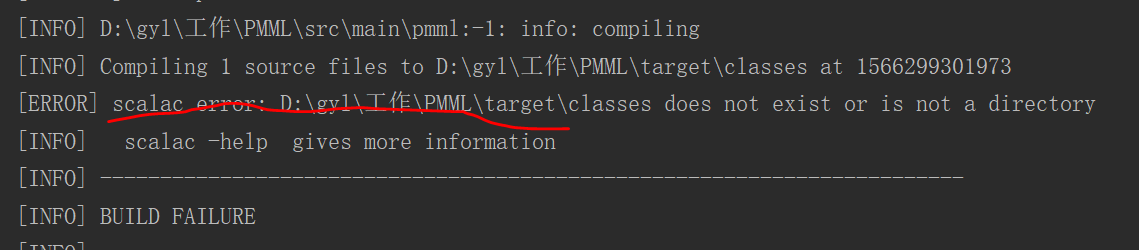
要把它改成这样：



3、下面的bug是因为没有写这句话：

Import org.apache.spark.sql.types.\_

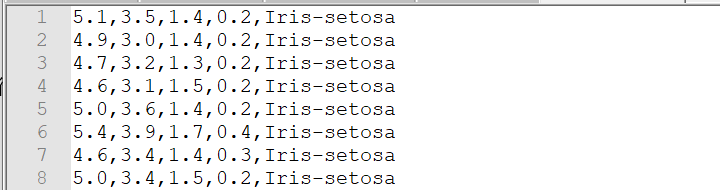


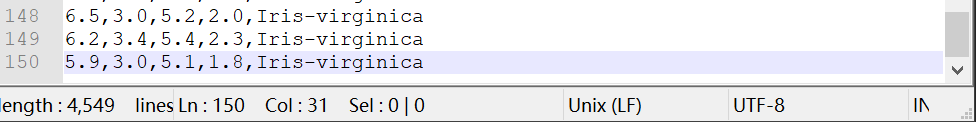
4、

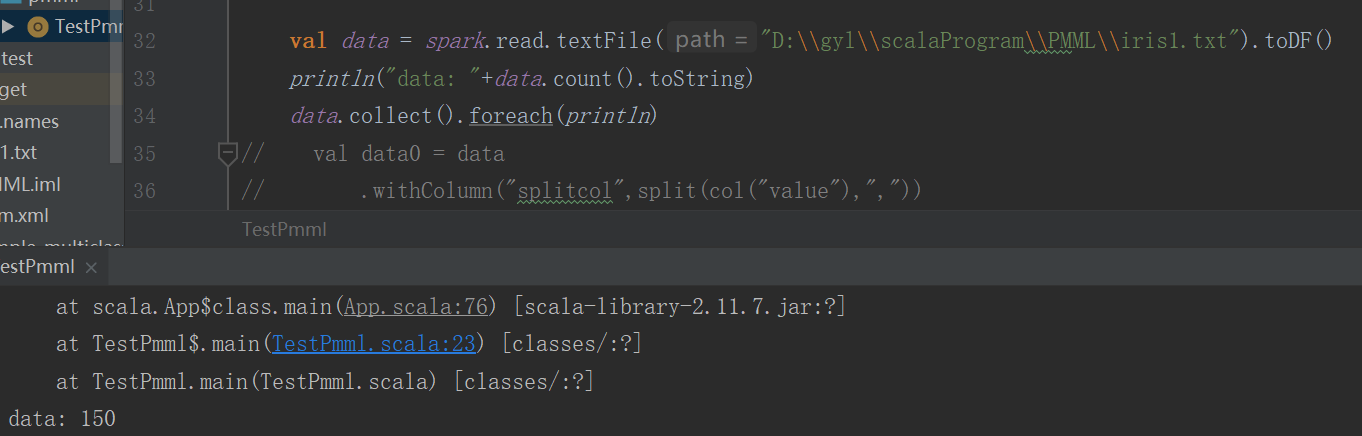
错误原因：工作目录中含有中文。

4、spark中textFile读取txt文件时，txt文件内容行数与读取出来进行count后的行数不一致。

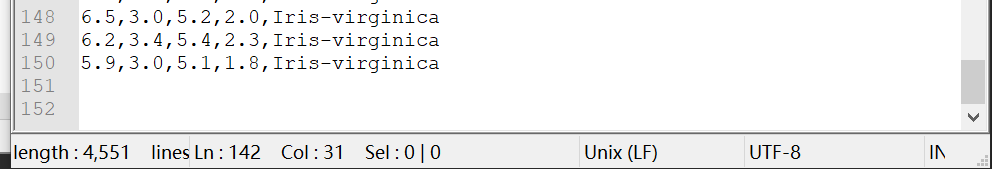
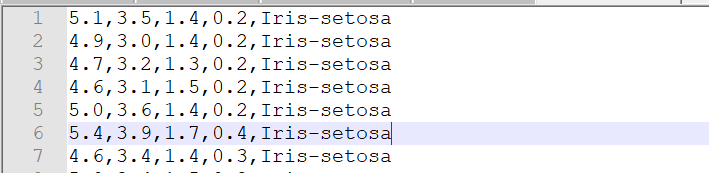
（1）txt文件最后没有空行，总共有150行

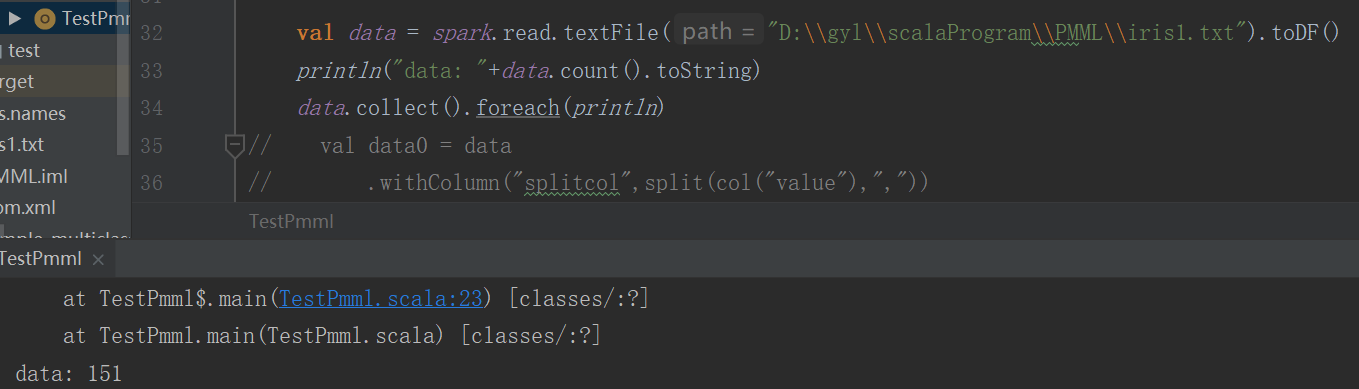






（2）txt文件最后有空行，spark中textFile读取了151行



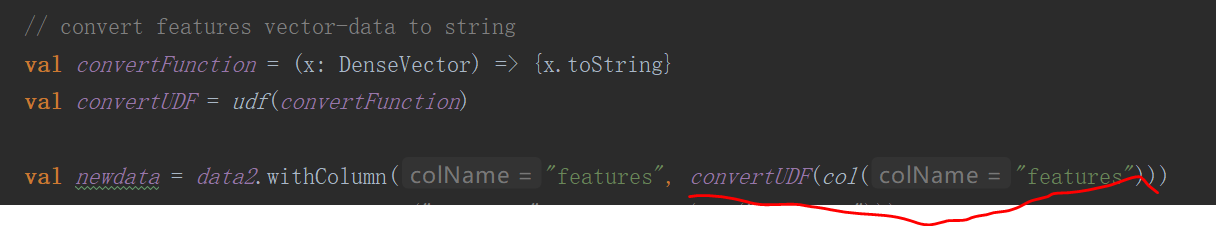


再看看数据，最后一个空行被toDF操作转换为[]

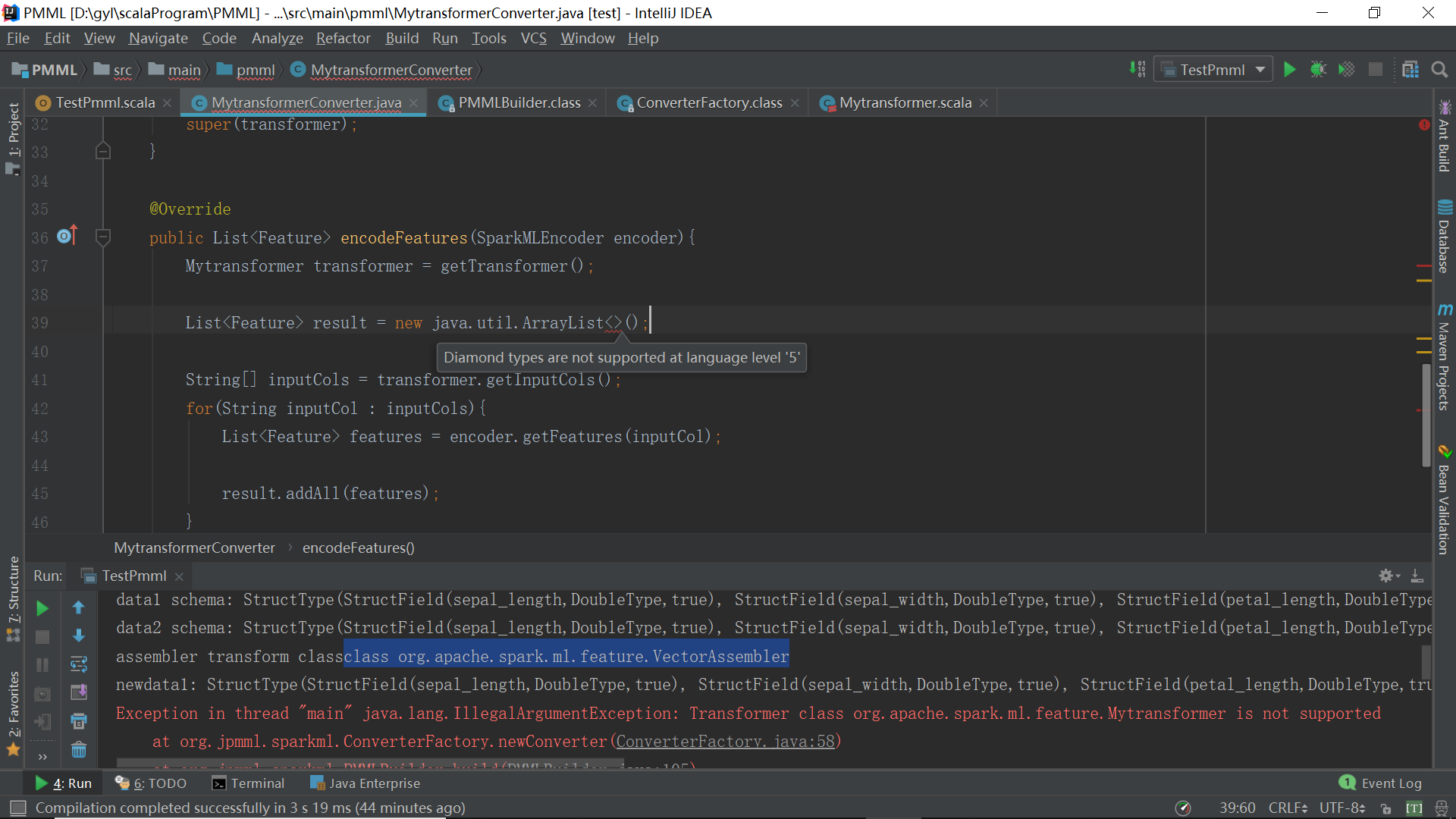


5、spark的dataframe中withColumn里函数不要写成这样col(“features”).apply(convertUDF)，会出问题（这里的apply的用法不同）

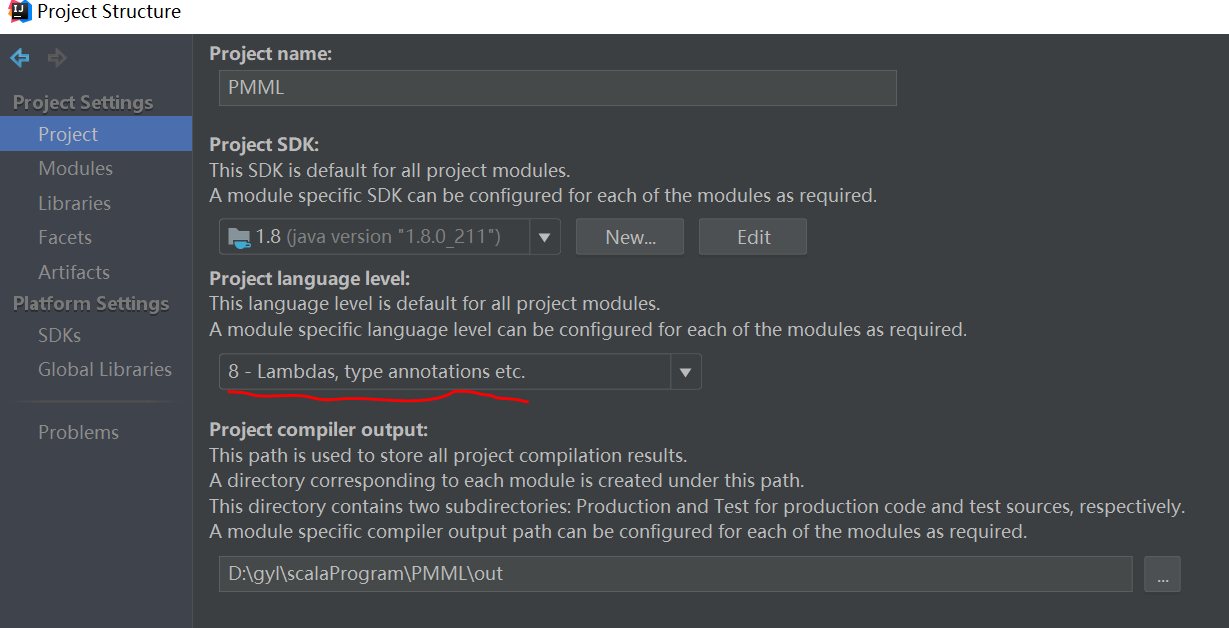
要写成下面这样就没问题：



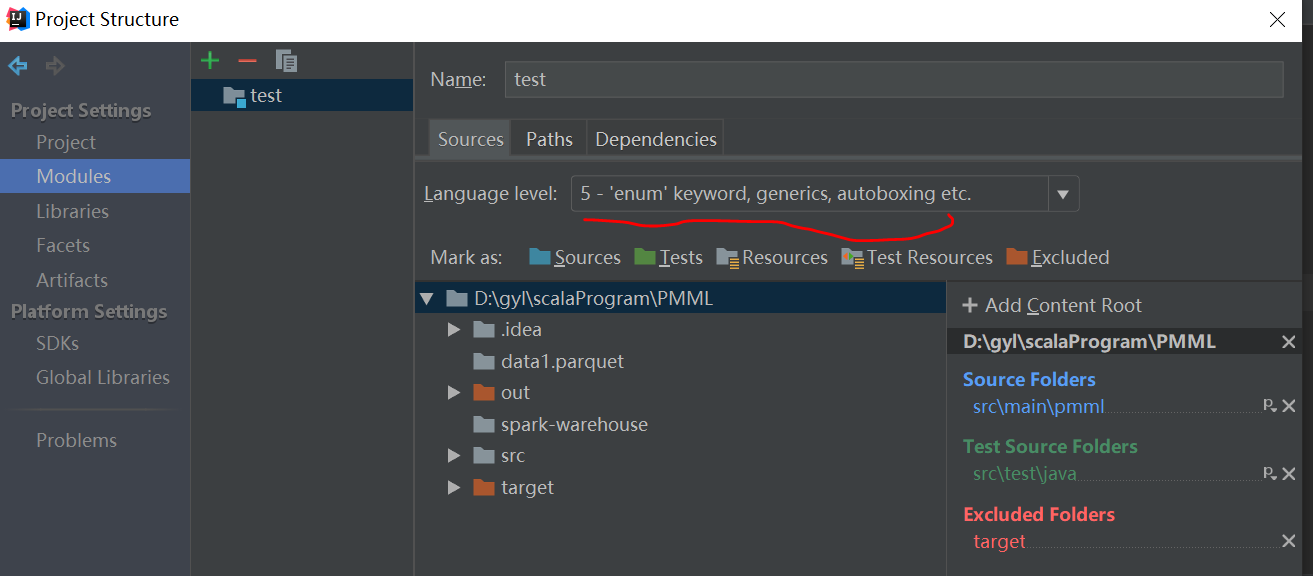
6、Idea中JDK为1.8，还提示Diamond types are not supported at this language level。



解决方法：通过查看项目设置，发现project的java level 也是8。

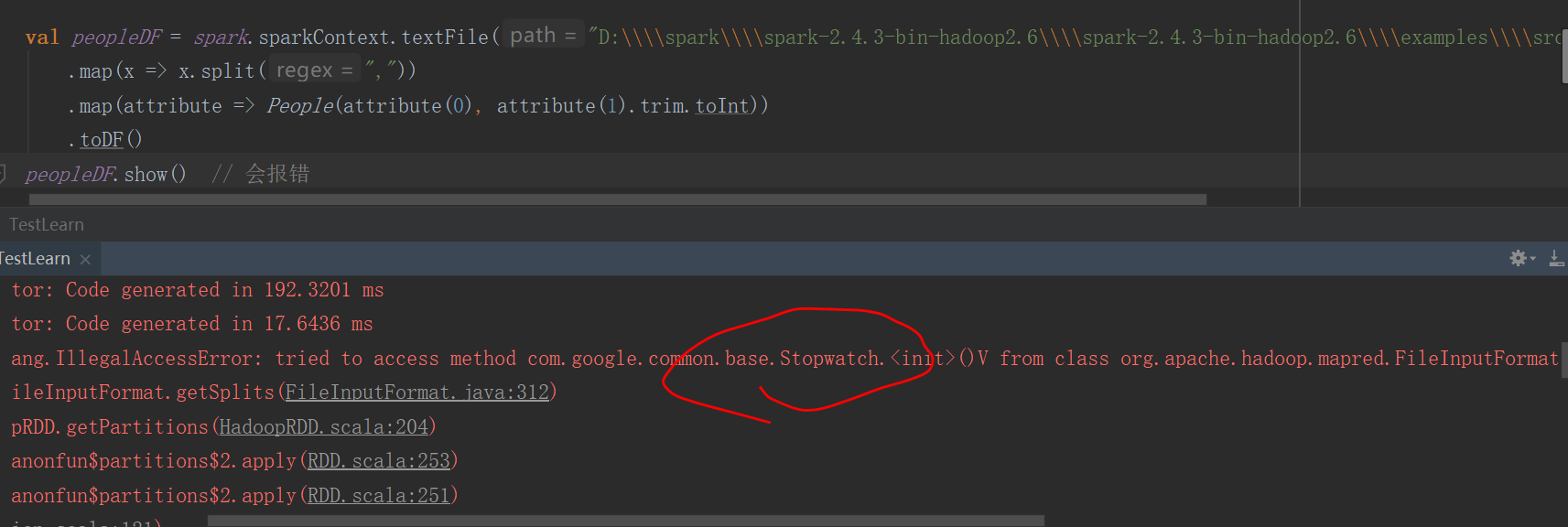


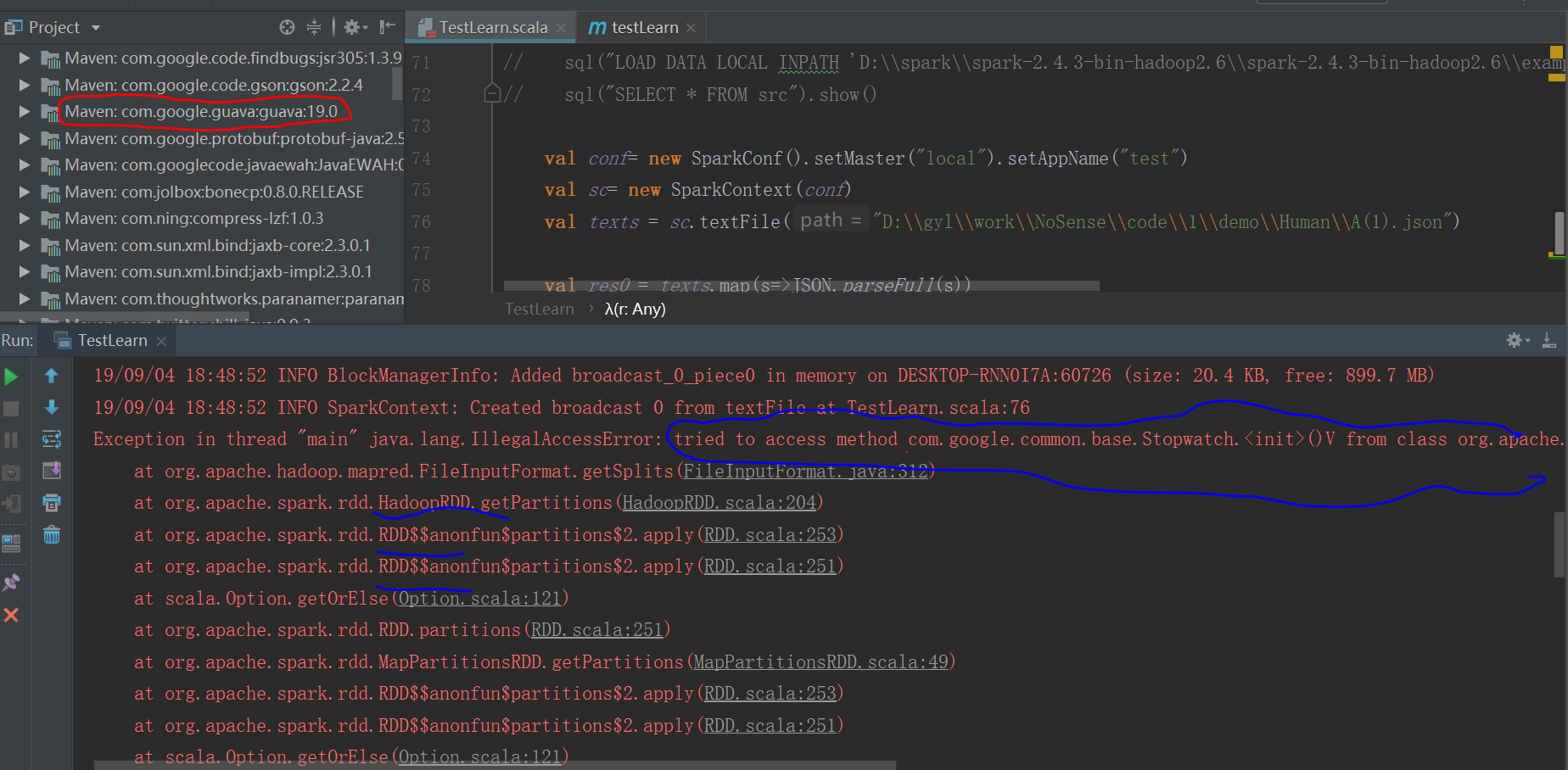
然后继续检查其他模块 如modules ，发现了问题所在。



将其改为8就可以了。

7、一个非常简单的程序，出现了一个这样的错误：





解决办法：

1. 网上(<https://www.cnblogs.com/ylcoder/p/6285041.html>)说是hdfs目录下有相关的hadoop jar包，删掉就能用，试了一下，并没有用。
2. 网上(<https://cloud.tencent.com/developer/article/1398995>)说是jar包冲突，删除本地guava库（这种解决方法是存在多个版本guava库的时候，而我这里只有一个版本，并且我也没有进行hbase查询，所以并没有用）。
3. 这个才是真正的答案(<http://www.luyixian.cn/news_show_4252.aspx>)，导致错误的原因的确是guava包的问题。查看pom.xml文件，发现自己并没有写guava的依赖，所以应该是别的地方引用到了。然后去idea的依赖包管理界面external Libraries查看存在版本为19.0的guava包，导致这个错误的原因就是版本19.0的guava包和spark不兼容，于是在pom.xml文件中添加以下依赖：

