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

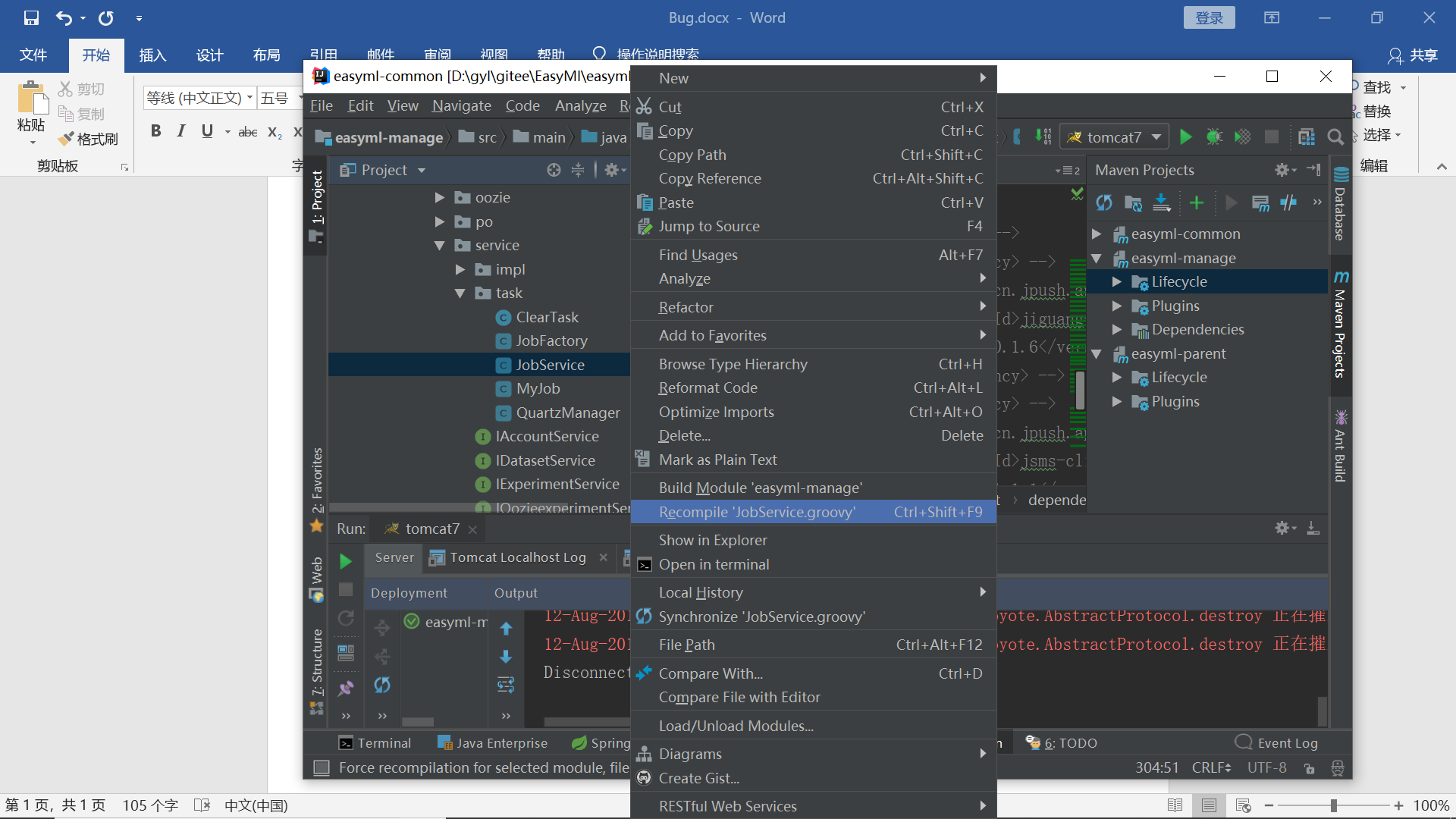
[ERROR] /D:/gyl/gitee/EasyMl/easyml-manage/src/main/java/com/trusfort/easyml/controller/interceptor/StartupListener.java:[22,5] 找不到符号