<dependency>

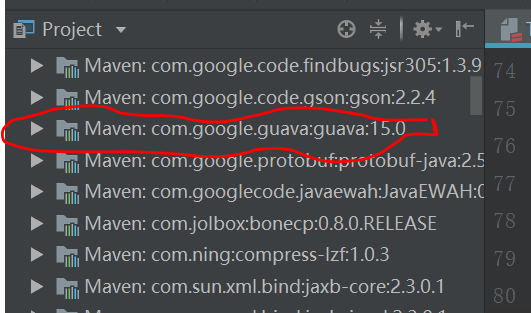
<groupId>com.google.guava</groupId>

<artifactId>guava</artifactId>

<version>15.0</version>

</dependency>

等依赖加载完后，能看到external Libraries里guava包的版本变为15.0，再次运行，发现不会报错了。



8、错误：Unable to instantiate SparkSession with Hive support because Hive classes are not found.

解决方法：在pom.xml文件中添加依赖：

<dependency>

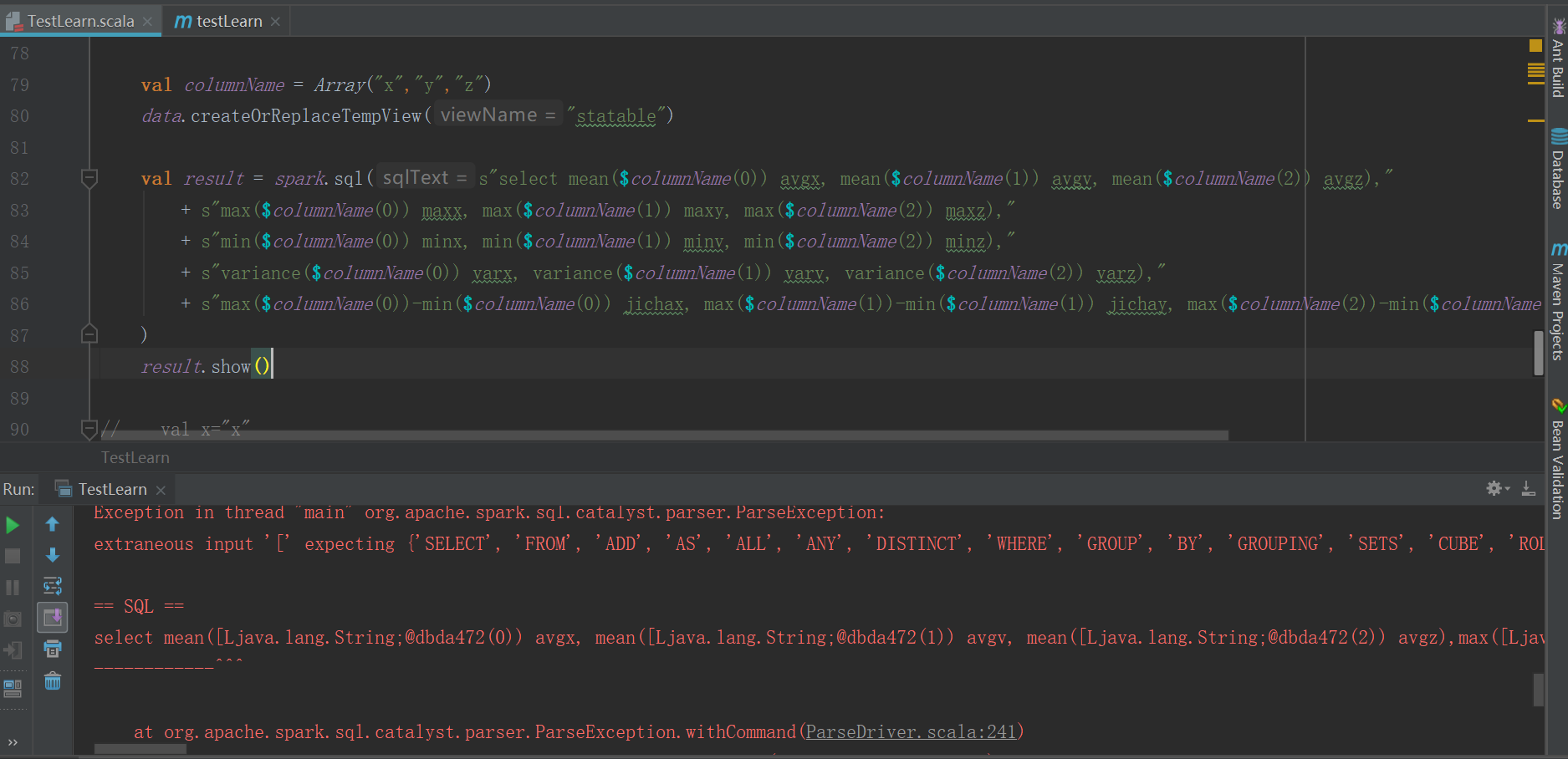
<groupId>org.apache.spark</groupId>

<artifactId>spark-hive\_2.11</artifactId>

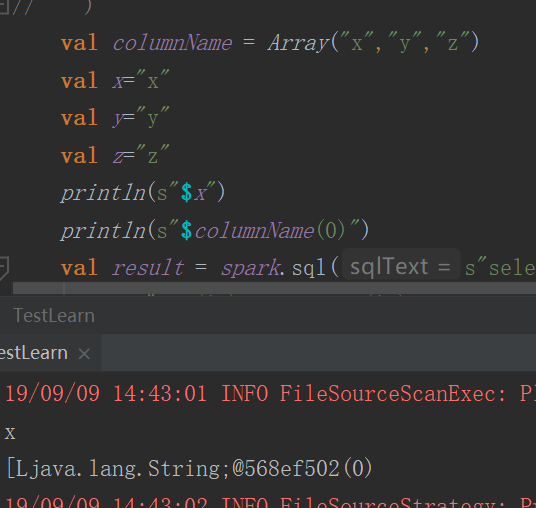
<version>${spark.version}</version>

</dependency>

9、在对dataframe的每列计算统计特征时，报如下错误：

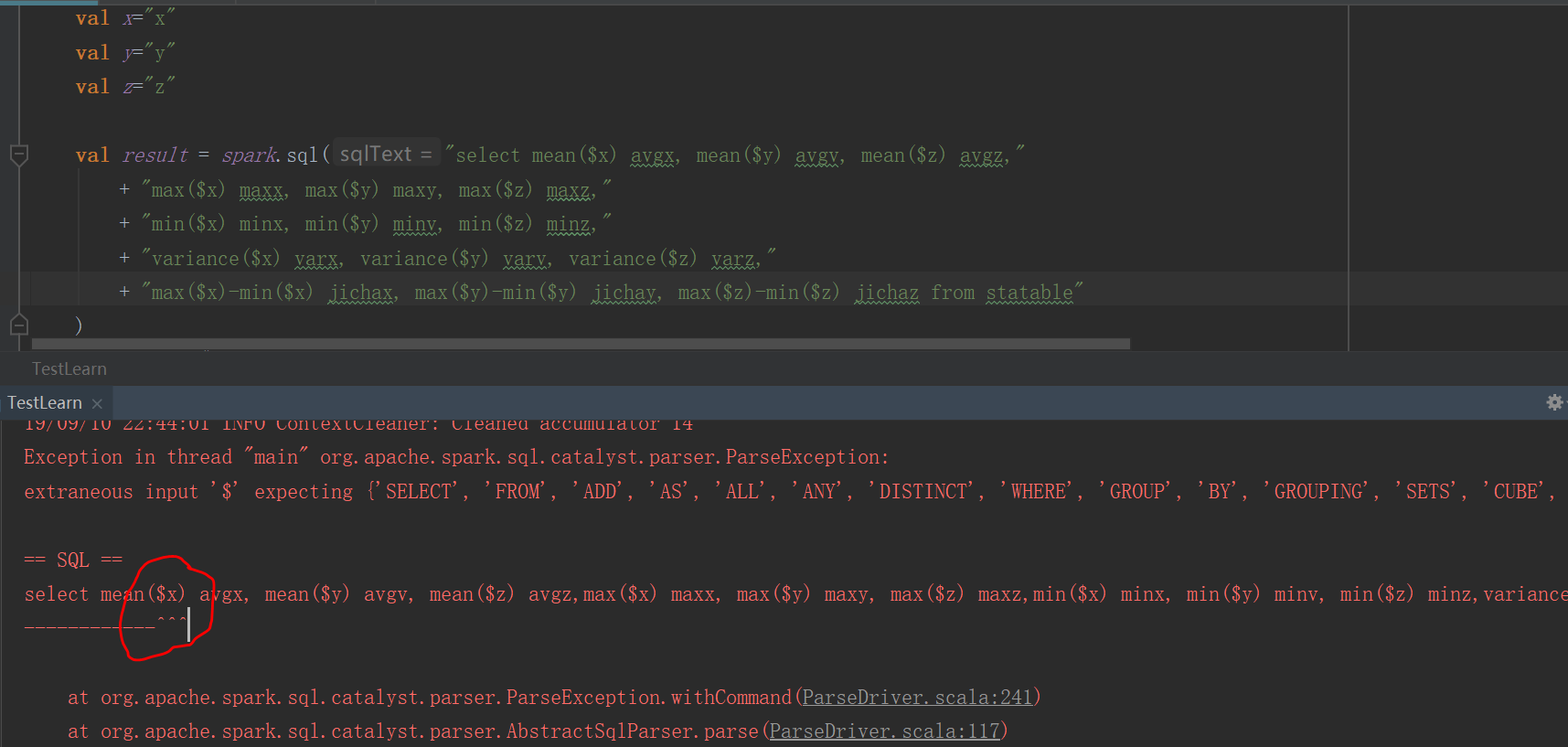


错误原因：scala中“$”后面接一个变量名以提取变量值，而这里columnName(0)的值是字符串。



解决方案：将”$”去掉也没用，因为将“$”去掉后，在双引号里面的columnName(0)就表示一个字符串，而不是”x”。最后的解决方法是分别定义，然后用”$”引用。

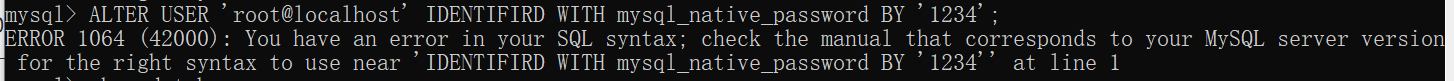
10、关于spark.sql中字符串前面s的问题，如果不加s，则会报错：无法解析变量



解决方法：在字符串前面加个s

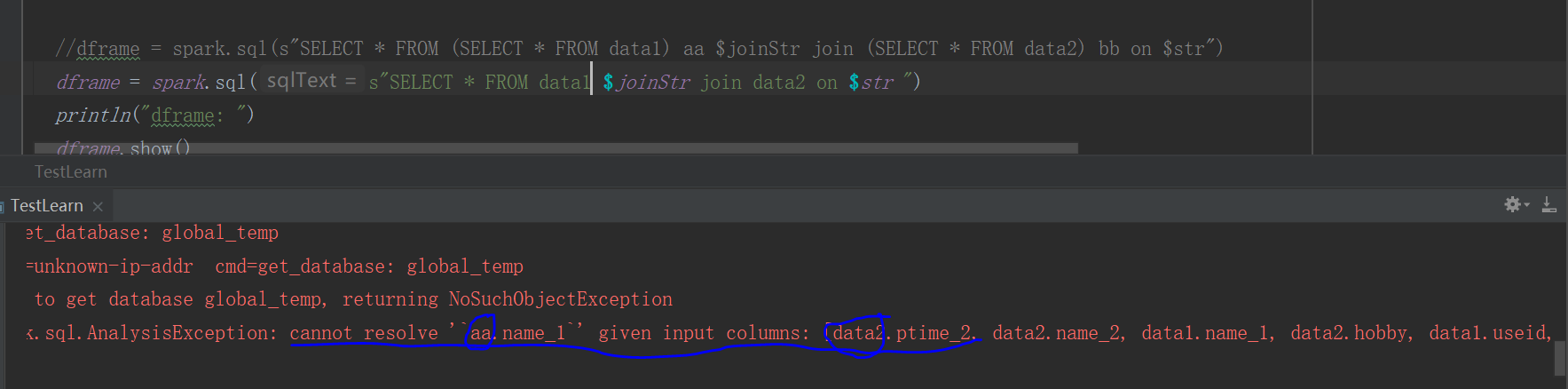


11、mysql中第一次修改密码出现的问题：



出现该错误的原因是root@localhost加了单引号，将该单引号去掉就ok，即：ALTER USER root@localhost IDENTIFIED  BY '1234'; 。(注意：最后的分号一定要有)

12、sql问题：出现这个问题的主要原因是$str中的值与data1和data2的列不一样，在str中数据表是aa和bb，而实际给的是data1和data2.



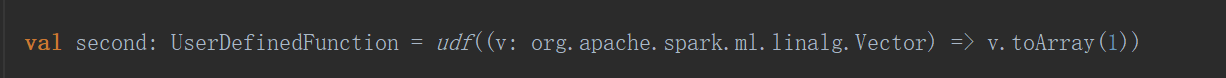


解决方法：为表加别名。

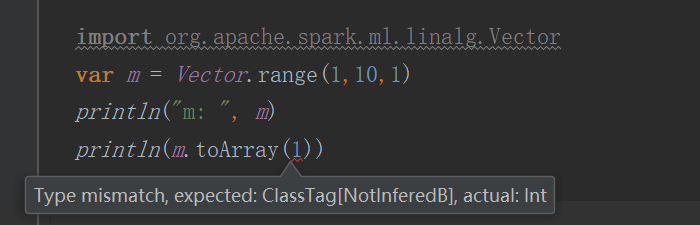


13、scala中vector和spark中vector的问题：

问题背景：看代码时发现toArray带了参数，而实际spark的API中Vector.toArray是不带参数的，如下图，竟然没有报错：

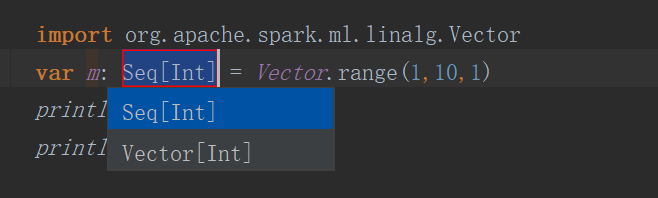


然后自己敲了一下，发现toArray带了参数报错，如下图，为什么会这样？

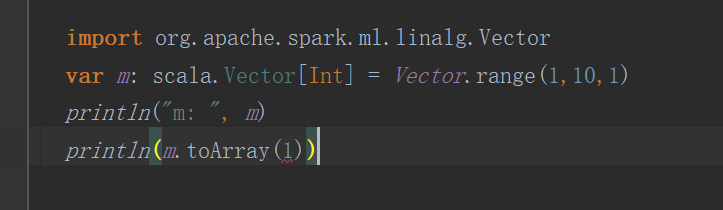


然后，在变量m处按快捷键Alt+Enter查看变量类型，如下图，有两种类型，无论选哪种类型，带参数的toArray都会报错。

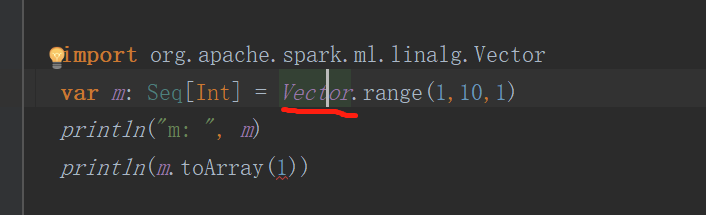
类型1：Seq[Int]



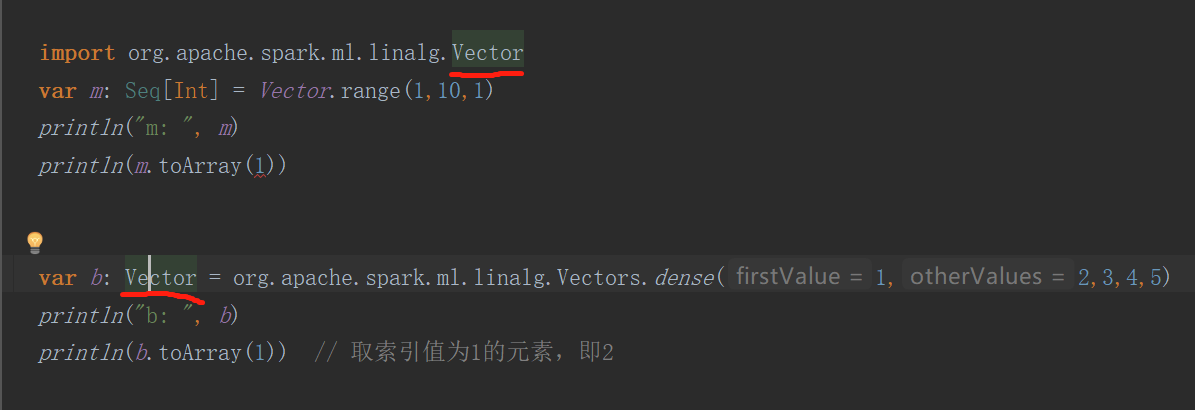
类型2：scala.Vector[Int]



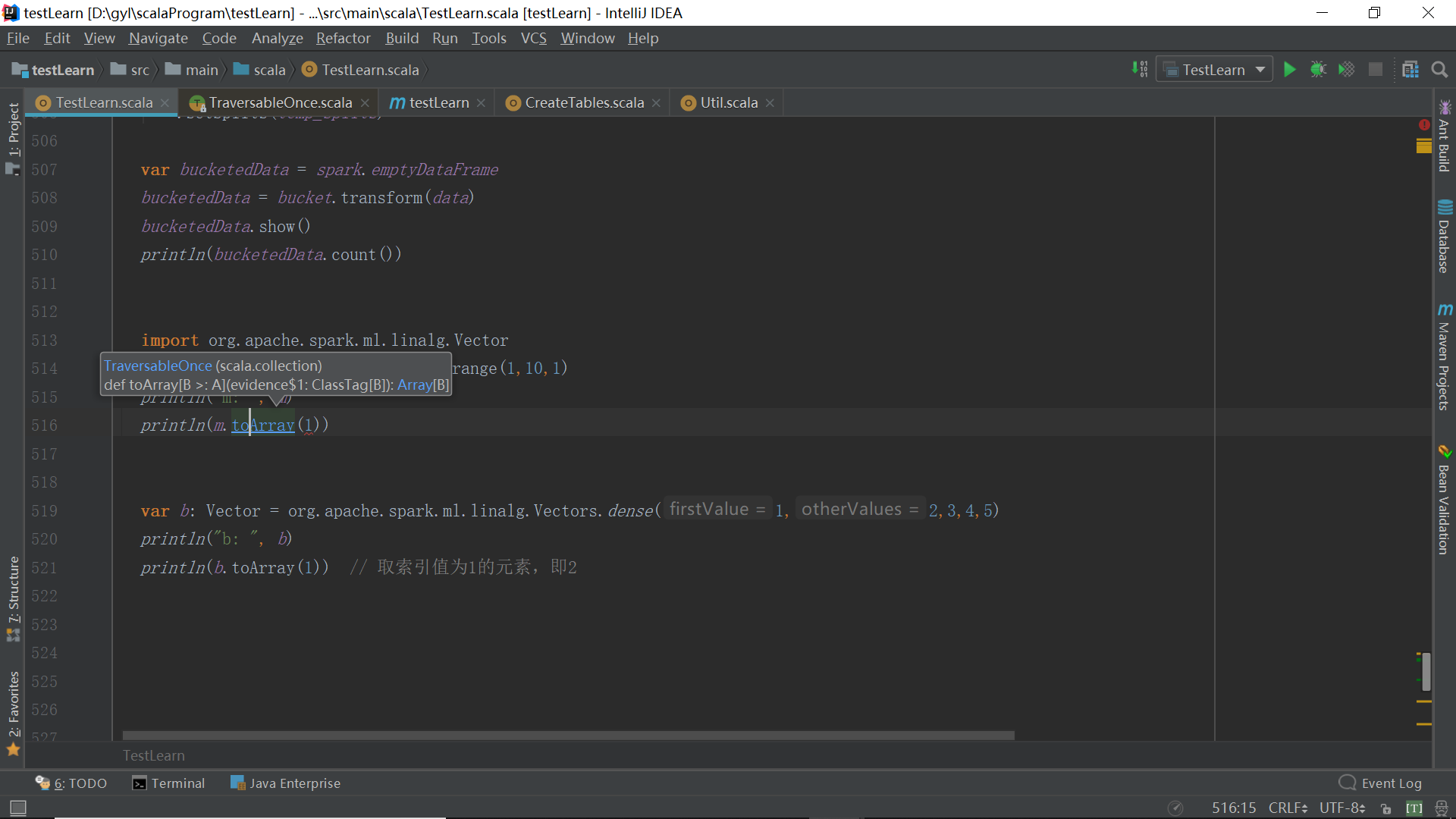
然后，点击Vector发现该Vector并不是import进来的Vector，如下图



于是，直接用org.apache.spark.ml.linalg.Vector定义一个Vector变量，如下图

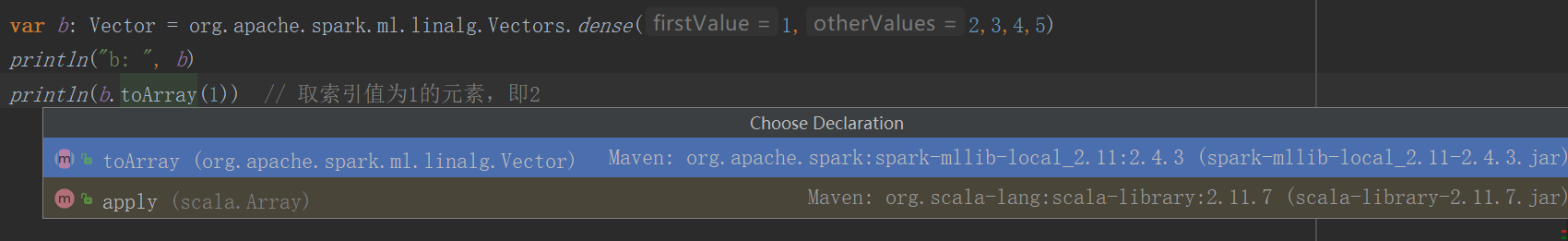


为什么直接把包名带上定义的变量b是Vector型，而变量m是Seq[Int]或scala.Vector[Int]。是因为变量m里的Vector实际是scala中的容器，并不是spark中的vector。在m.toArray(1)上按Ctrl+鼠标点击toArray也可以看出该toArray函数是容器中的函数，如下图



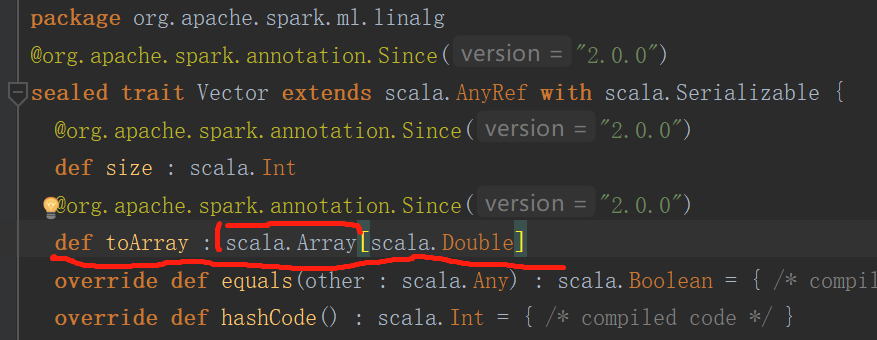
从上面可以看出，即使将spark中Vector给import进来，但是直接写vector还是调用的scala中的容器vector。

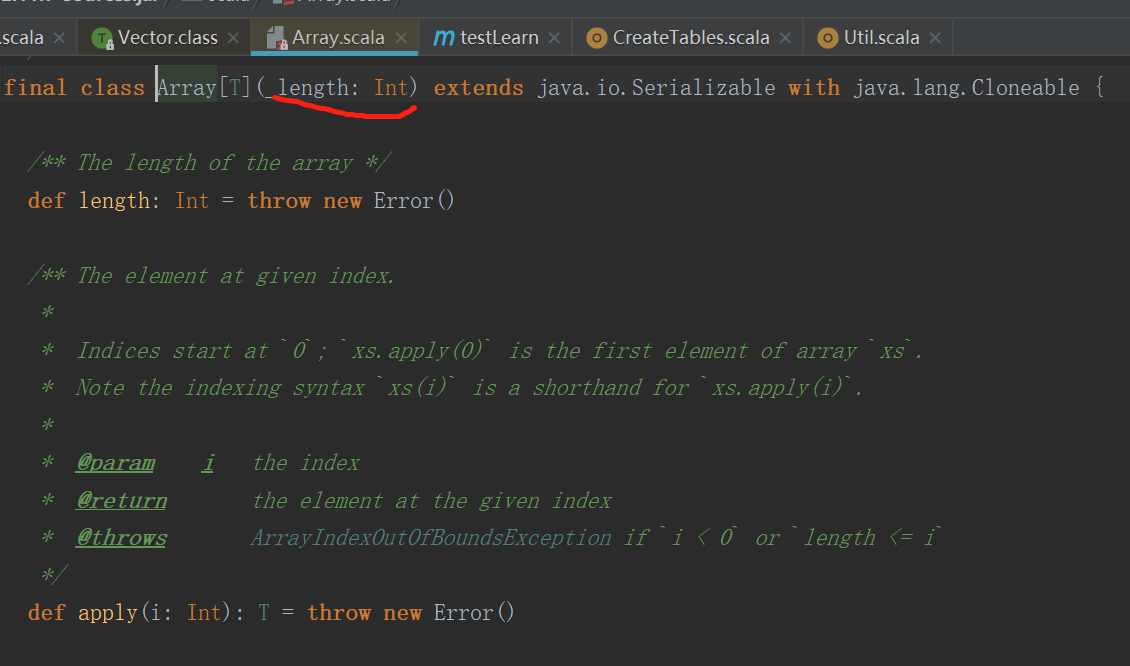
到这里，问题还没解决，就是实际spark的API中Vector.toArray是不带参数的，这里带参数为啥不报错？



原因如下：

在上图toArray处按快捷键Ctrl+鼠标点击可以看到，Spark中vector的toArray方法是继承scala.Array，而scala.Array是可以带参数的，如下：





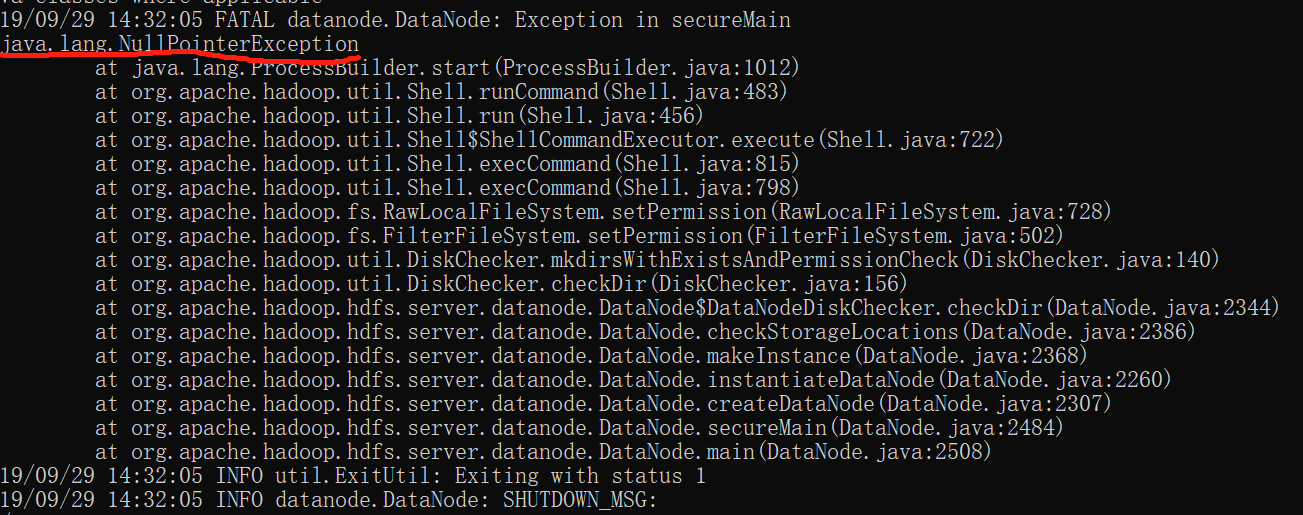
不过，说明一个问题，就是spark中vector的toArray方法是可以带参数的，参数表示元素的索引，scala中vector容器的toArray方法不能带参数。

14、windows中cmd里运行start-dfs.cmd启动hadoop，报错:

**找不到或无法加载主类 org.apache.hadoop.hdfs.server.namenode.NameNode**

解决方法：按照网上解决方法查了一下hadoop-env.cmd文件并修改，发现根本不是hadoop-env.cmd的问题。束手无策时，考虑到开始安装hadoop时对各种文件配置完全不懂，就是按照网上教程胡配一通的，所以就删除了现有hadoop文件夹，重新安装hadoop文件并进行配置(具体配置方法看厦门大数据实验室的配置教程即可)，然后就有了第15个问题。

15、（1）windows中cmd里运行start-dfs.cmd启动hadoop，报错:



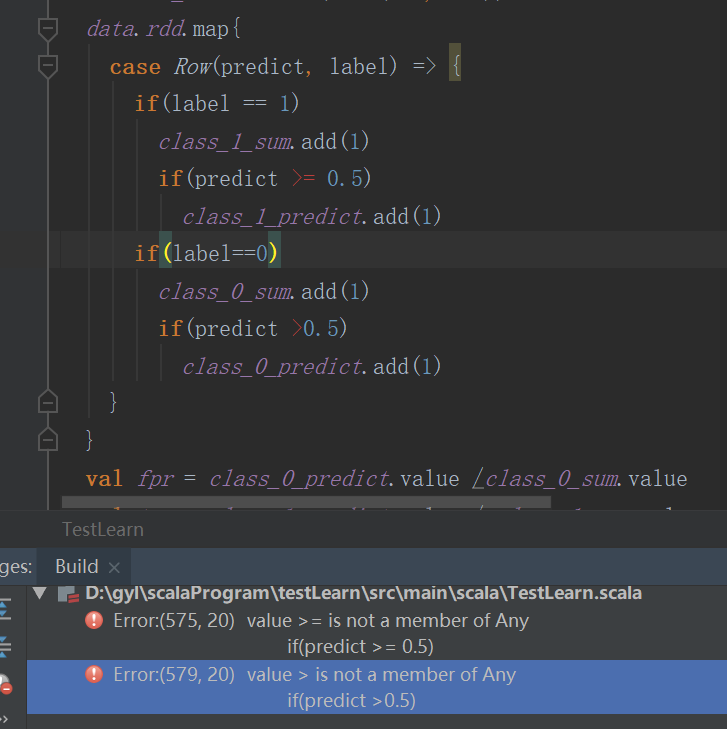
（2）idea中运行程序报错：



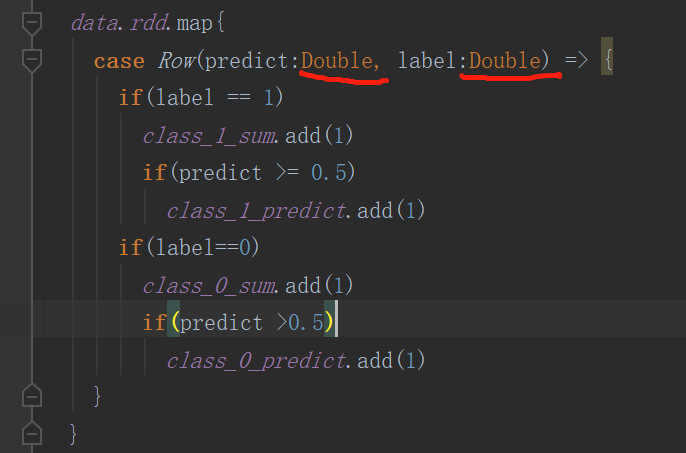


出现以上问题的原因是hadoop缺乏在windows上运行的必要组件，像winutils.exe等，这些组件的下载地址：<https://github.com/steveloughran/winutils>，从winutils-master\hadoop-2.7.1目录中直接拷贝bin文件夹，覆盖掉hadoop-2.7.1/文件夹中的bin目录即可。

16、出现如下报错：比较符号无法解析



解决方法：根据提示，比较符两边的值是Any类型**(对于一个未定义类型的变量，scala默认为Any型)**，解决办法如下：

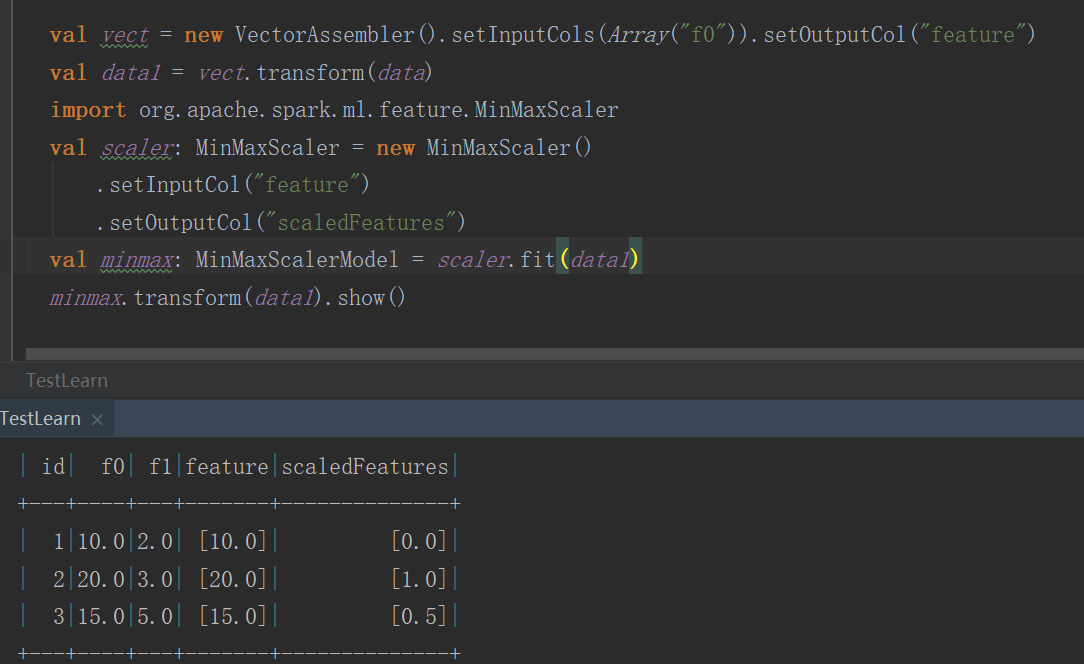


17、对数据进行minMax放缩的时候报错，意思是dataframe中列的格式不对。

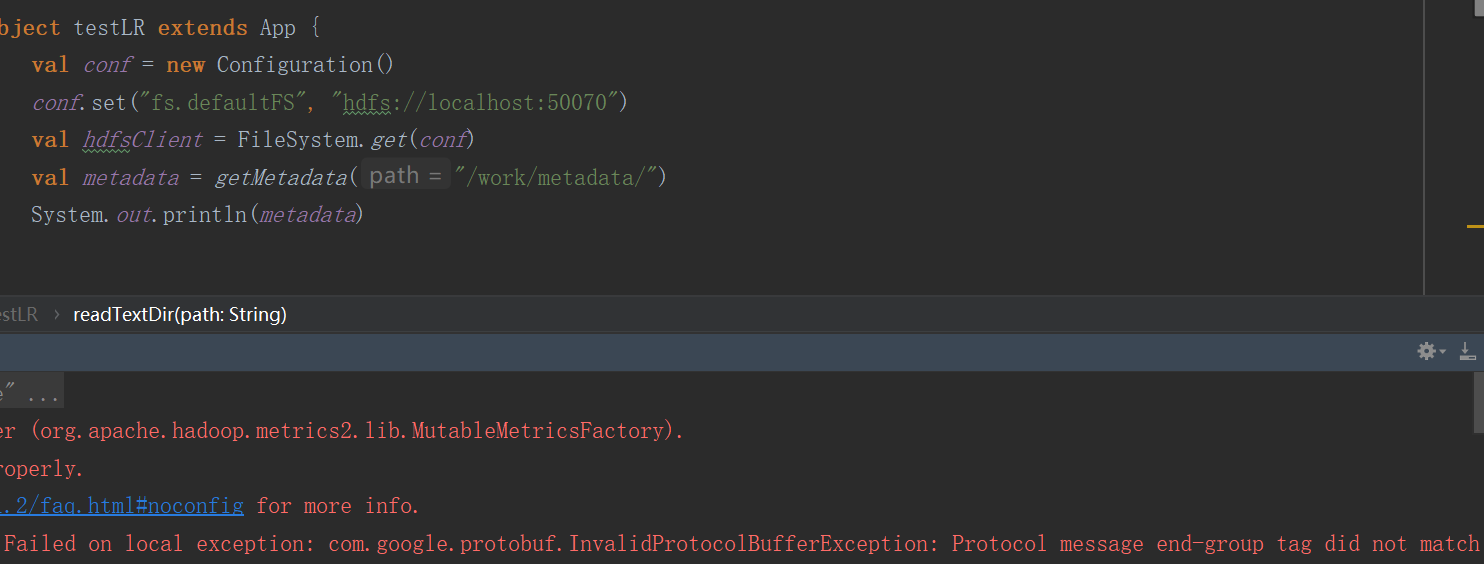
Exception in thread "main" java.lang.IllegalArgumentException: requirement failed: Column f0 must be of type struct<type:tinyint,size:int,indices:array<int>,values:array<double>> but was actually array<double>.