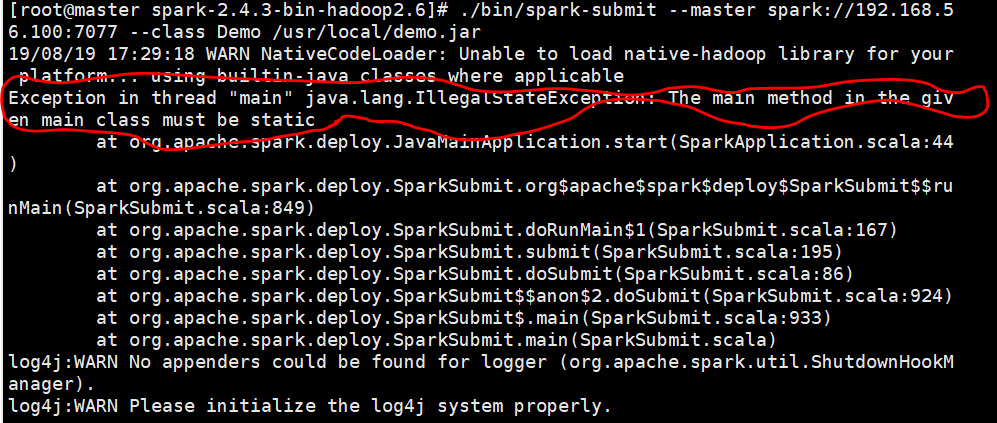
[ERROR] 符号: 类 JobService

[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：

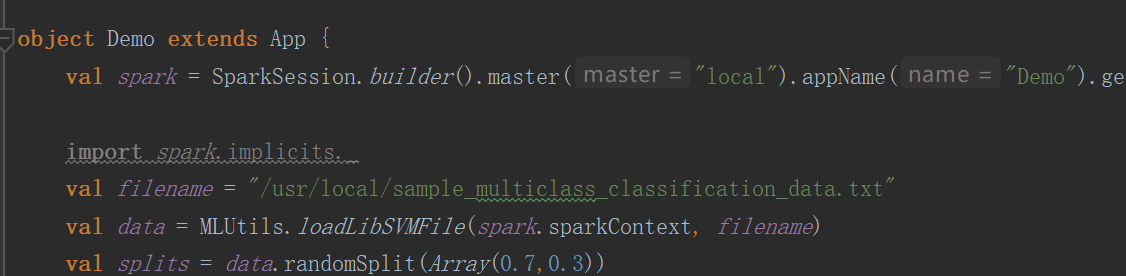
Class Demo {

def main(args : Array[String]) {

}

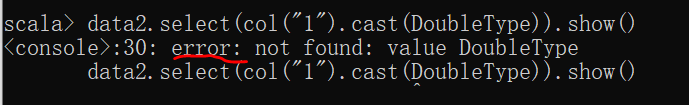
}

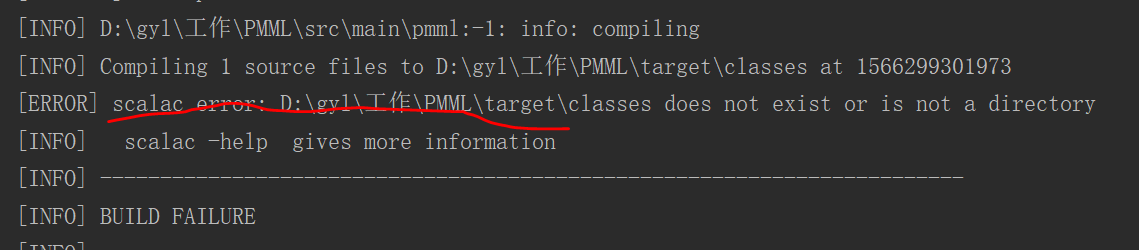
要把它改成这样：



3、下面的bug是因为没有写这句话：

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

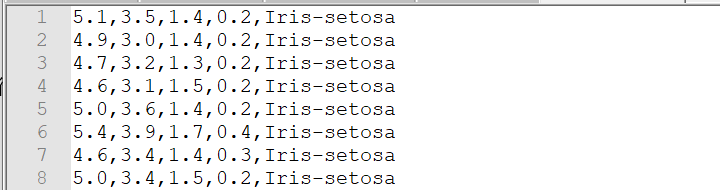


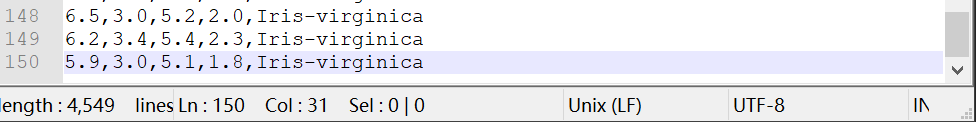
4、

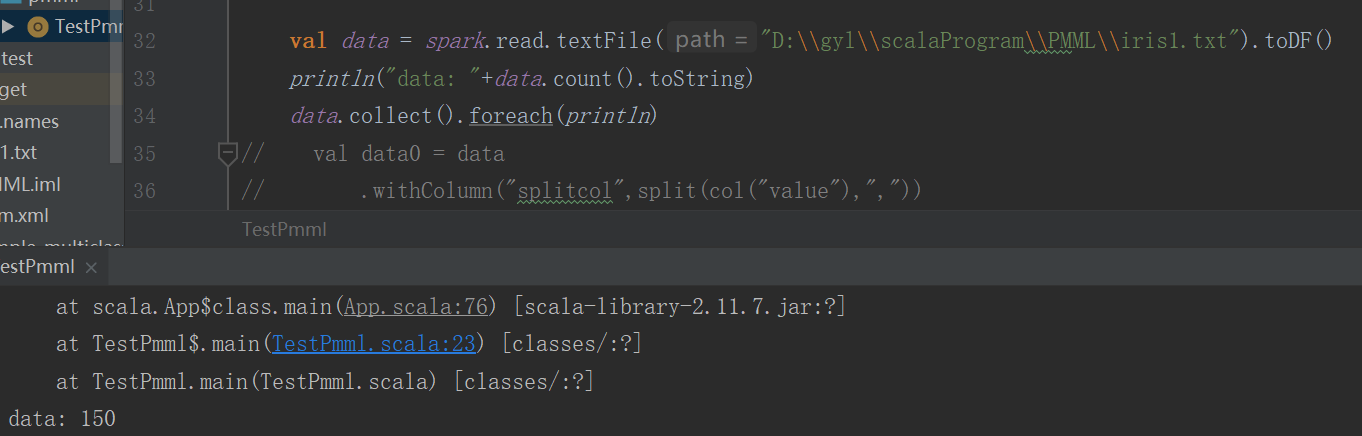
错误原因：工作目录中含有中文。

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。

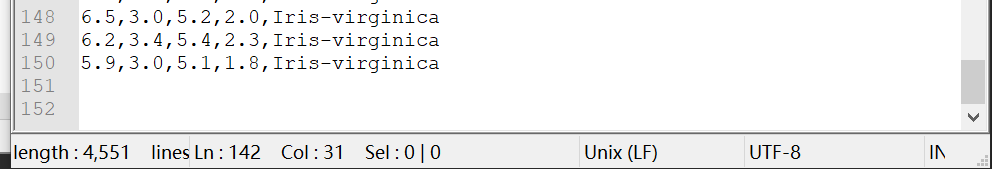
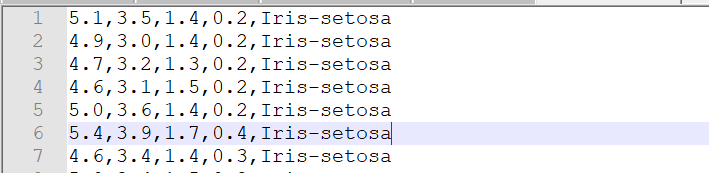
（1）txt文件最后没有空行，总共有150行