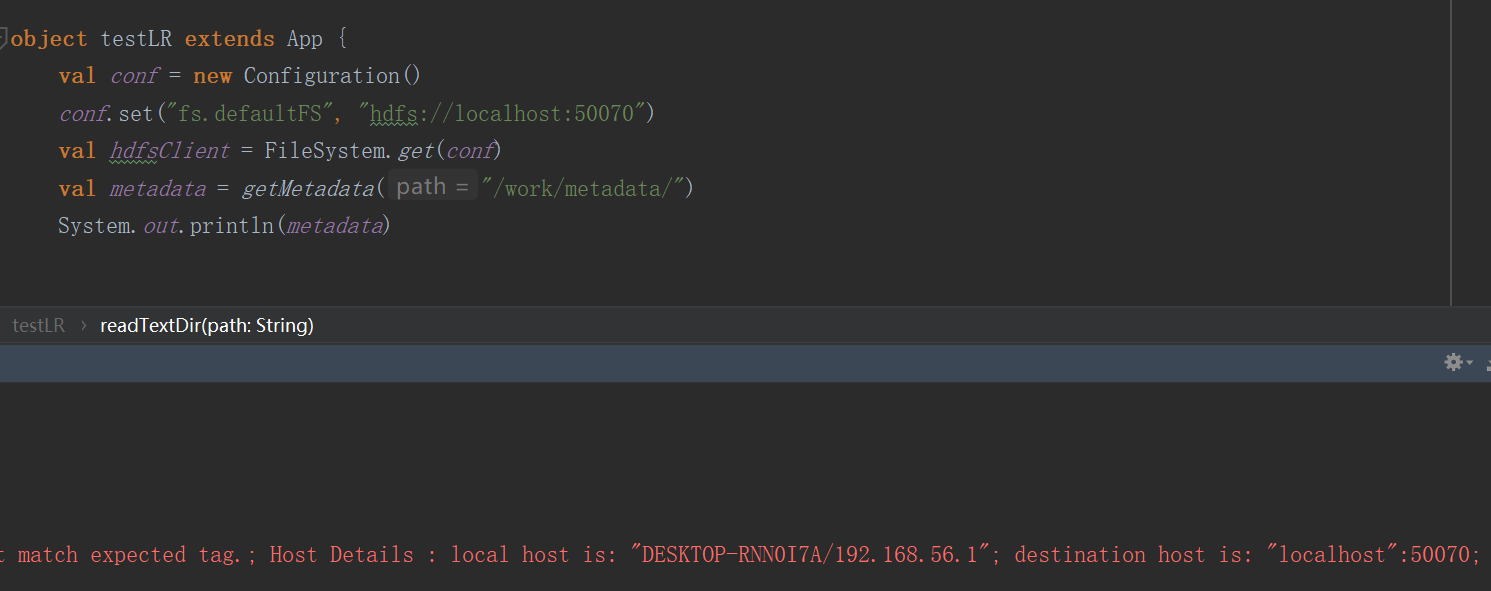


解决方法：利用VectorAssembler做个转换操作。

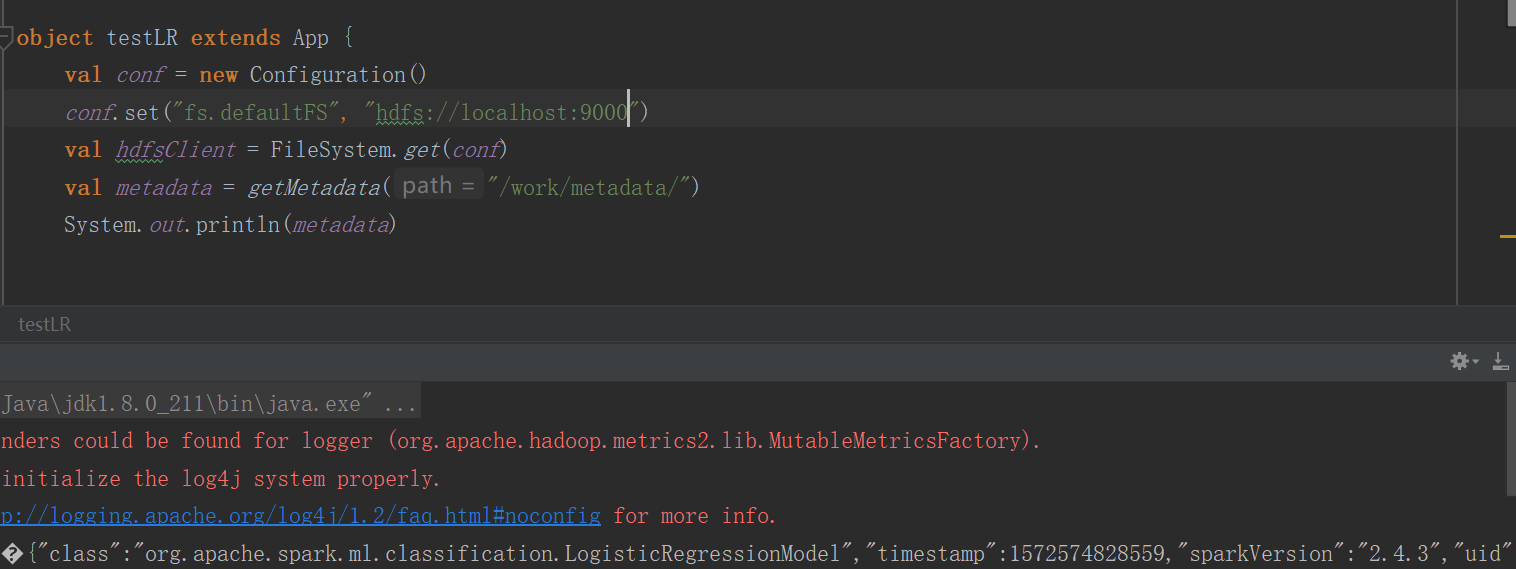


18、读取hdfs上的文件报错，

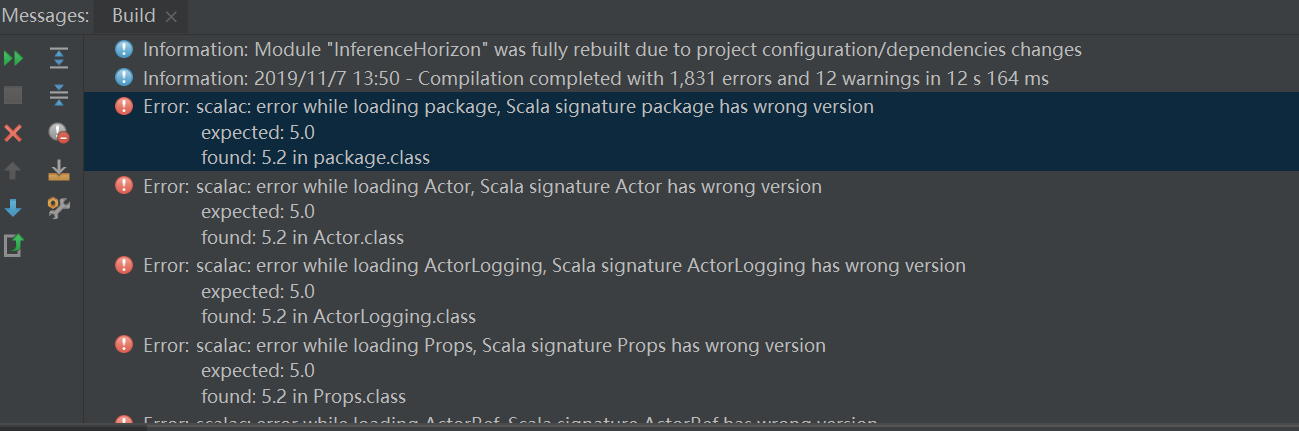




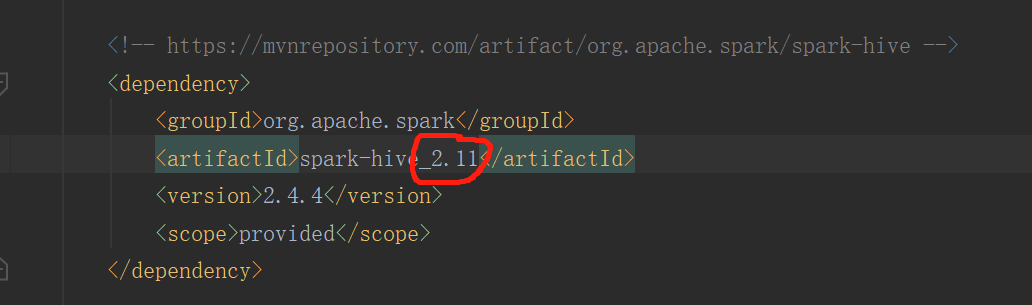
报错原因：hdfs的port配置出错，这是因为宿主机和hdfs有个port映射，宿主机的port是9000，映射到hdfs的port是50070，所以在配置的时候，应该设置port是宿主机的port。改完运行后



19、



出现下面问题主要是scala版本对应有问题，比如：scala版本是2.11的，但是在pom.xml文件中写spark-hive依赖时是这样写的：spark-hive\_2.12，所以运行时报上述错误，重新改成spark-hive\_2.11即可。



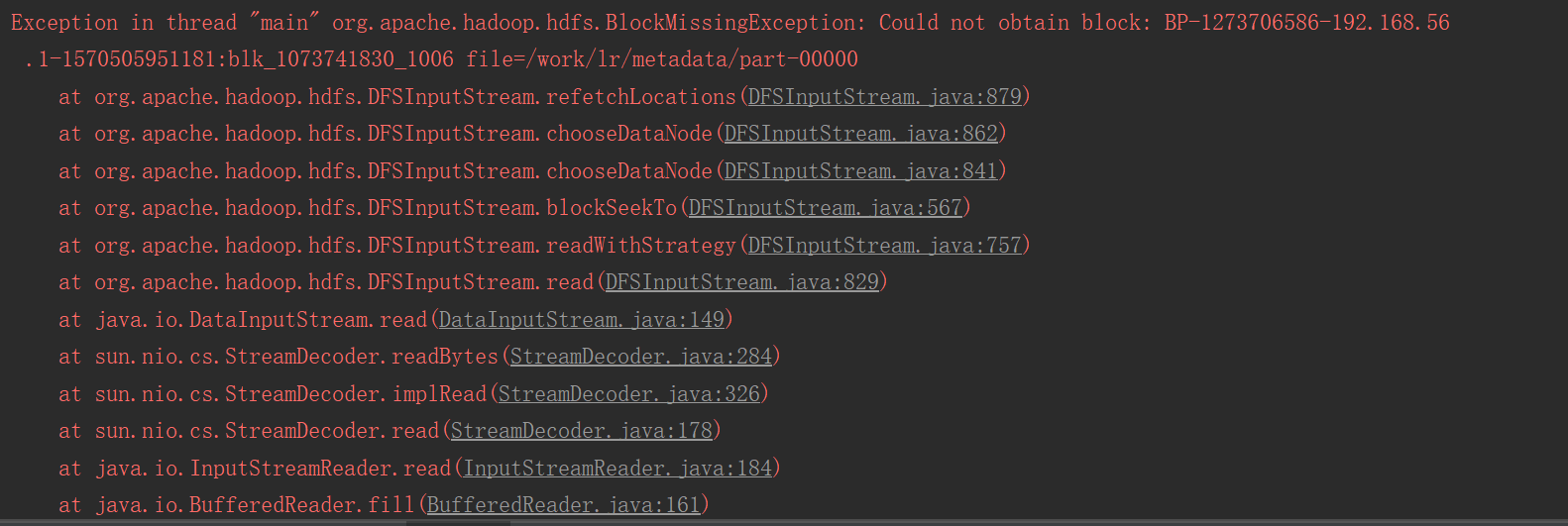
20、Error: value create is not a member of object org.apache.spark.ml.util.Instrumentation，实际中create的确是org.apache.spark.ml.util.Instrumentation中的方法。

出现这类问题有两种可能原因及解决方法：

1. 没有相应的依赖，添加依赖即可；
2. 依赖的版本问题，我这里是因为依赖里的版本(2.4.4)太高了，改成2.3.2就好了。（参考<https://www.oipapio.com/question-3924425>）

21、Error：java: returnResource(redis.clients.jedis.Jedis) 在 redis.clients.jedis.JedisPool 中是 protected 访问控制。

出现这个问题的原因：pom.xml中redis.clients.jedis的版本问题，版本太新了，将其从3.0.0降成2.9.0后就全好了。

22、

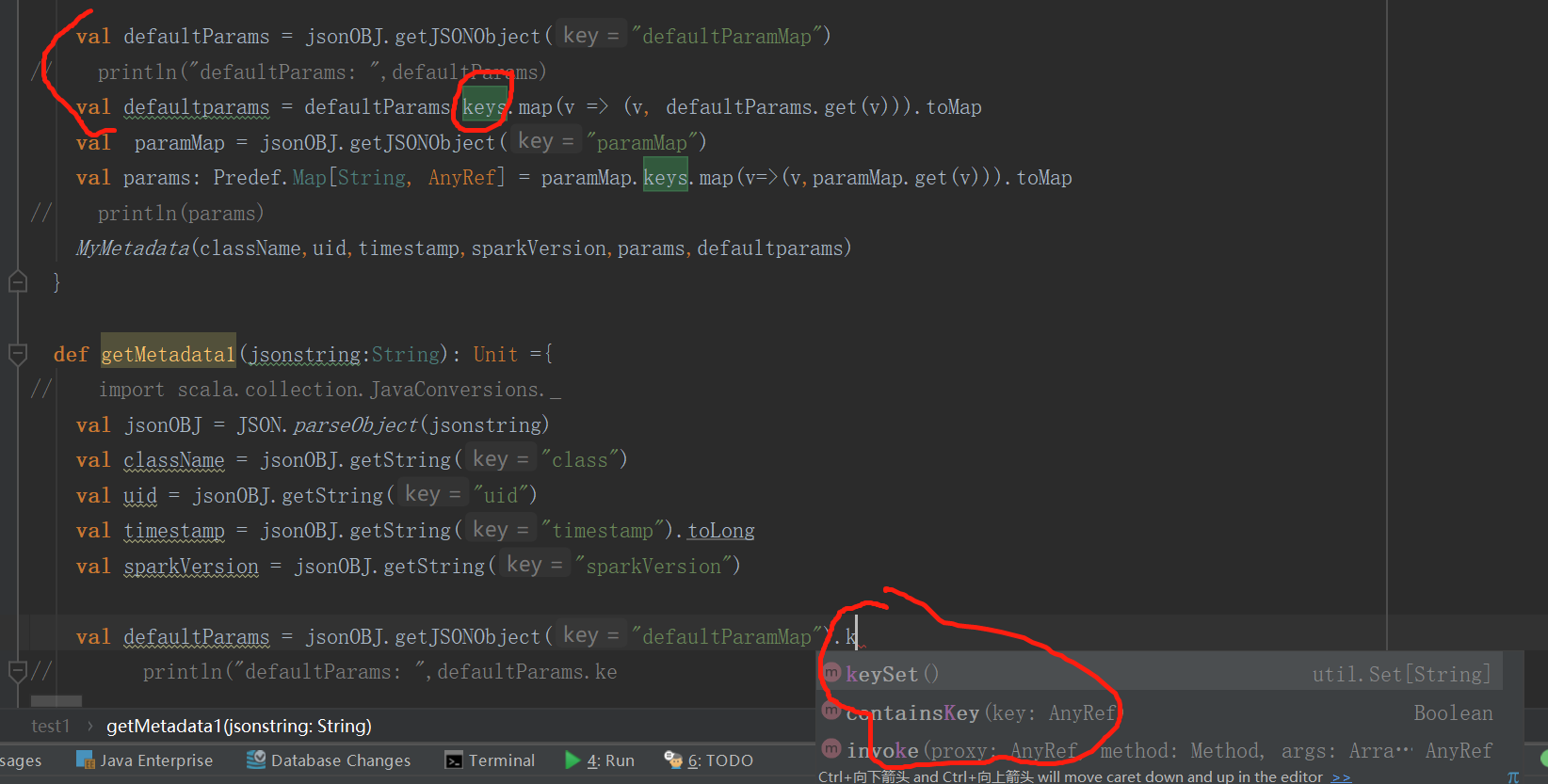
错误原因：在浏览器上看到datanode节点挂，可能是电脑睡眠导致的。

23、 [Error:(345, 17) constructor LinearSVC Model in class LinearSVC Model cannot be accessed in object test1 val model = new LinearSVCModel(metadata.uid, coefficientsVector, intercept)](https://www.google.com/search?q=Error:(345,+17)+constructor+LinearSVC+Model+in+class+LinearSVC+Model+cannot+be+accessed+in+object+test1+val+model+%3D+new+LinearSVCModel(metadata.uid,+coefficientsVector,+intercept)&spell=1&sa=X&ved=0ahUKEwiX7L6PkeHlAhWiHqYKHV_zDW4QBQgtKAA)

出现这个问题的原因：class LinearSVCModel的使用是有作用域的，其作用域限定为org.apache.spark。

解决方法：必须将test1这个对象添加作用域org.apache.spark，添加方法为在test1.scala首行添加package org.apache.spark。

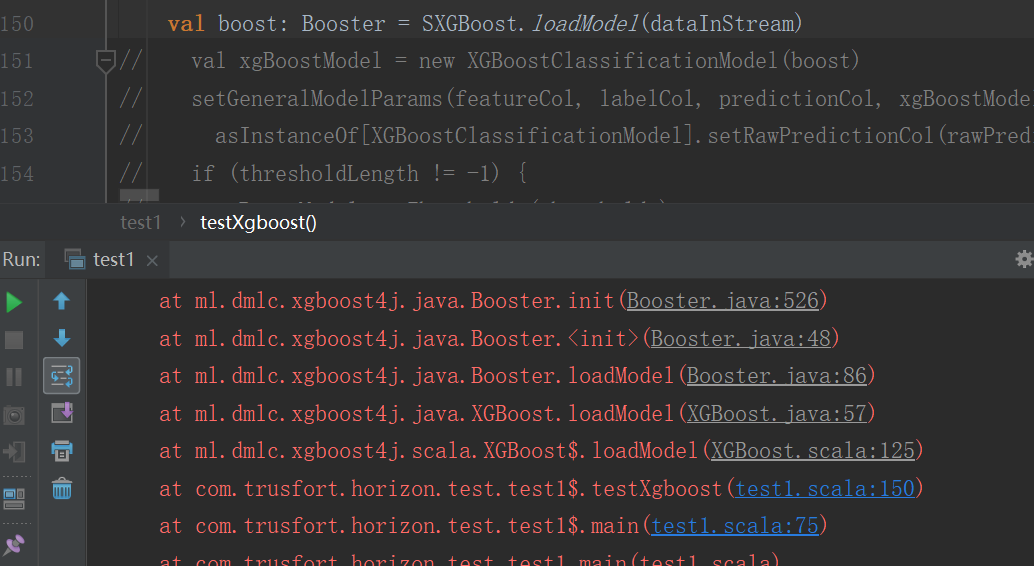
24、

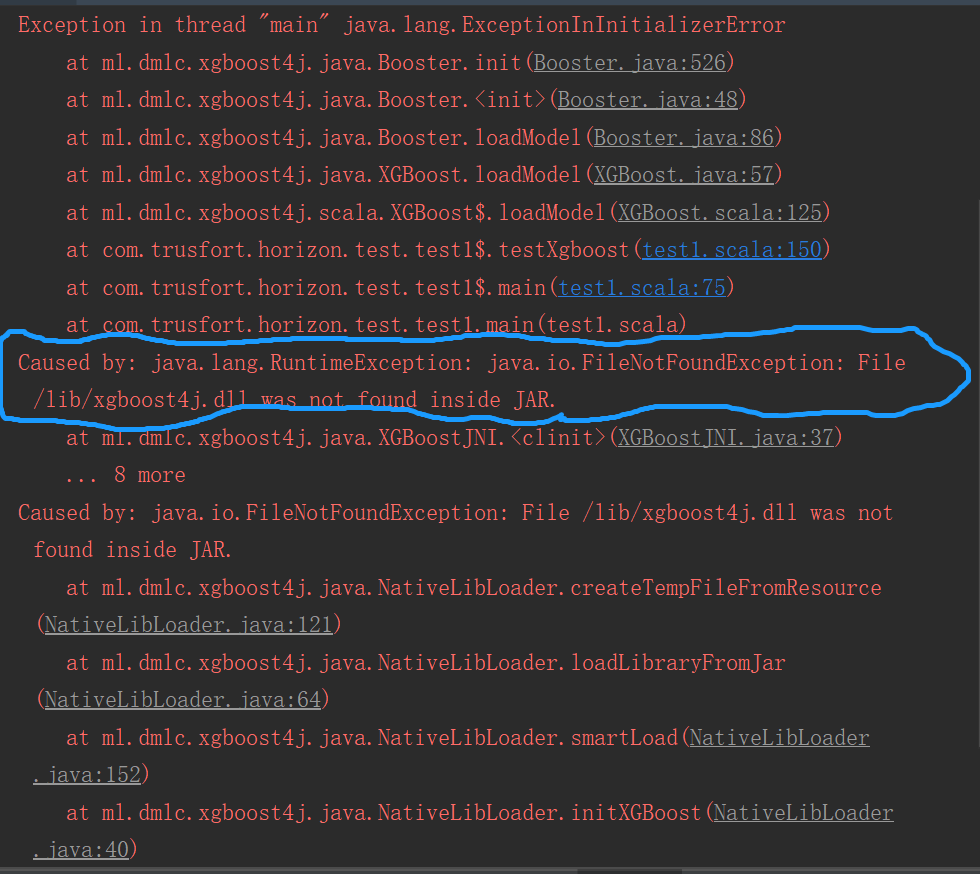


错误描述：解析JSON字符串的时候，两段代码是一样的，但是发现上面的defaultparams存在方法keys，但是下面的defaultparams没有方法keys。

解决方法：出现上述错误的原因是这里存在java和scala的转换，下面的代码里没有引入scala.collection.JavaConversions.\_导致的，引入之后，问题解决了。

25、代码运行到加载XGBoost模型时报如下Jar包错误：java.lang.RuntimeException: java.io.FileNotFoundException: File /lib/xgboost4j.dll was not found inside JAR.





检查下pom中XGBoost依赖，

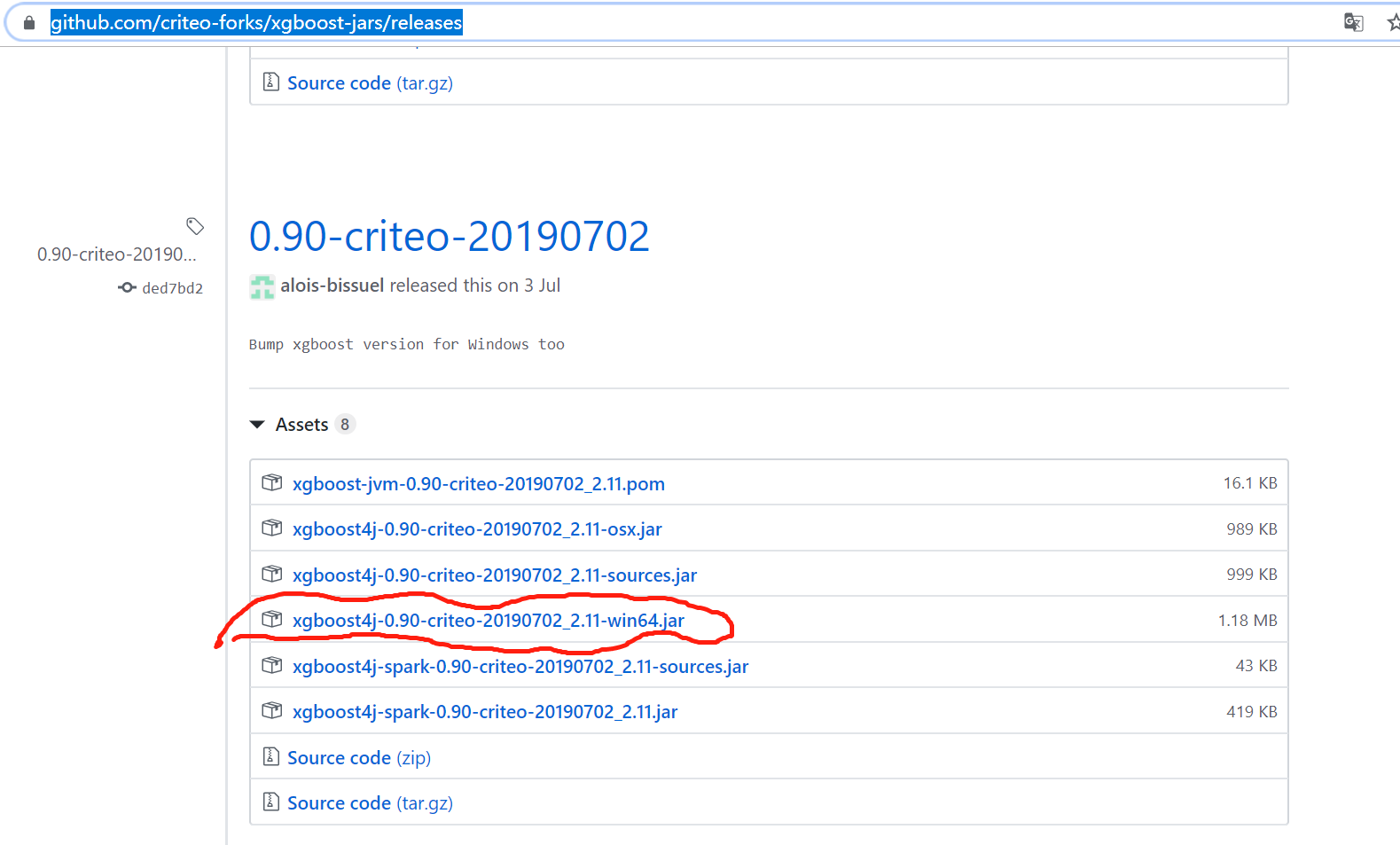


Google一下这个问题：

（1）这个链接(<https://github.com/dmlc/xgboost/issues/3683>)的解释：jar包问题，如下



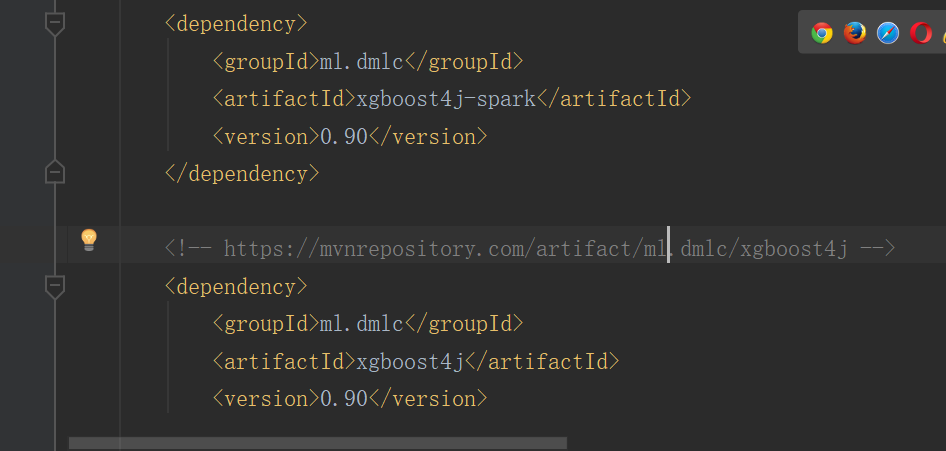
因为报错也是说缺少xgboost4j.dll文件，所以比较符合这个解释，于是根据提示的链接<https://github.com/criteo-forks/xgboost-jars/releases>下载了下图红框里的jar包。



（2）可以看看其它的解释(没试过里面的方法)，比如知乎里有人说XGBoost支持spark版本2.3.0以上，但是我的spark版本是2.4.3的，所以应该不是这个问题。

解决方法：

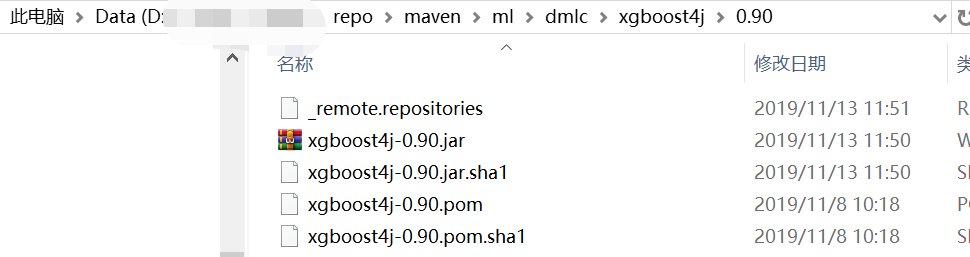
（1）首先，因为上面下载了xgboost4j-0.90-…-win64.jar包，所以在pom中将版本更新到0.90。

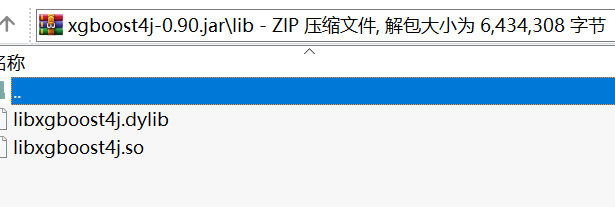


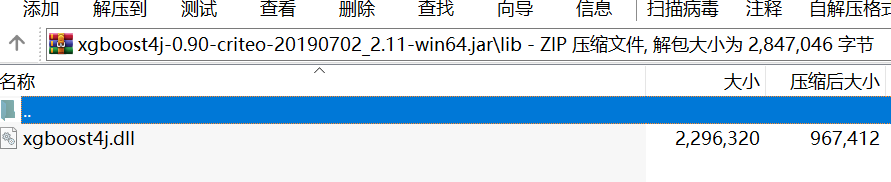
（2）找到maven中jar包的存放位置，如图，下面0.90文件夹就是更新后重新下载的，



进入0.90的文件夹，有下面几个文件，用winrar打开xgboost4j-0.90.jar进入lib文件夹会发现的确缺少xgboost4j.dll，而打开xgboost4j-0.90-…-win64.jar包会发现有该文件。

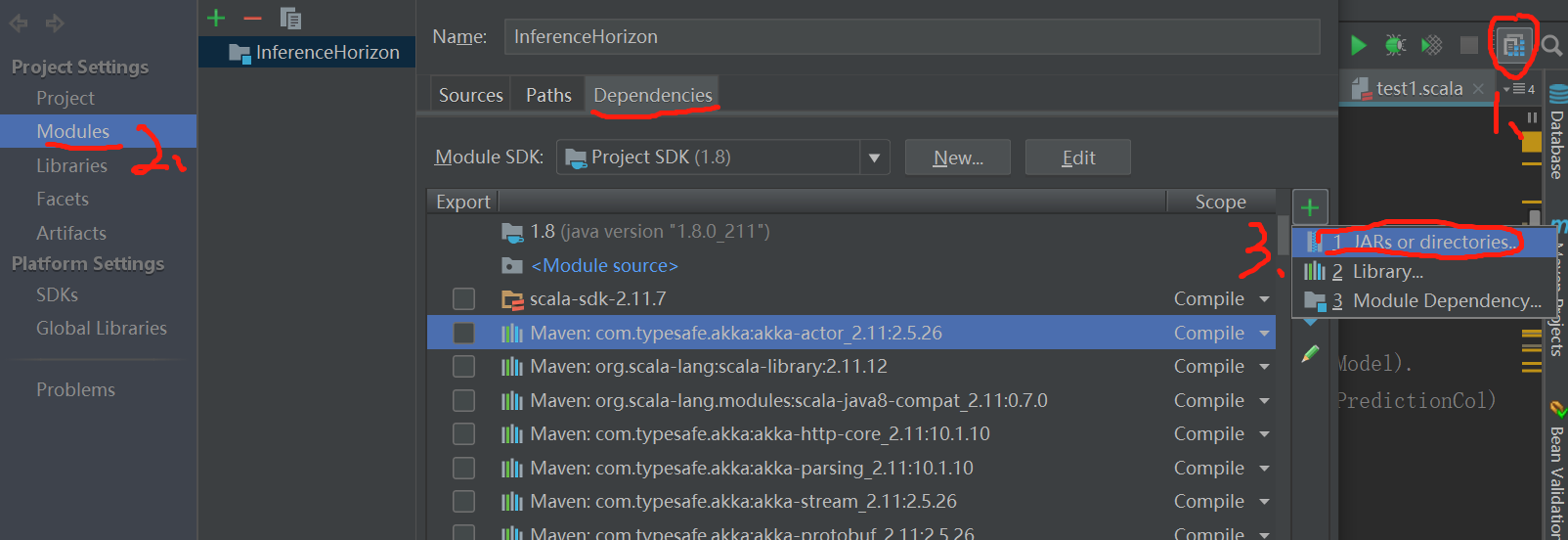


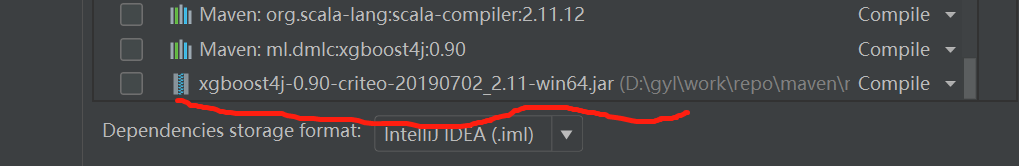




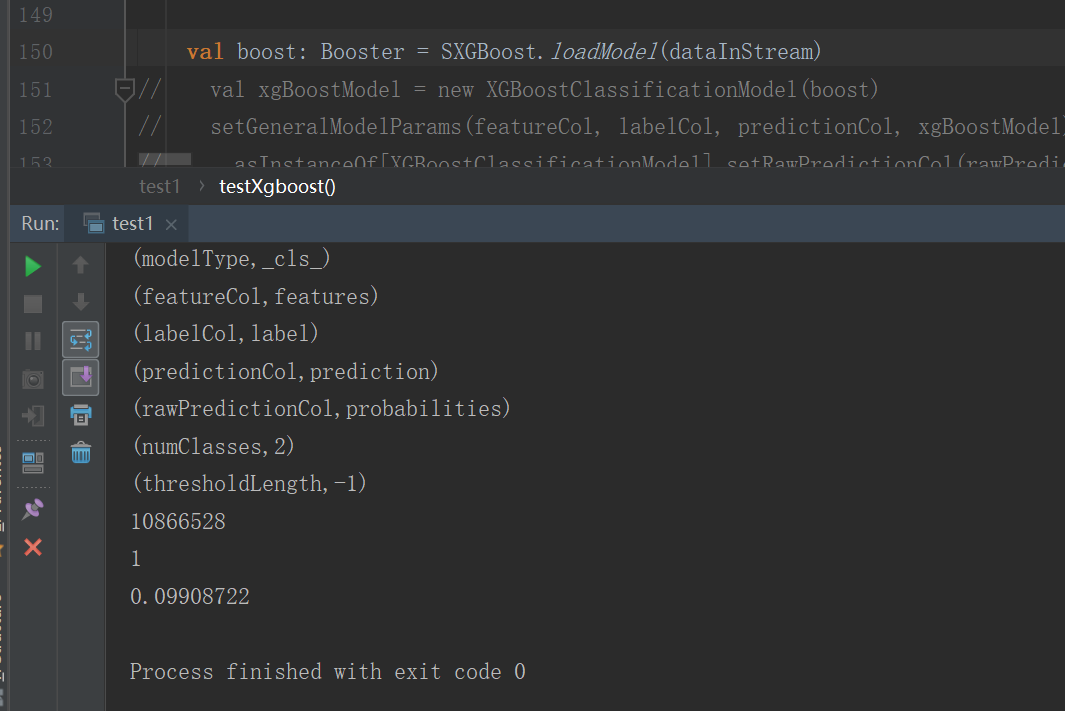
（3）将下载的xgboost4j-0.90-…-win64.jar包放到该0.90文件夹下。（其实放不放都无所谓）

（4）接下来，就是在idea中导入该jar包，按以下步骤操作，然后能看到被导入了。

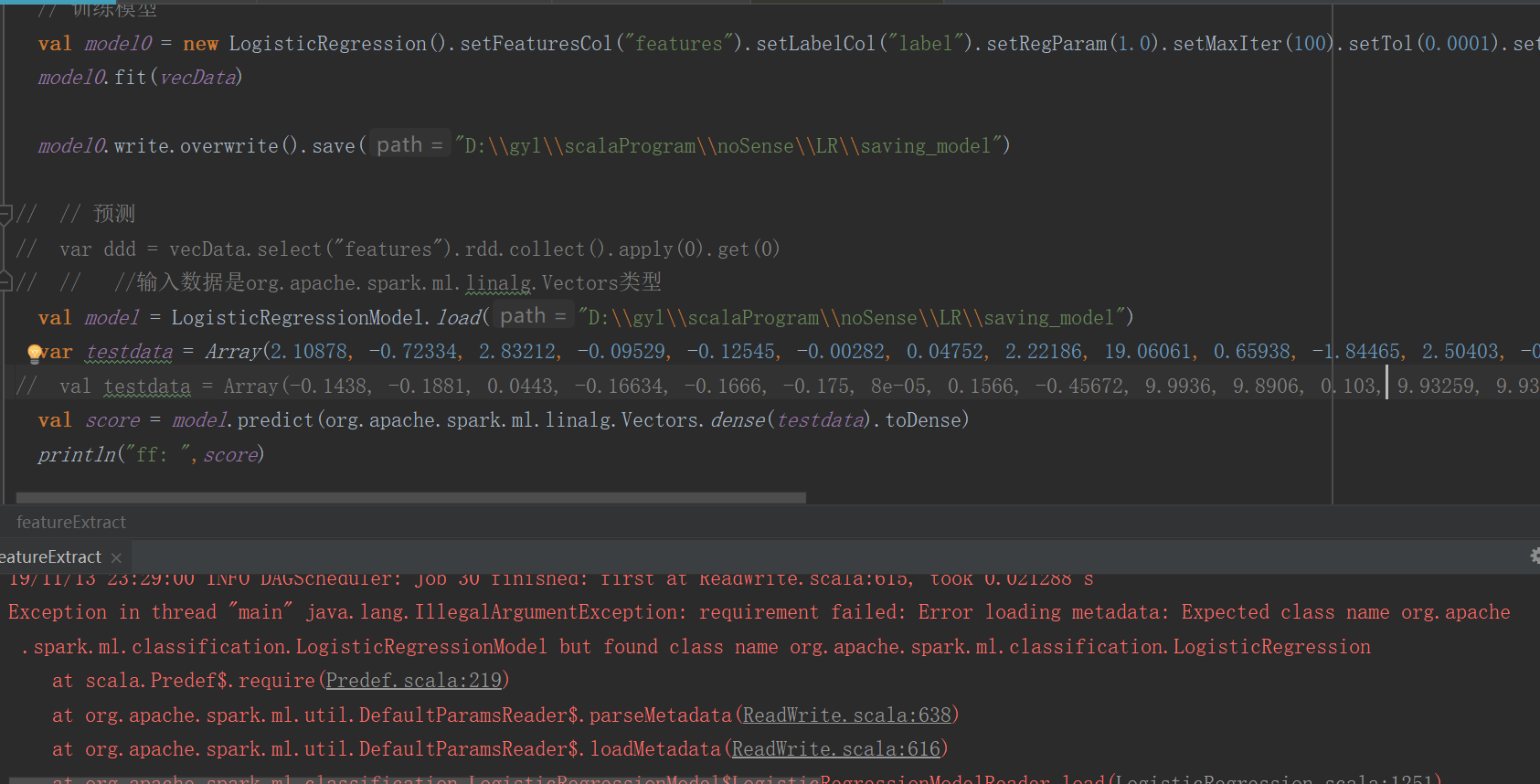




（5）最后，问题被解决了。

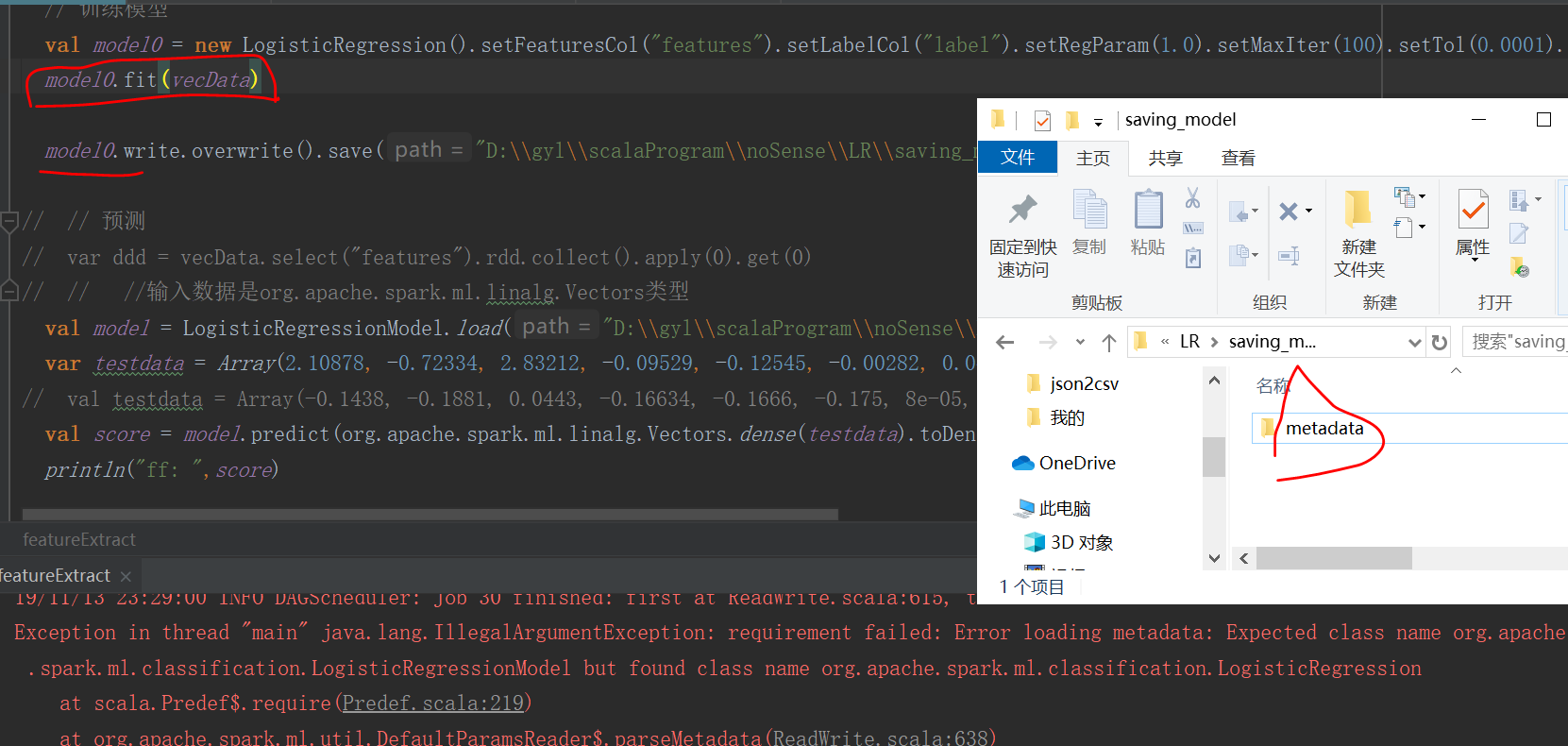


26、训练模型后进行预测时报下图中的错误：Exception in thread "main" java.lang.IllegalArgumentException: requirement failed: Error loading metadata: Expected class name org.apache.spark.ml.classification.LogisticRegressionModel but found class name org.apache.spark.ml.classification.LogisticRegression。