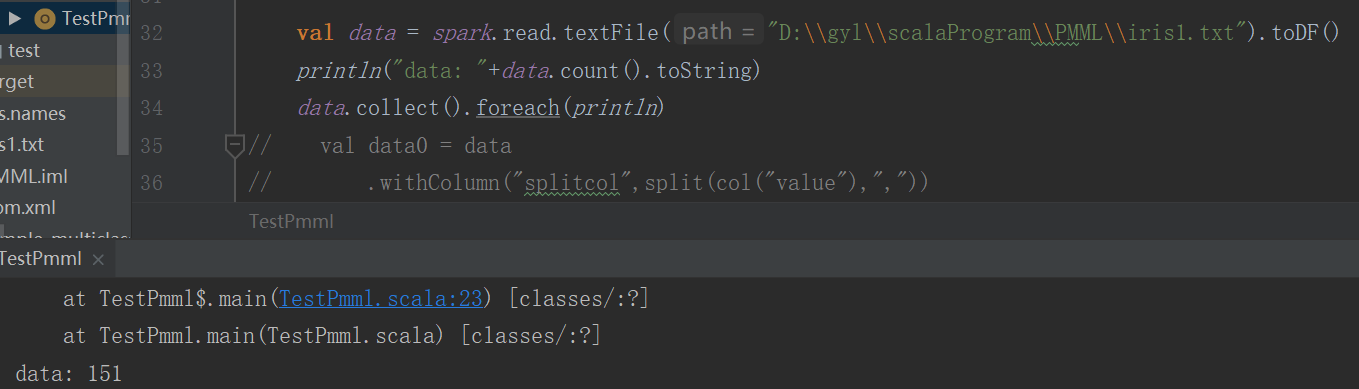






（2）txt文件最后有空行，spark中textFile读取了151行





再看看数据，最后一个空行被toDF操作转换为[]

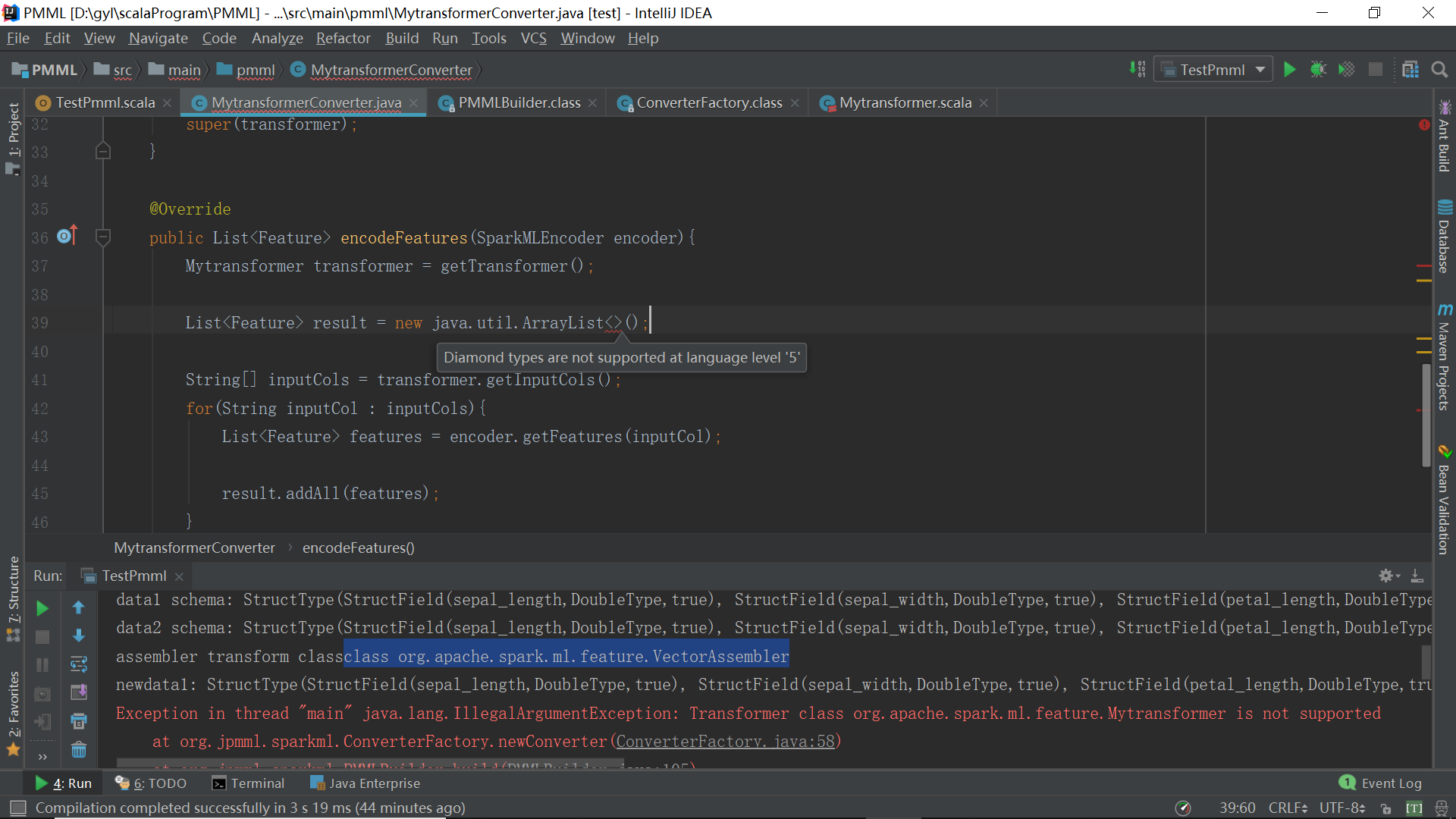


5、spark的dataframe中withColumn里函数不要写成这样col(“features”).apply(convertUDF)，会出问题（这里的apply的用法不同）