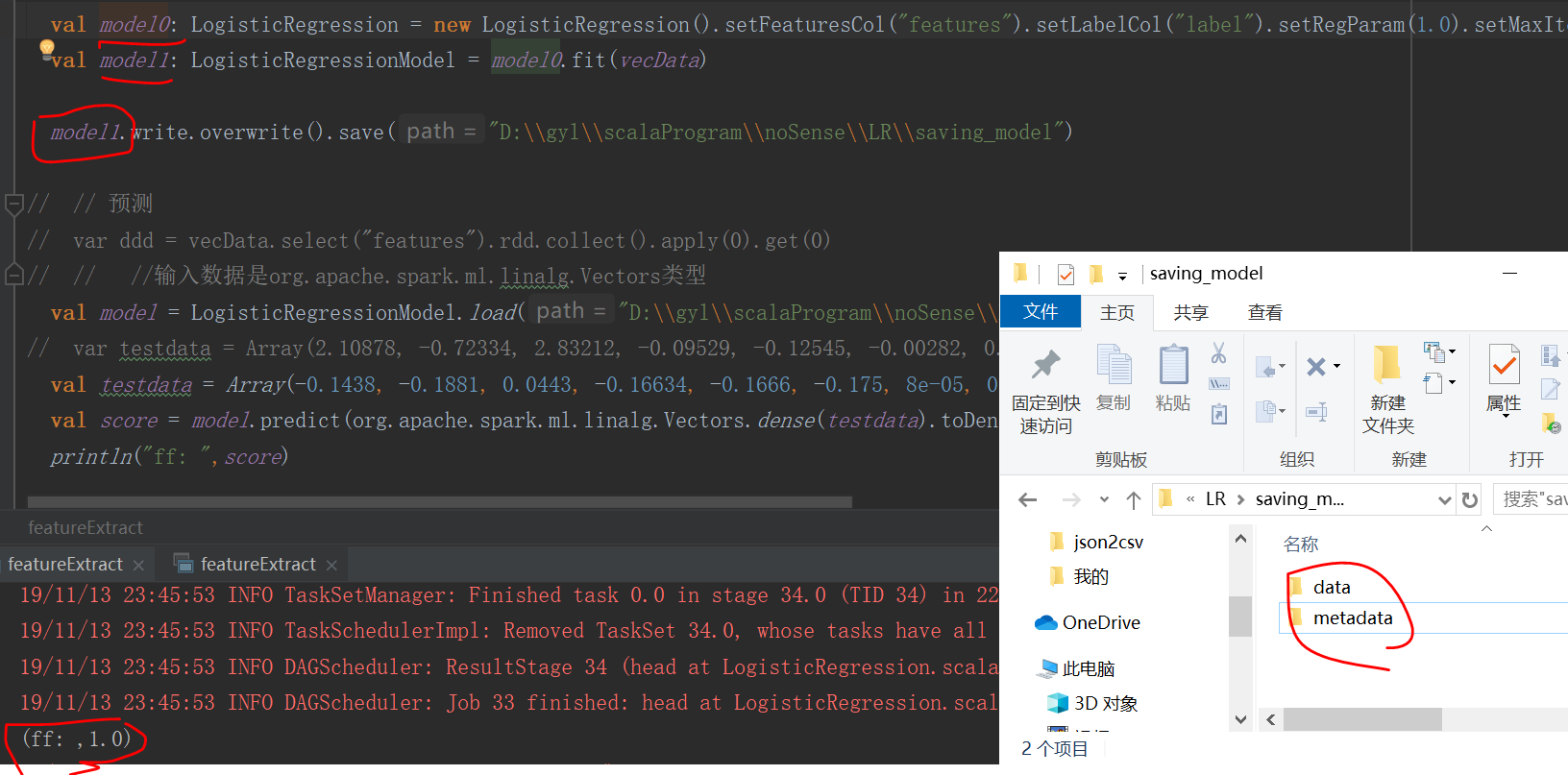


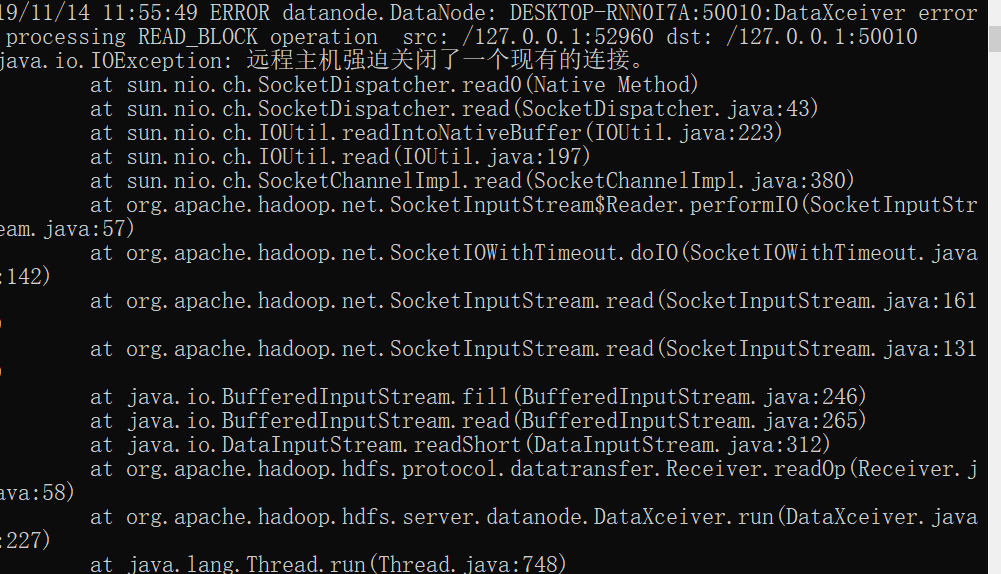
错误描述：保存模型的时候，其实是保存的LogisticRegression类型，并没有保存LogisticRegressionModel类型，查看保存路径的文件夹也只有metadata文件，没有data信息。

错误原因：定义的model0变量是LogisticRegression类型的，虽然使用了fit拟合数据，但此时并没有重新定义变量来接受其fit之后的返回值(返回值类型是LogisticRegressionModel类型)，所以write保存的时候就只是保存了模型的元数据信息model0变量。

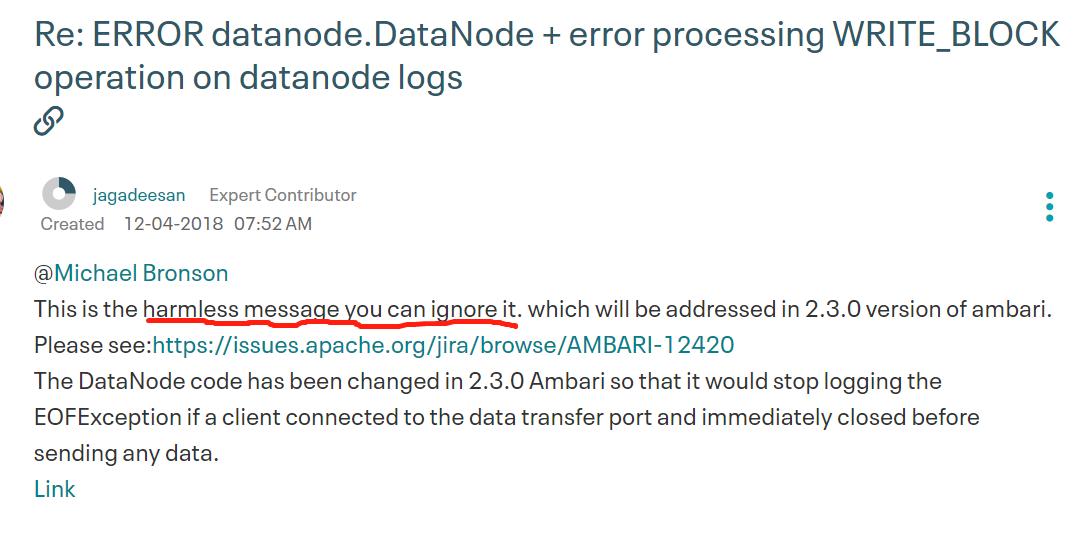


解决方法：定义一个新的变量model1来接收变量model0拟合数据后的返回值，并将保存该返回值，如下图。



27、

解决方法：<https://community.cloudera.com/t5/Support-Questions/ERROR-datanode-DataNode-error-processing-WRITE-BLOCK/m-p/238072>



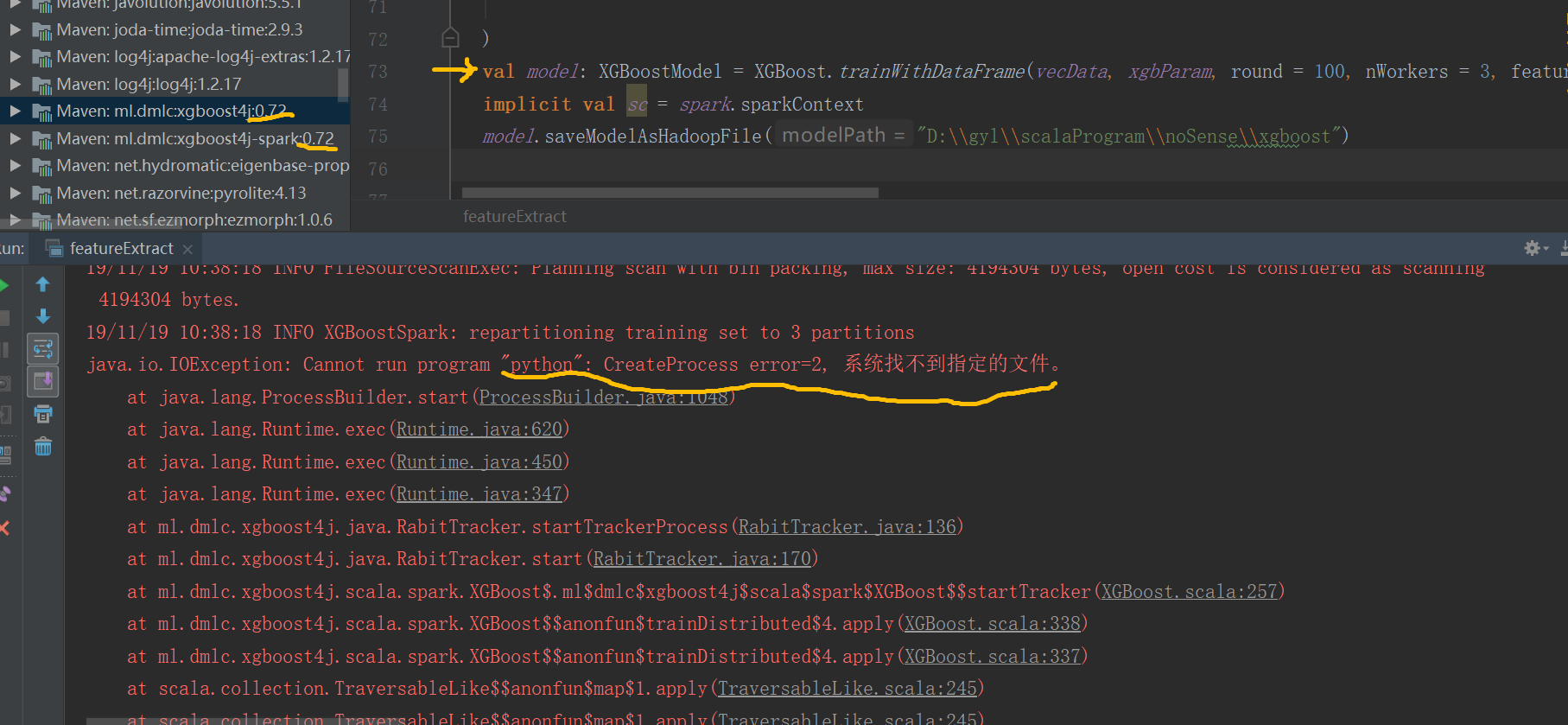
问题：

1、因为训练模型的数据是没有考虑精度的，而测试数据考虑了精度，只选取小数点后五位，发现模型预测结果出错，将格式统一后，预测正确。（python）

2、根据1中格式统一后的数据放到spark上进行训练时，预测同样的数据出错，也就是同样的训练数据和预测数据，在python上训练得到的模型的预测结果是正确的，而在spark上训练得到的模型的预测结果是错误的。

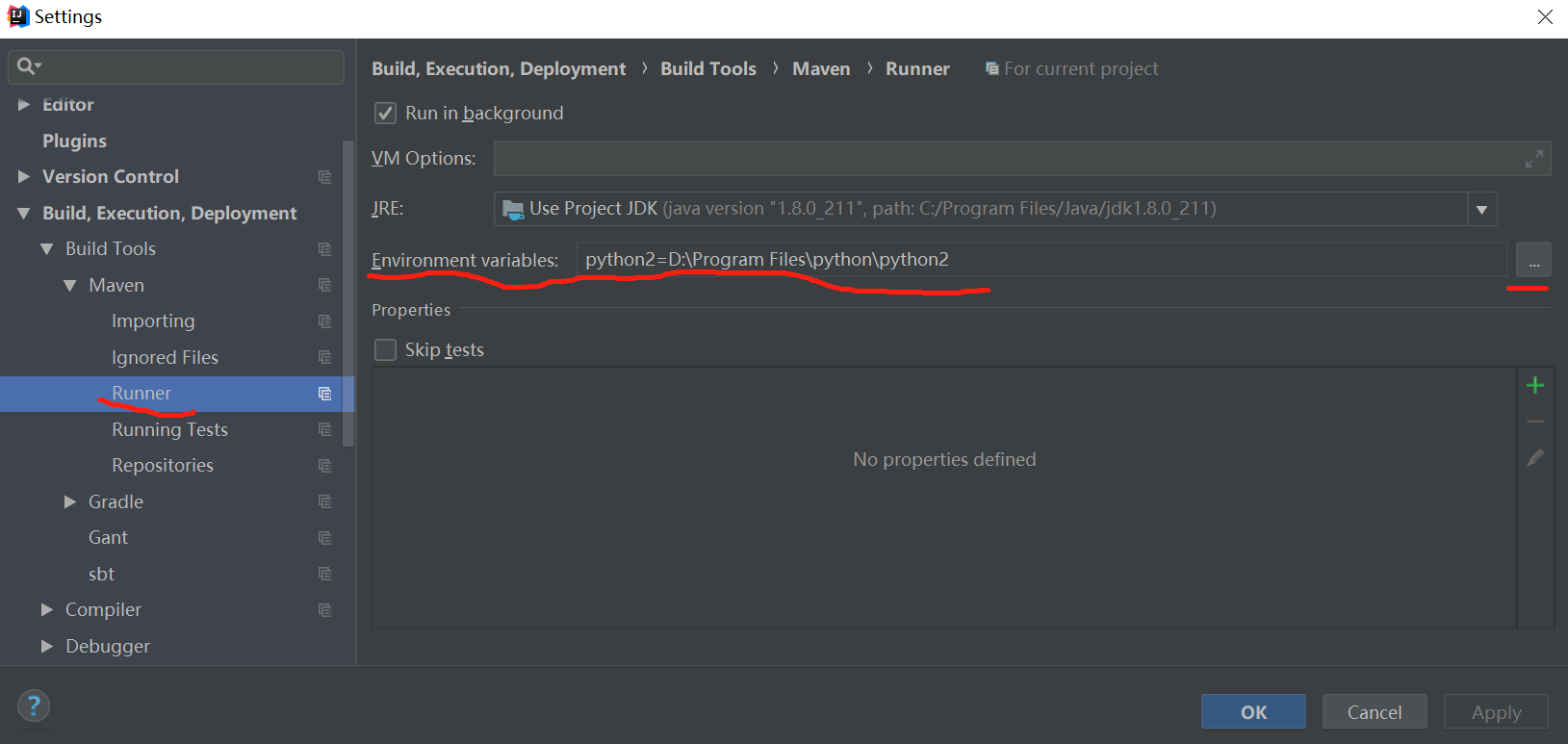
问题解决：spark的模型未设置参数，而python里的参数有默认值；通过将spark里的参数按照python里默认的参数设置好就ok了。

28、（1）spark上训练XGBoost模型报错：java.io.IOException: Cannot run program “python”: CreateProcess error=2，系统找不到指定的文件。XGBoost包版本为0.72.

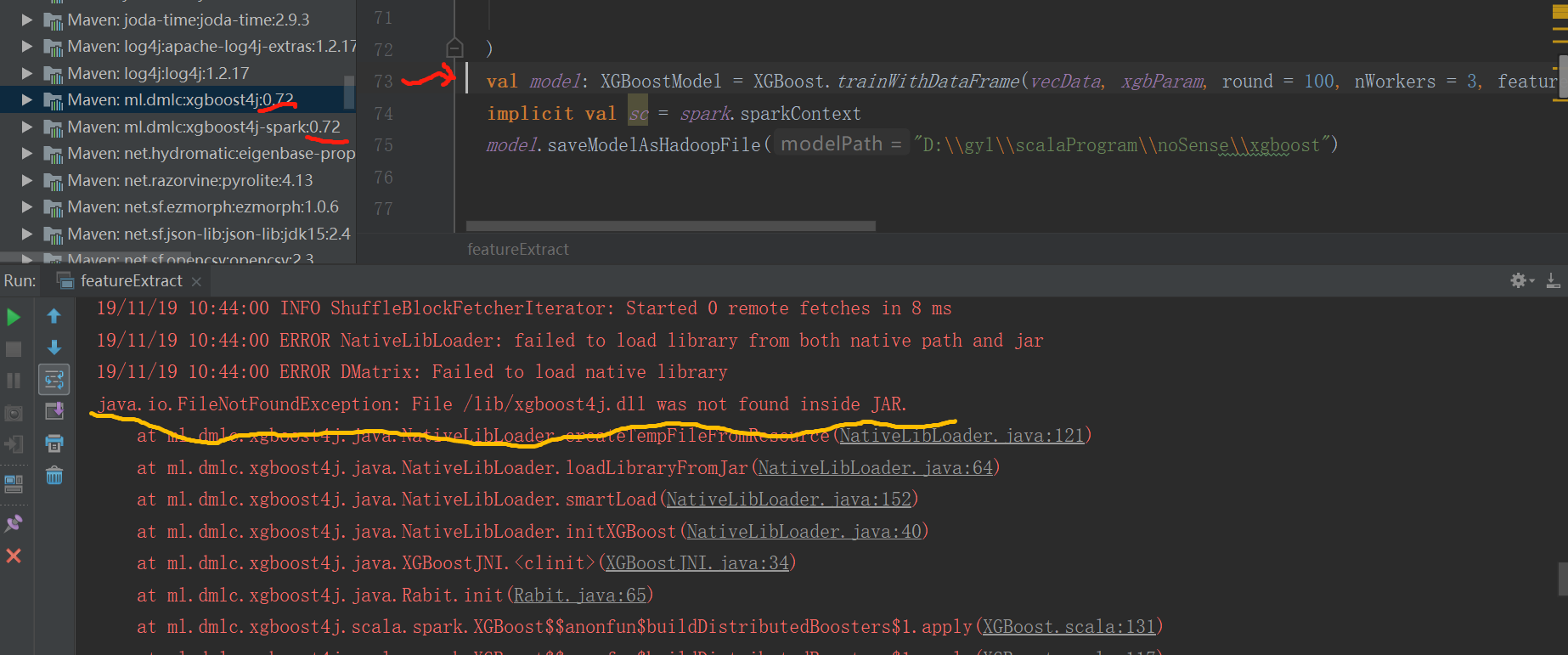


出现上述错误的原因：idea中没有配置python环境。

解决方法：在idea中按如下添加python路径，注意python路径中必须是python.exe，如果是python2.exe或python3.exe，都会报错。



（2）按上述方法添加python环境变量后，再次运行又报如下错误，这个错误和第25个错误一样： java.io.FileNotFoundException: File /lib/xgboost4j.dll was not found inside JAR.。