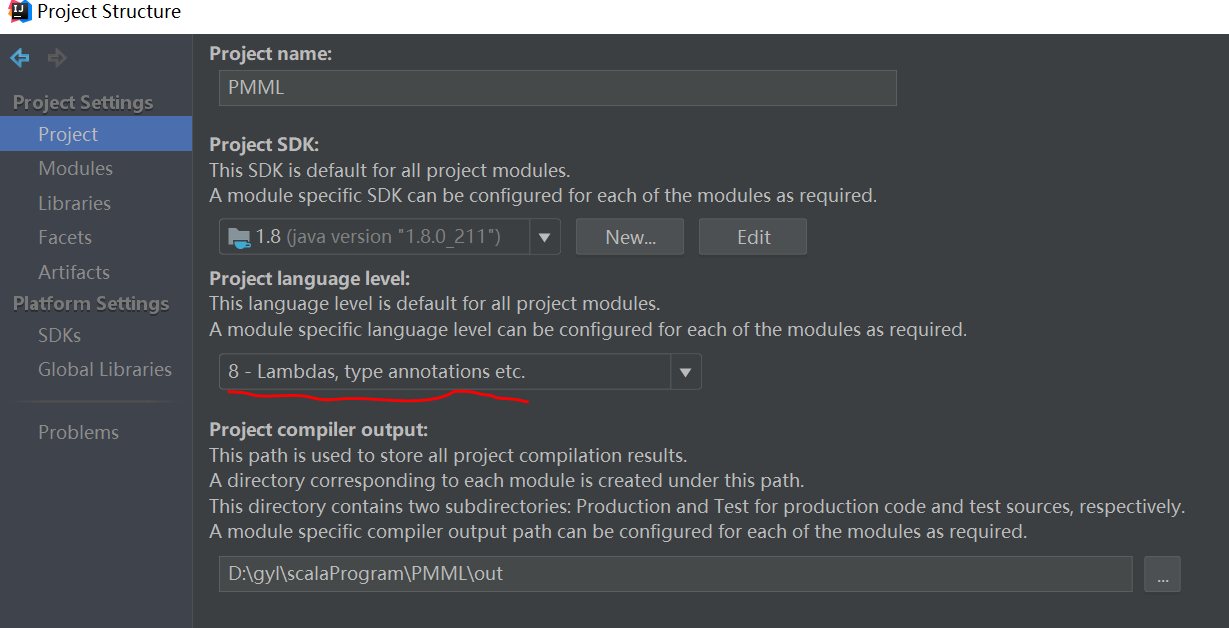
要写成下面这样就没问题：



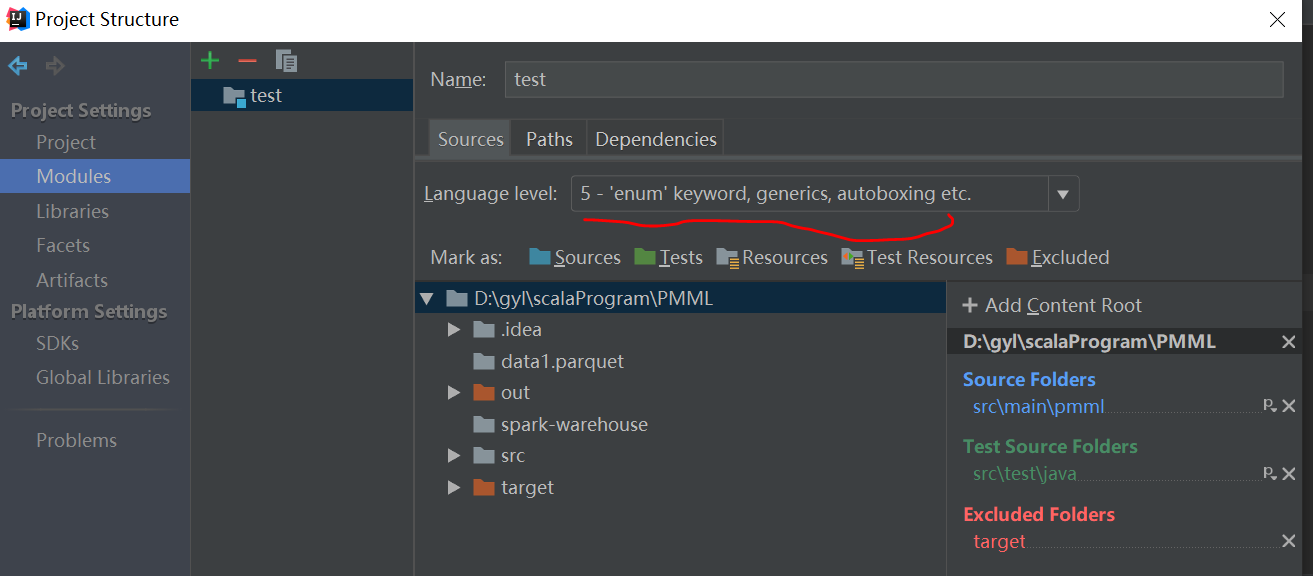
6、Idea中JDK为1.8，还提示Diamond types are not supported at this language level。



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

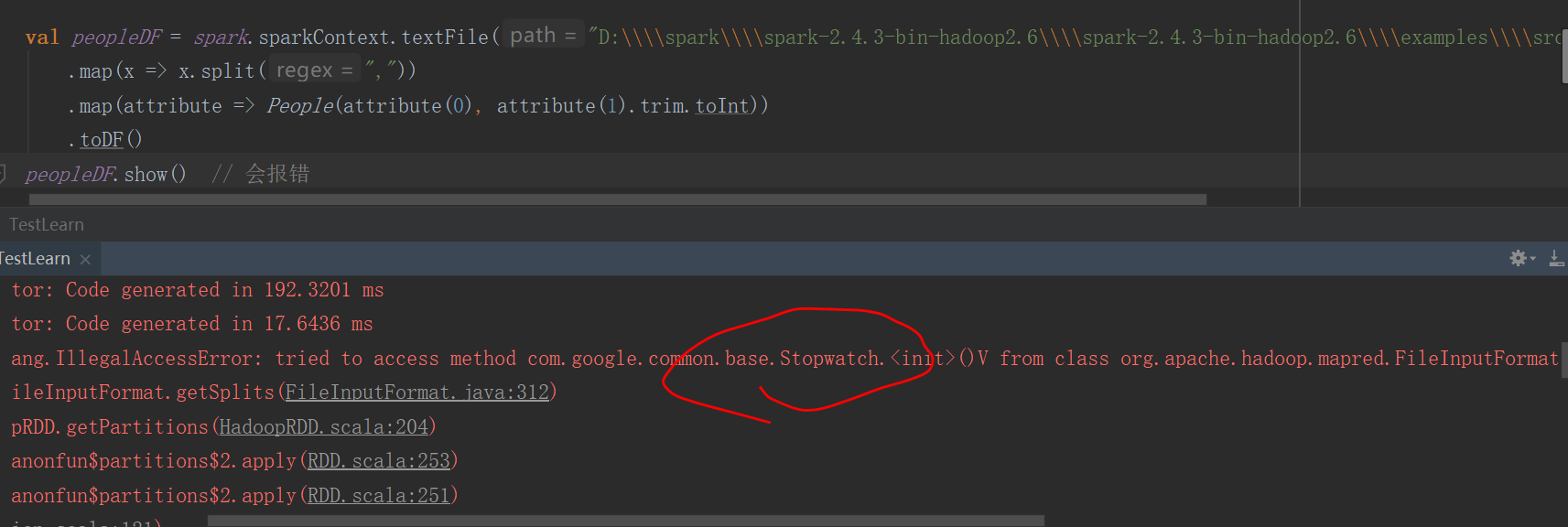


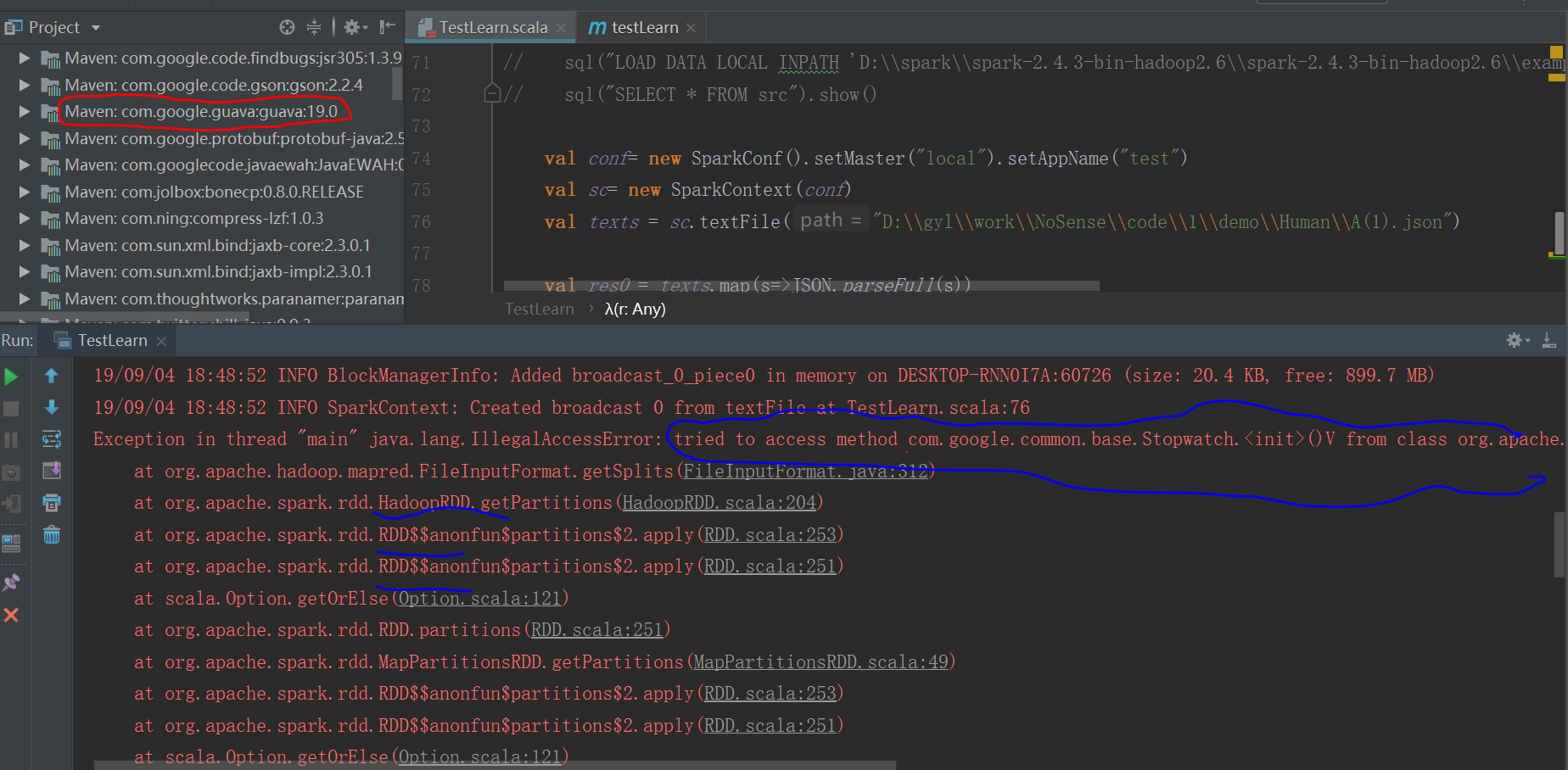
然后继续检查其他模块 如modules ，发现了问题所在。



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>

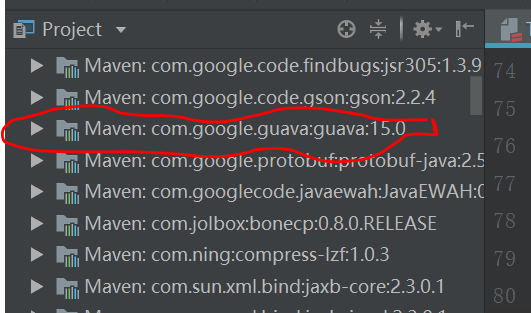
<groupId>com.google.guava</groupId>