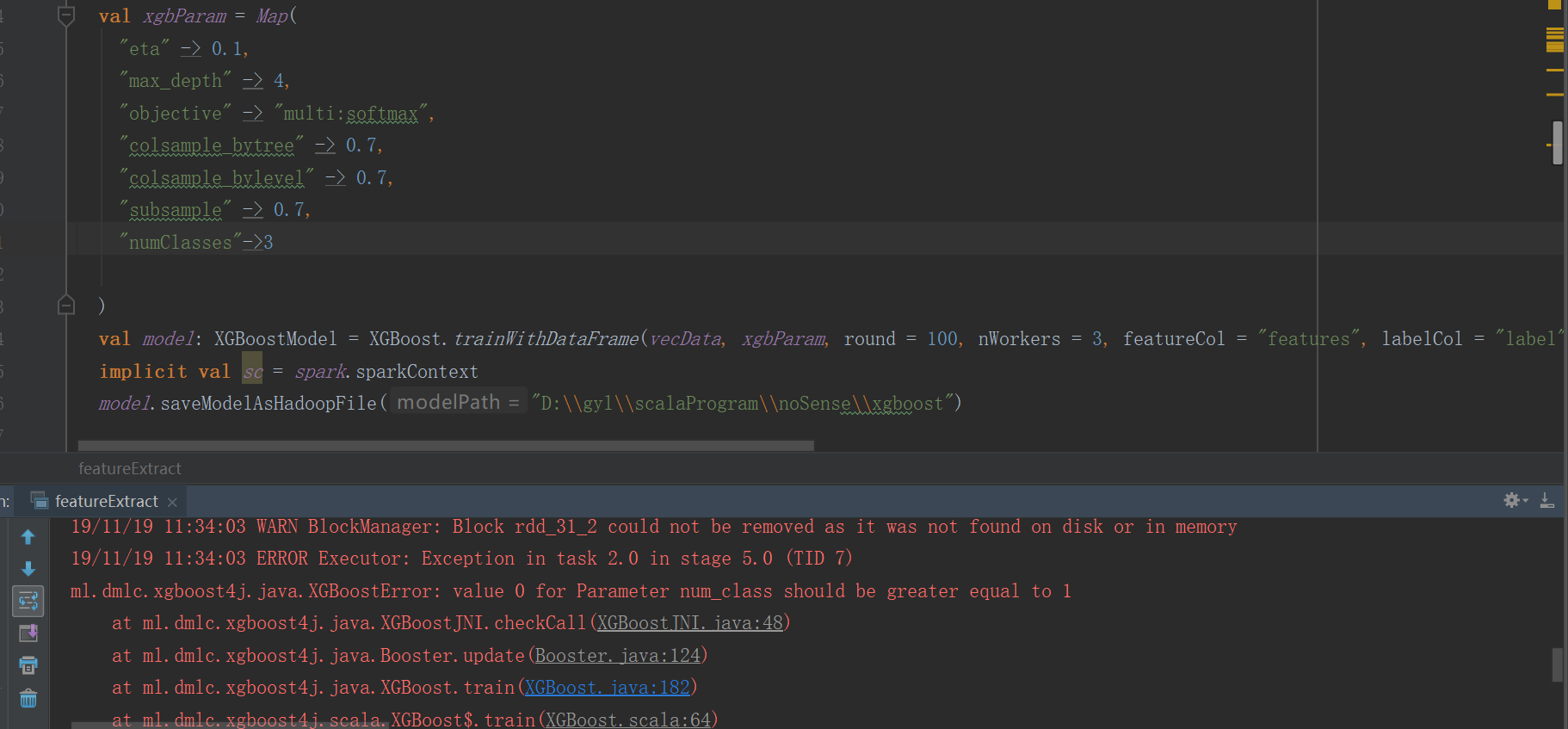
29、（1）ml.dmlc.xgboost4j.java.XGBoostError:C:\projects\xgboost-jars\xgboost\src\objective\regression-obj.cc:103:Check failed:Loss::CheckLabel(y) label must be in [0, 1] for logistic regression。



错误原因：xgboost参数设置时，objective参数设置为了binary:logistic，而实际数据是多标签的。

解决方法：将objective参数设置为multi:softmax。

（2）ml.dmlc.xgboost4j.java.XGBoostError: value 0 for Parameter num\_class should be greater equal to 1.

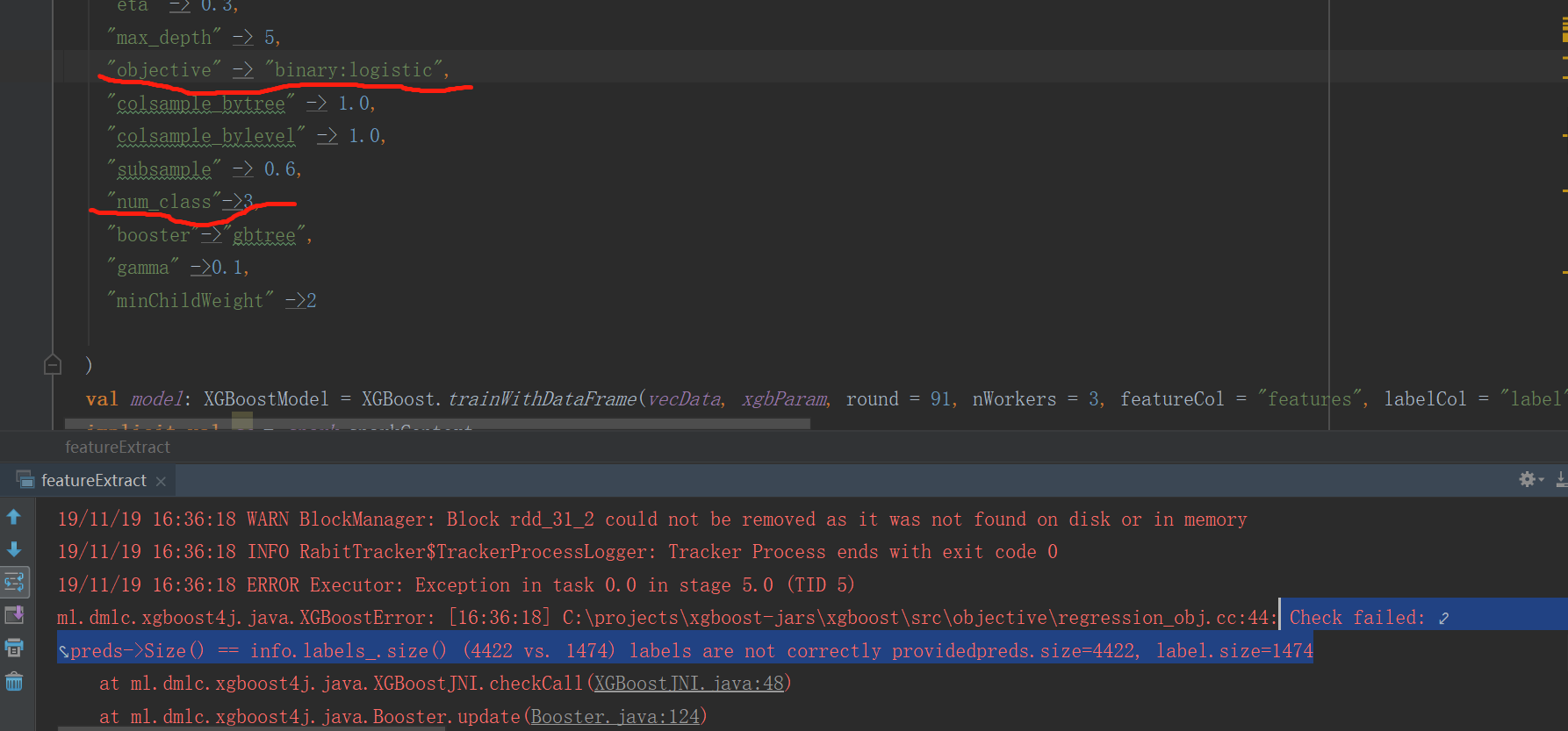


错误原因：参数设置错误，应该设置”num\_class”，而不是”numClasses”。

解决方法：将”numClasses”->3改成”num\_class”->3，参考：<https://github.com/dmlc/xgboost/issues/3186>。其它参数设置见下图：



30、ml.dmlc.xgboost4j.java.XGBoostError: [16:36:18] C:\projects\xgboost-jars\xgboost\src\objective\regression\_obj.cc:44: Check failed: preds->Size() == info.labels\_.size() (4422 vs. 1474) labels are not correctly providedpreds.size=4422, label.size=1474。



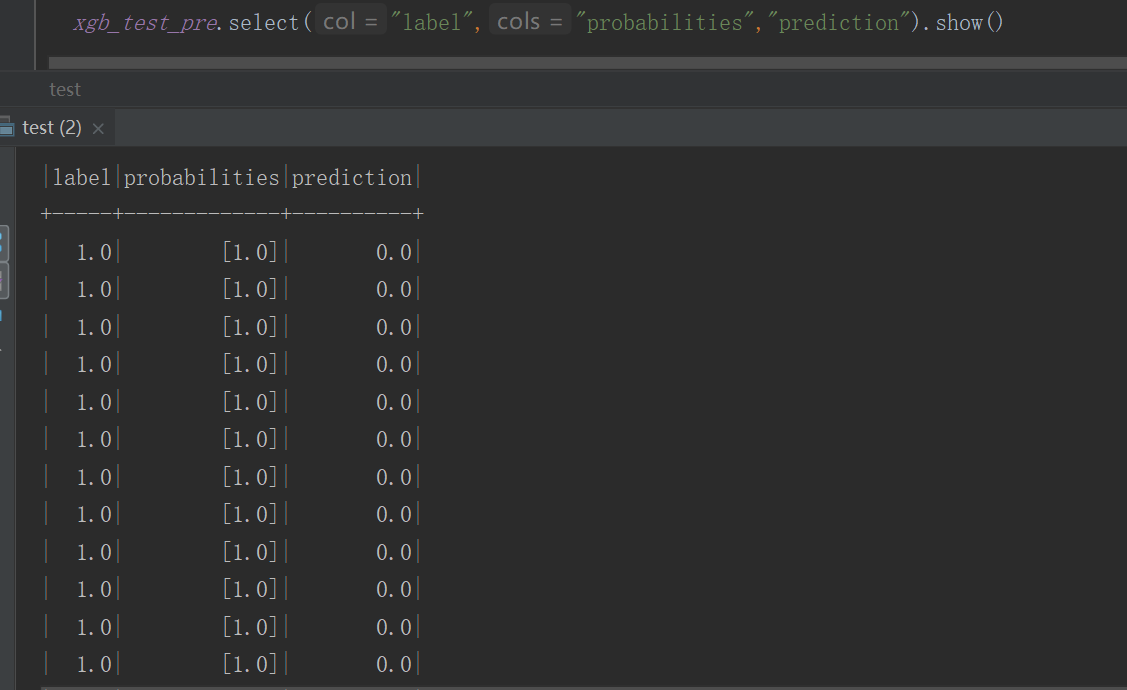
错误原因：参数设置错误，目标是二分类问题（"objective" -> "binary:logistic"），但是类别数设置为多分类（"num\_class"->3）。

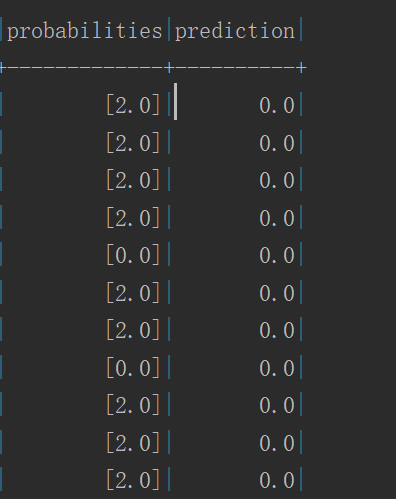
解决方法：将类别参数设置正确（"num\_class"->2），参考<https://github.com/dmlc/xgboost/issues/2563>.

31、待解决问题：spark中objective里softprob和softmax的区别？

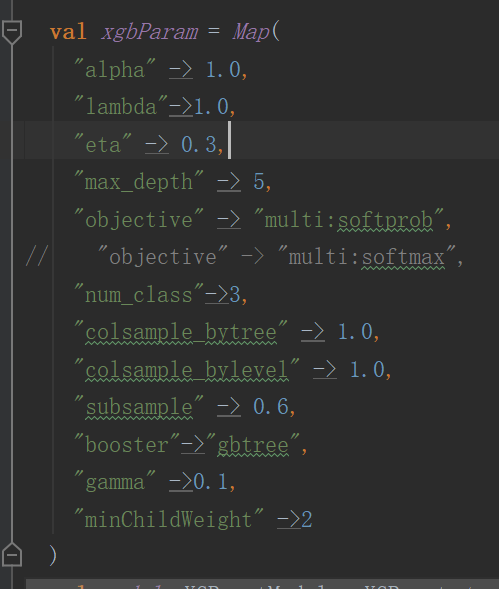
（1）当参数objective设置为softmax时，transform中probabilities和prediction的输出结果为

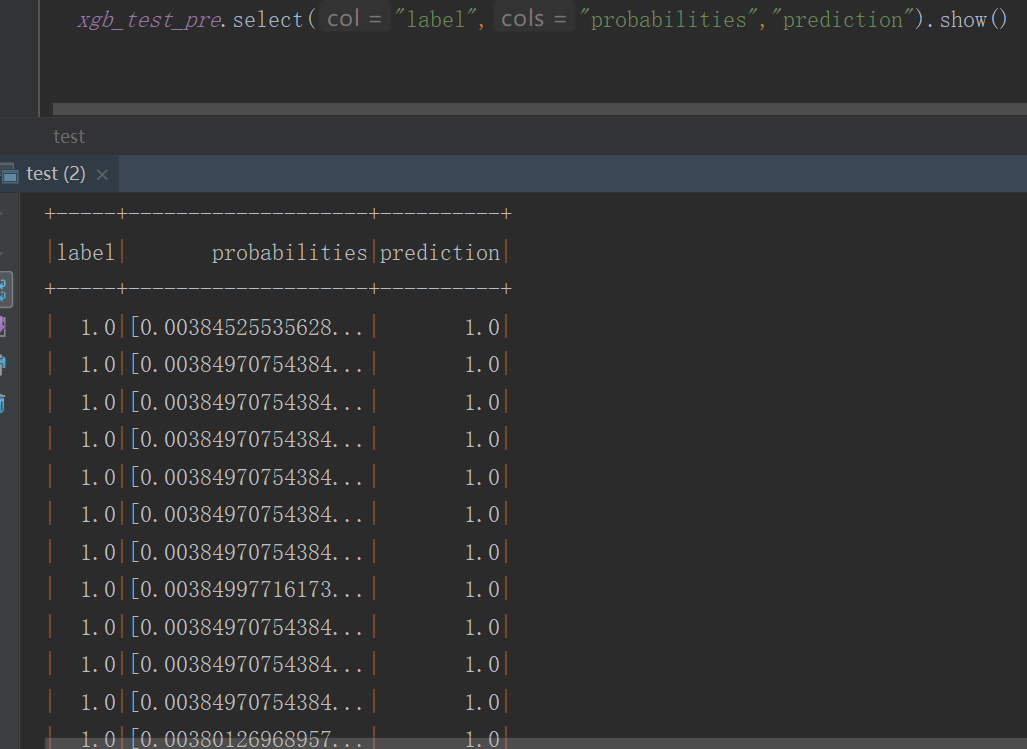




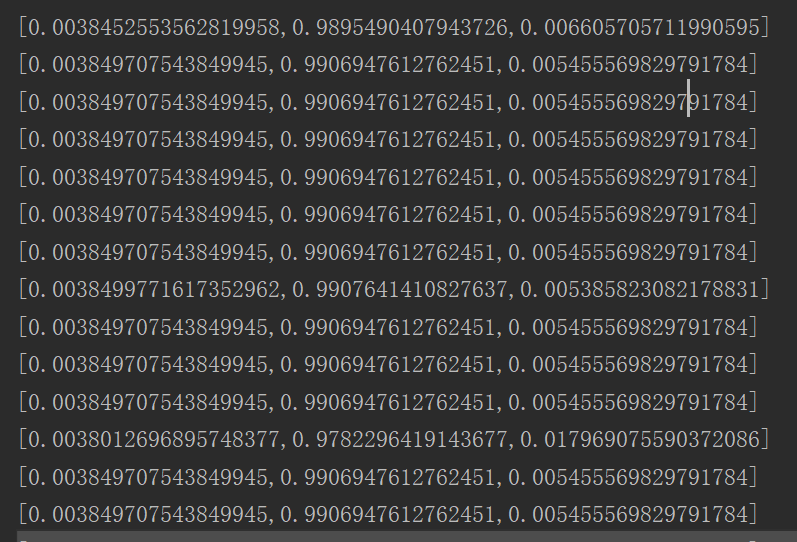


（2）当参数objective设置为softprob时，transform中probabilities和prediction的输出结果为





其中，probabilities列的输出为



区别为：当设置为softmax时，transform方法生成的probabilities值是分类结果值，所以其prediction都会是0(用predict方法的结果也会是0)；而当参数设置为softprob时，transform方法生成的probabilities值是各个类别的概率值。

32、（1）Error:(108, 46) could not find implicit value for parameter sparkContext: org.apache.spark.SparkContext

val model = XGBoost.loadModelFromHadoopFile("hdfs://localhost:9000/work/xgboost/saving\_model/saving\_model")

（2）Error:(108, 46) not enough arguments for method loadModelFromHadoopFile: (implicit sparkContext: org.apache.spark.SparkContext)ml.dmlc.xgboost4j.scala.spark.XGBoostModel.

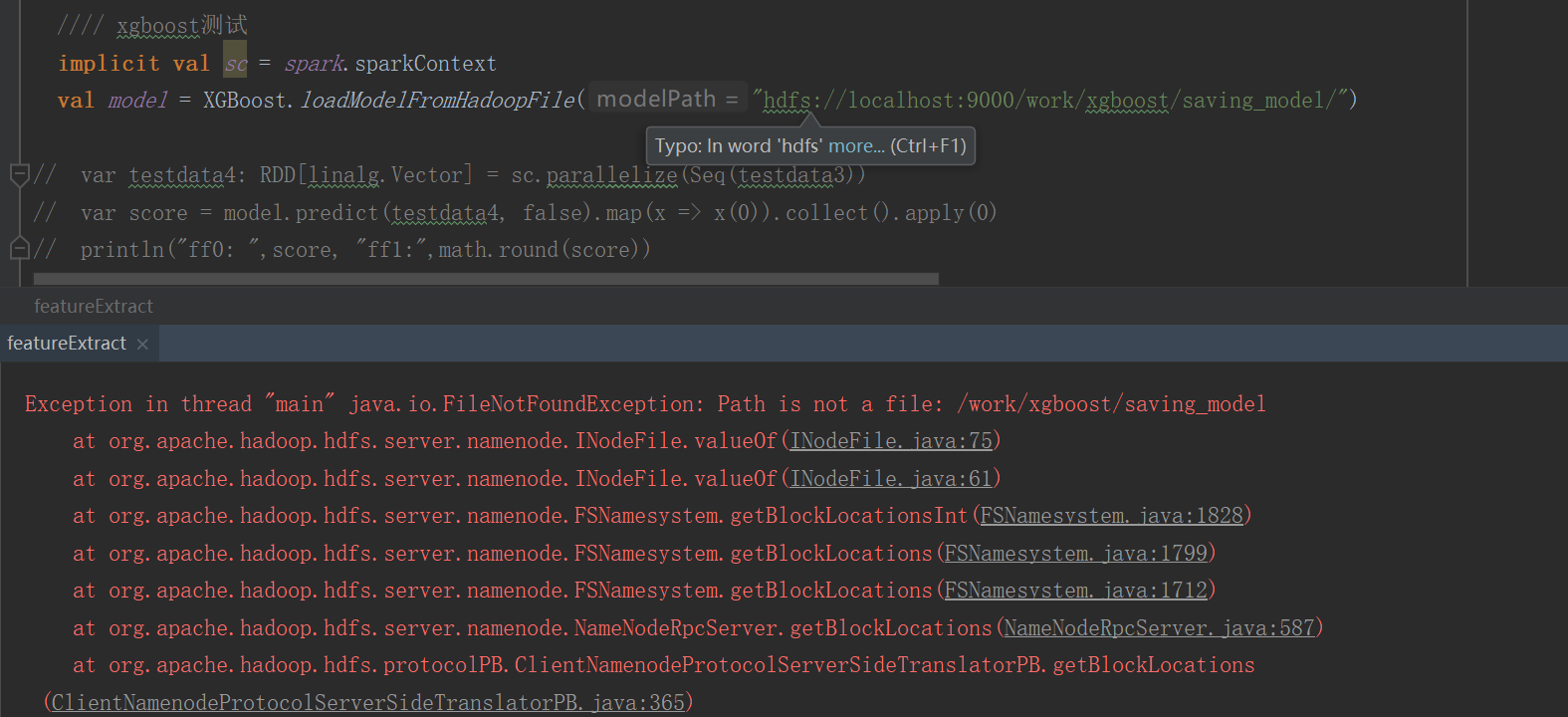
Unspecified value parameter sparkContext.

val model = XGBoost.loadModelFromHadoopFile("hdfs://localhost:9000/work/xgboost/saving\_model/saving\_model")