<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>

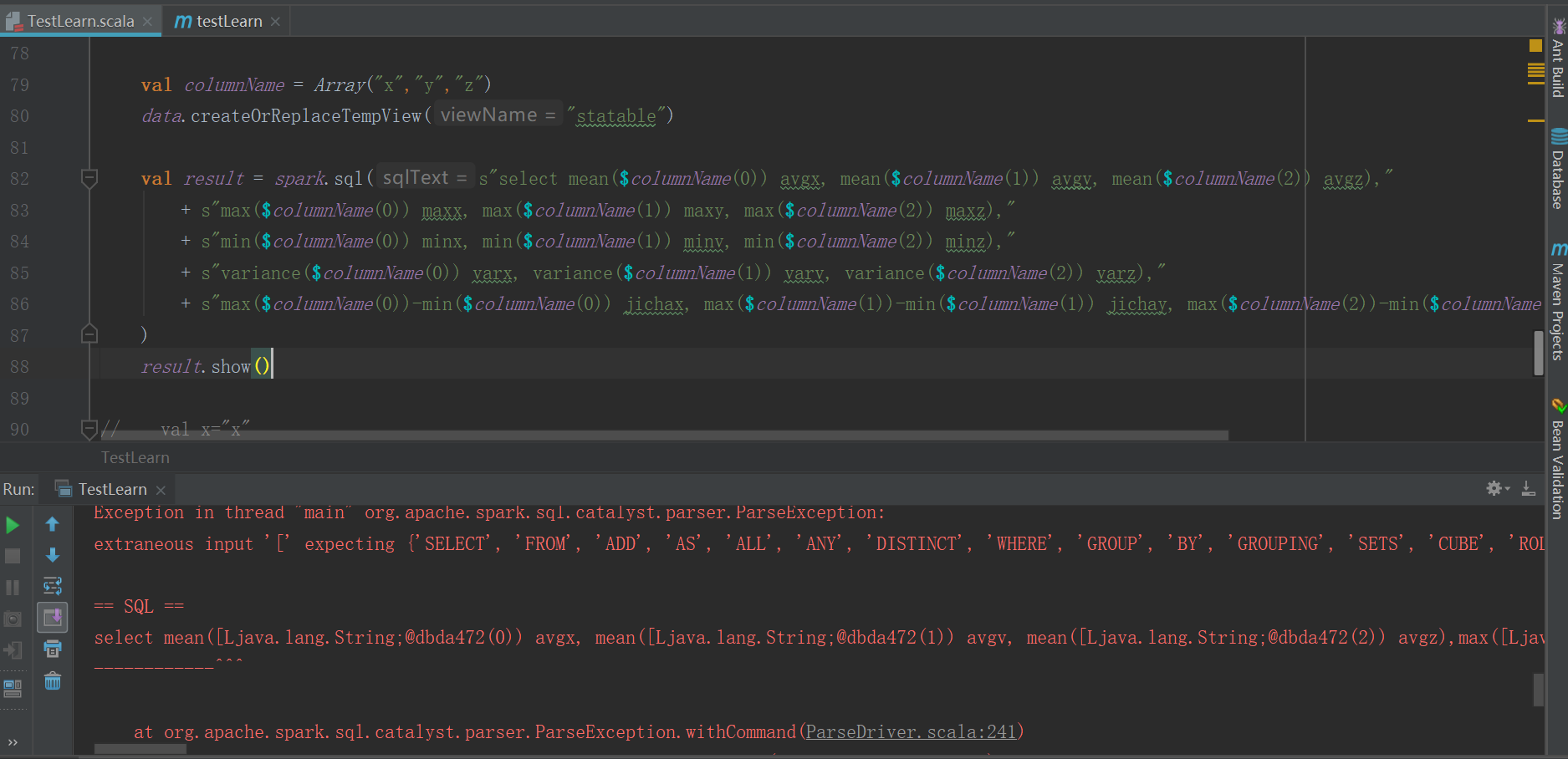
<groupId>org.apache.spark</groupId>

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

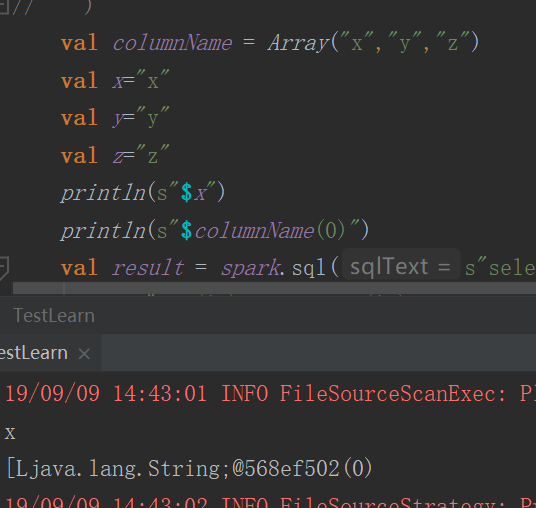
<version>${spark.version}</version>

</dependency>

9、在对dataframe的每列计算统计特征时，报如下错误：

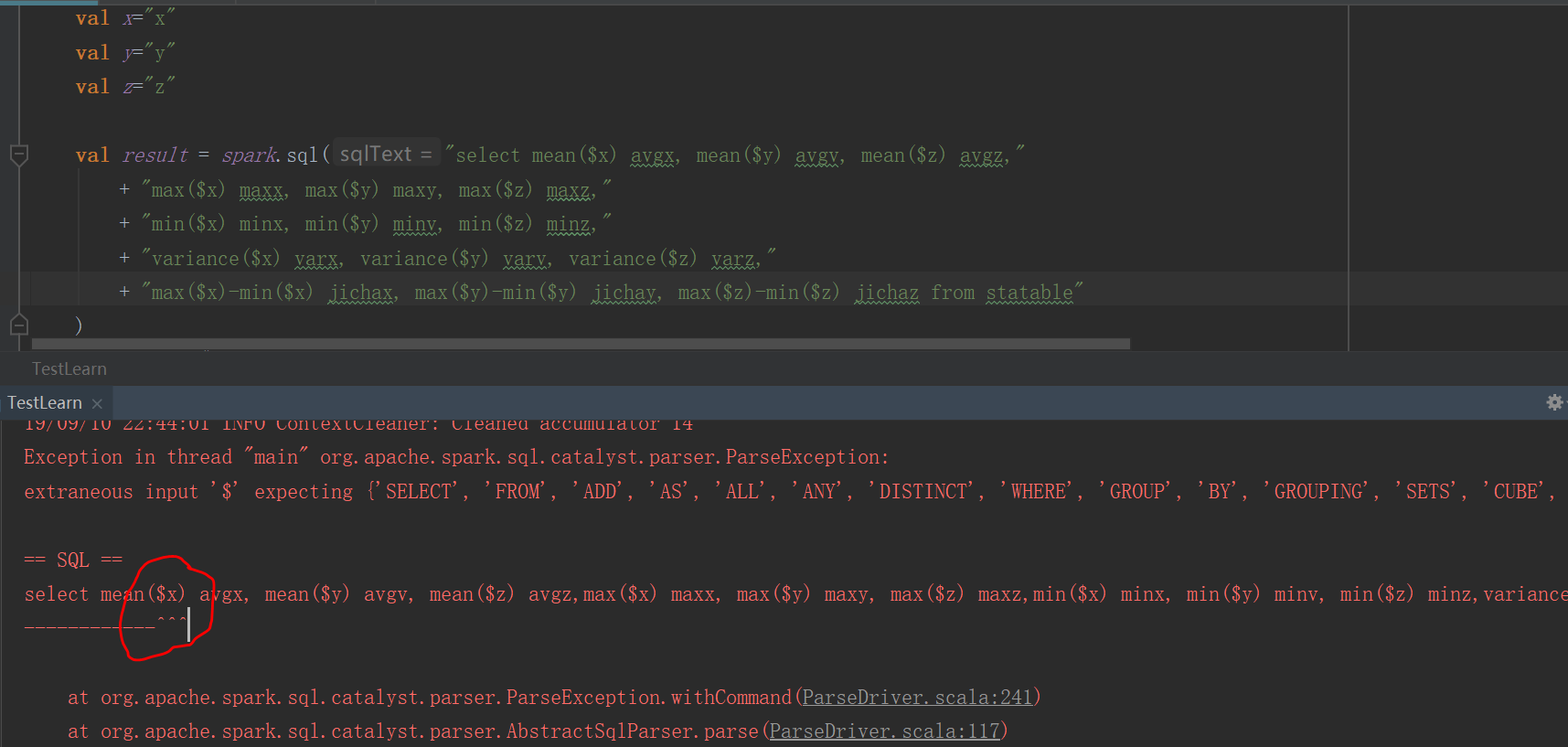


错误原因：scala中“$”后面接一个变量名以提取变量值，而这里columnName(0)的值是字符串。

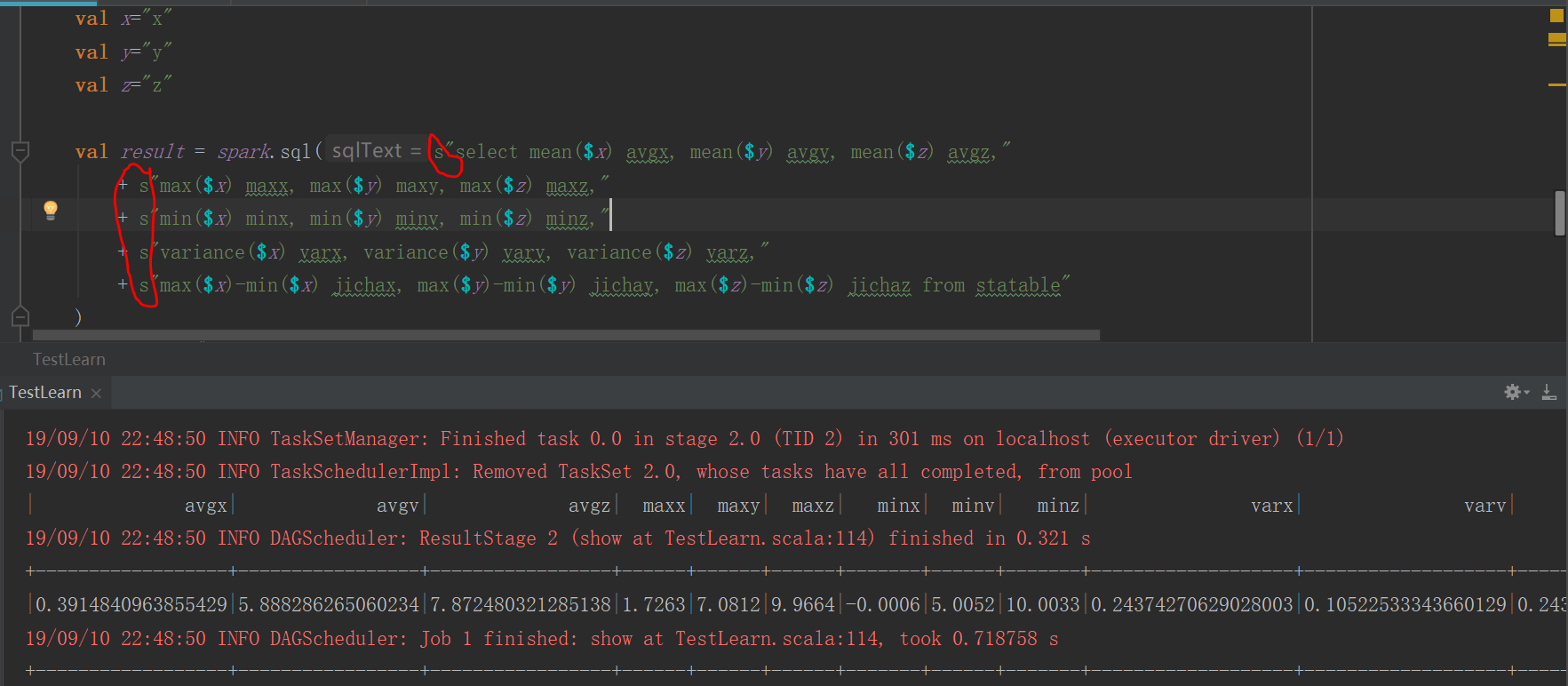


解决方案：将”$”去掉也没用，因为将“$”去掉后，在双引号里面的columnName(0)就表示一个字符串，而不是”x”。最后的解决方法是分别定义，然后用”$”引用。

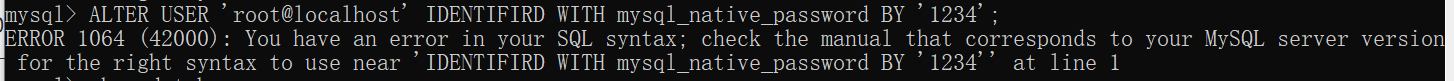
10、关于spark.sql中字符串前面s的问题，如果不加s，则会报错：无法解析变量