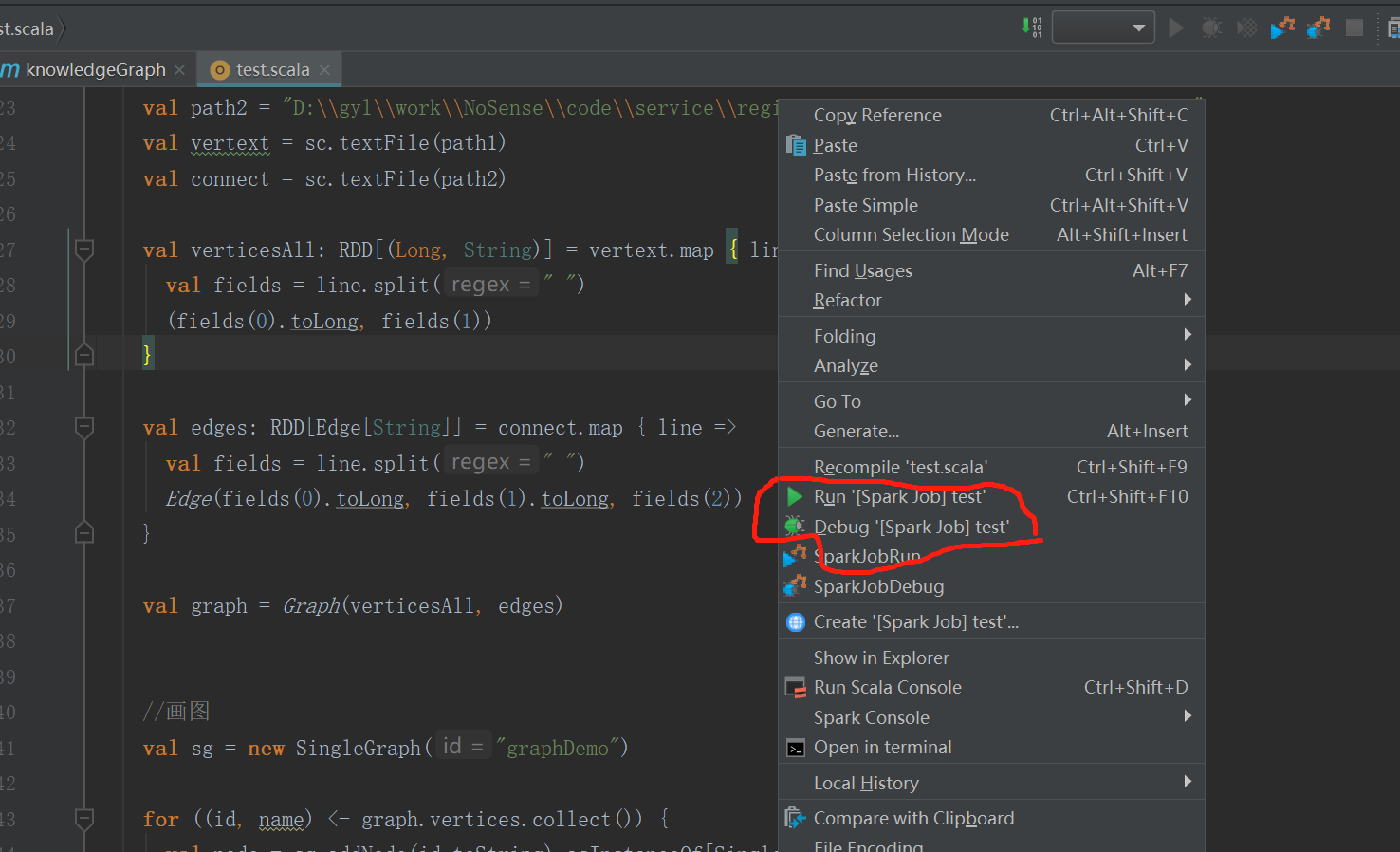
解决方法：加上这句话：implicit val sc = spark.sparkContext

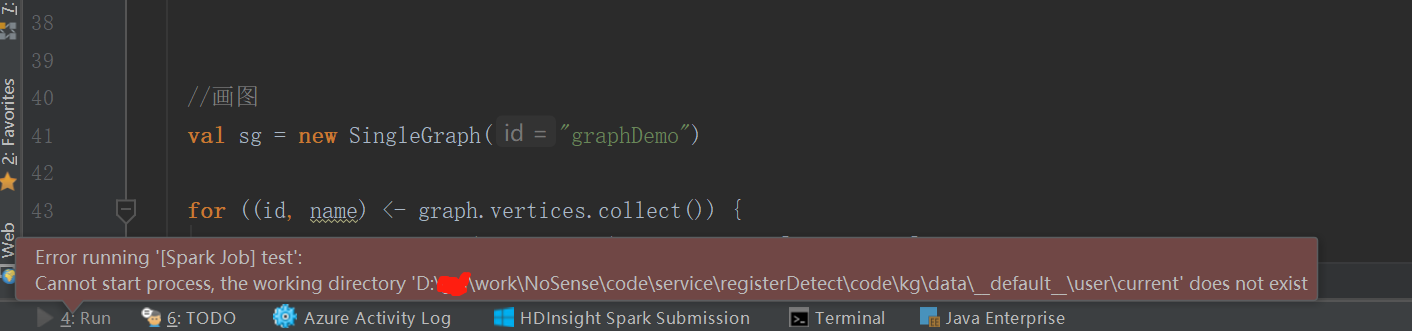
33、



解决方法：xgboost加载hadoop文件要具体到文件名。

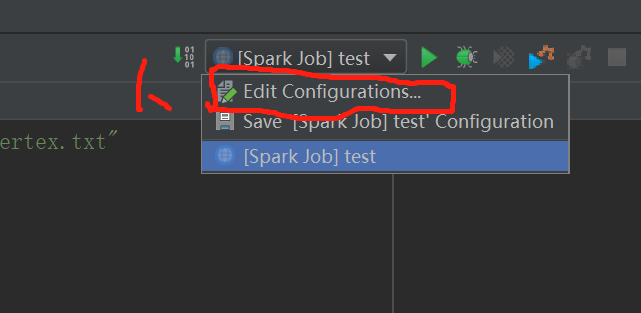
35、问题描述：右键点击Run的时候，test前面”[Spark Job]”，然后运行时报错：Error running '[Spark Job] test': Cannot start process, the working directory 'D:\work\NoSense\code\service\registerDetect\code\kg\data\\_\_default\_\_\user\current' does not exist.



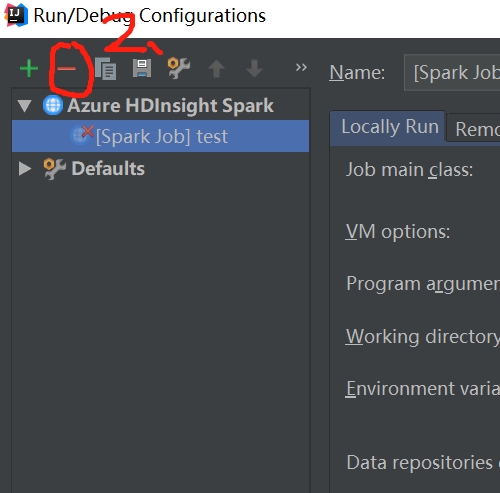


出现问题的原因：出现上述bug是因为没有配置好运行环境，具体做法如下：

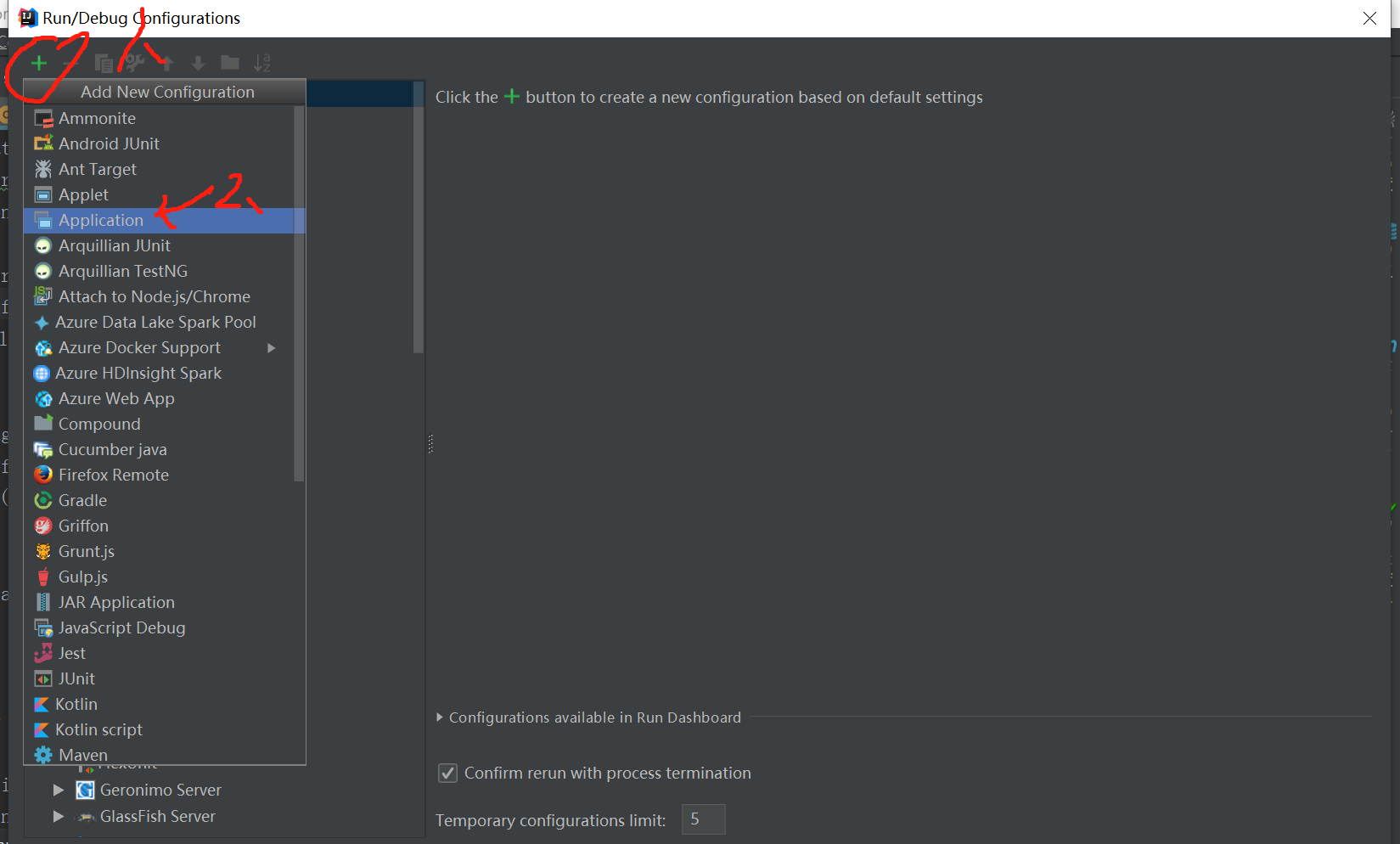
第一步：点击Edit Configurations；



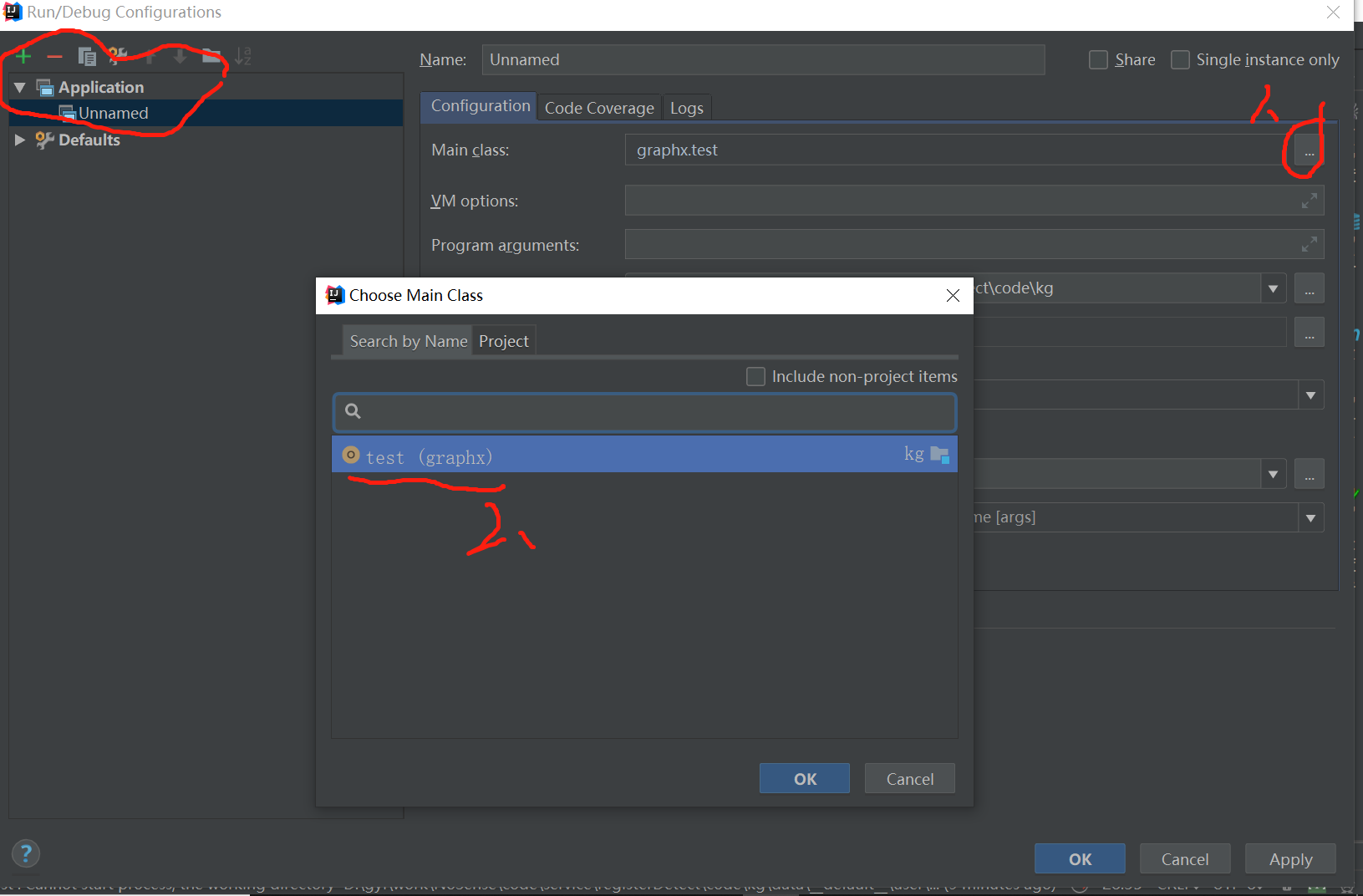
第二步：点击图中红色减号删除；



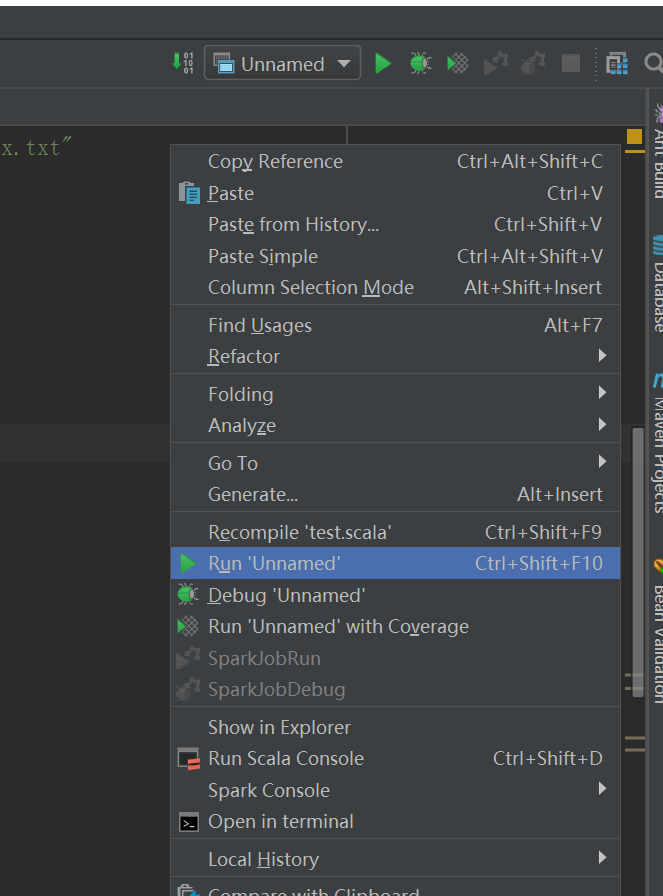
第三步，一定要选+号（之前在default里选择Application后，还是没用）



第四步，一定要把Main class改成程序里的主类名



出现下面的图，环境就ok了。



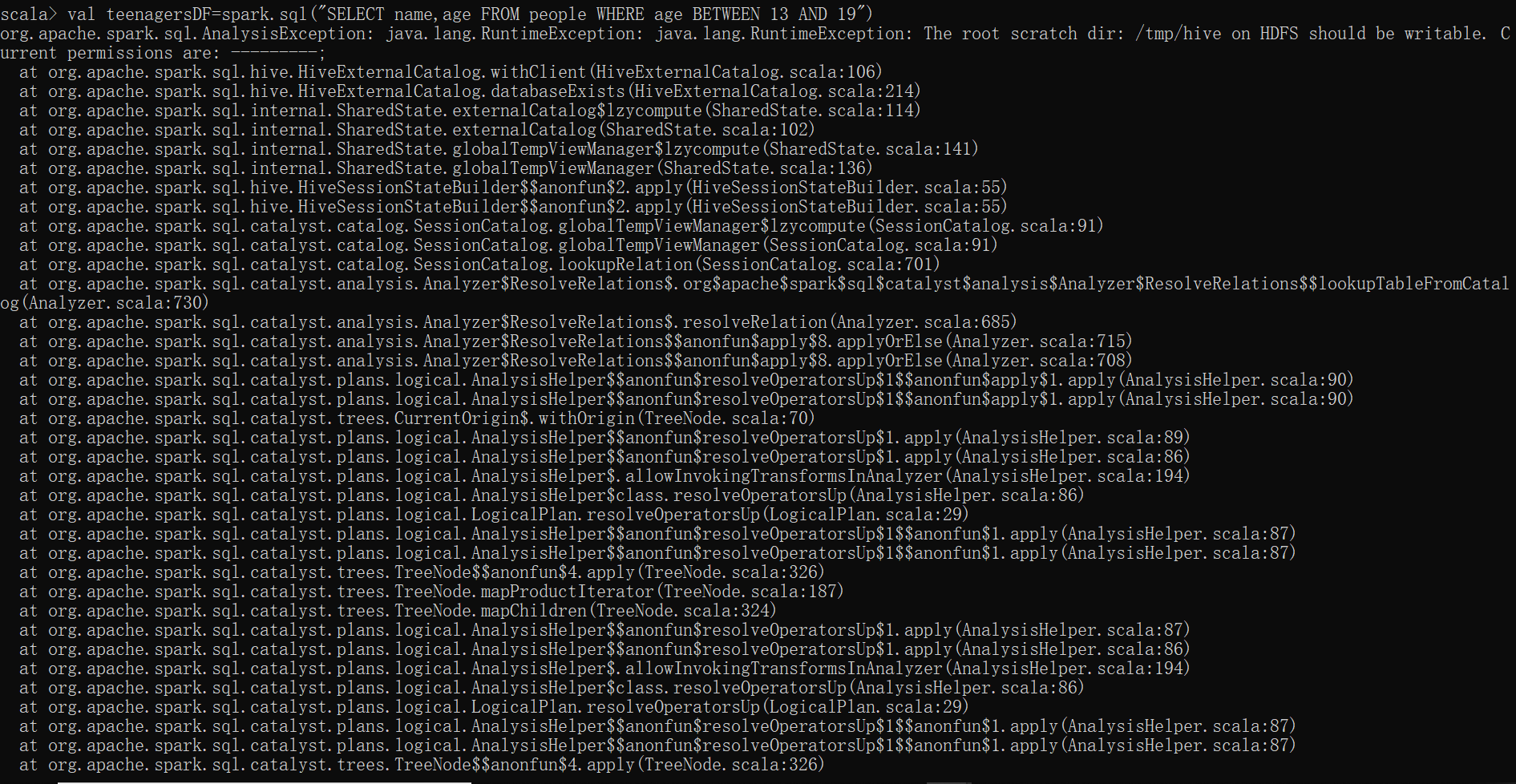
34、LabeledPoint函数的使用，在训练和预测方面有啥区别？

ml.dmlc.xgboost4j.LabeledPoint函数就是用来对训练数据进行标签化的。DMatrix中使用LabelsPoint主要是用来划分batch的。

35、scala+spark训练xgboost模型，特征数为30个，测试时，传递了50个特征，但是没有报错，而是返回了结果，python中测试后直接报错，这是为什么？

答：有可能spark-xgboost不会对特征数量进行判断。

36、org.apache.spark.sql.AnalysisException: java.lang.RuntimeException: java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable. Current permissions are: ---------;



Df0.join(df1,Seq(“FOMR\_CUST\_NUM”),”left”)问题