解决方法：在字符串前面加个s

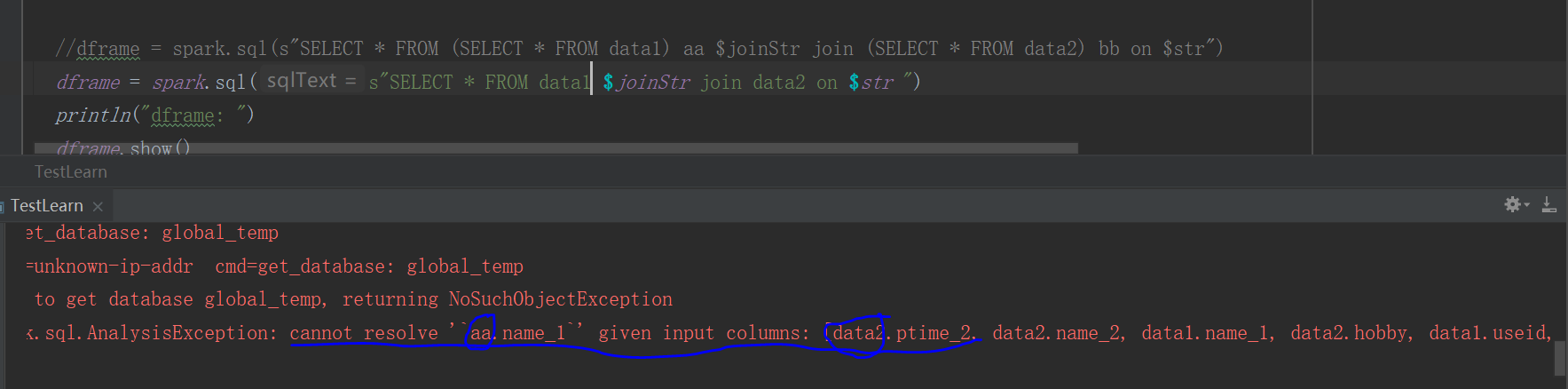


11、mysql中第一次修改密码出现的问题：



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.



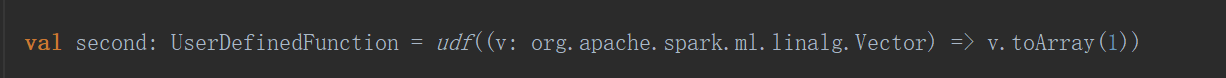


解决方法：为表加别名。

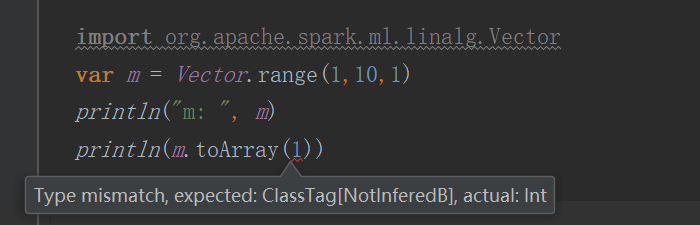


13、scala中vector和spark中vector的问题：

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：

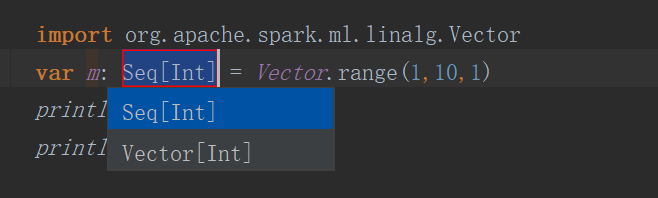


然后自己敲了一下，发现toArray带了参数报错，如下图，为什么会这样？

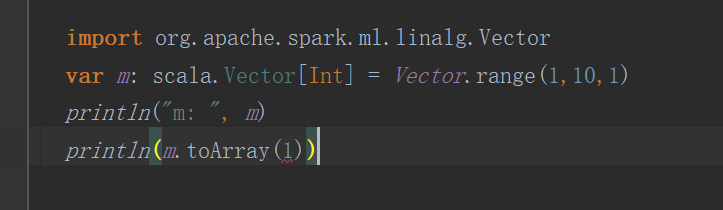


然后，在变量m处按快捷键Alt+Enter查看变量类型，如下图，有两种类型，无论选哪种类型，带参数的toArray都会报错。

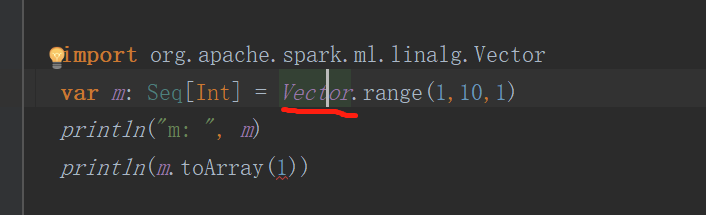
类型1：Seq[Int]



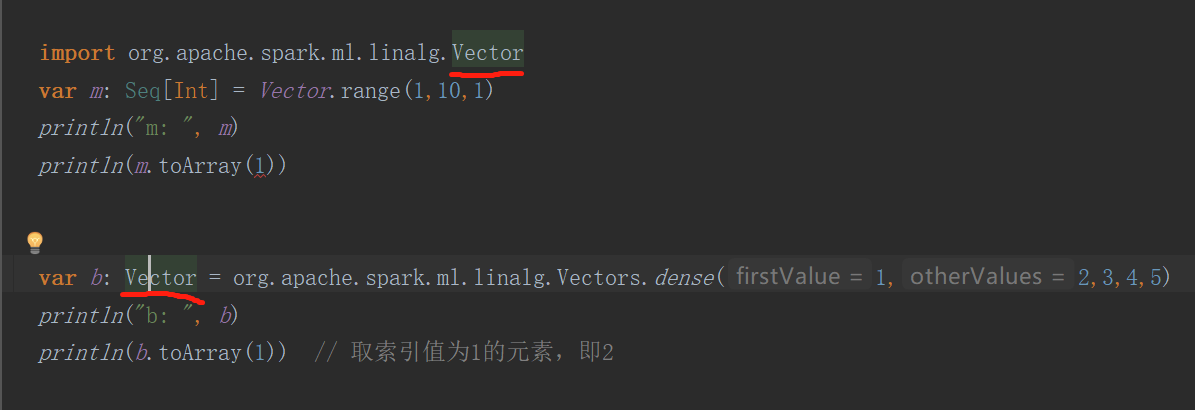
类型2：scala.Vector[Int]



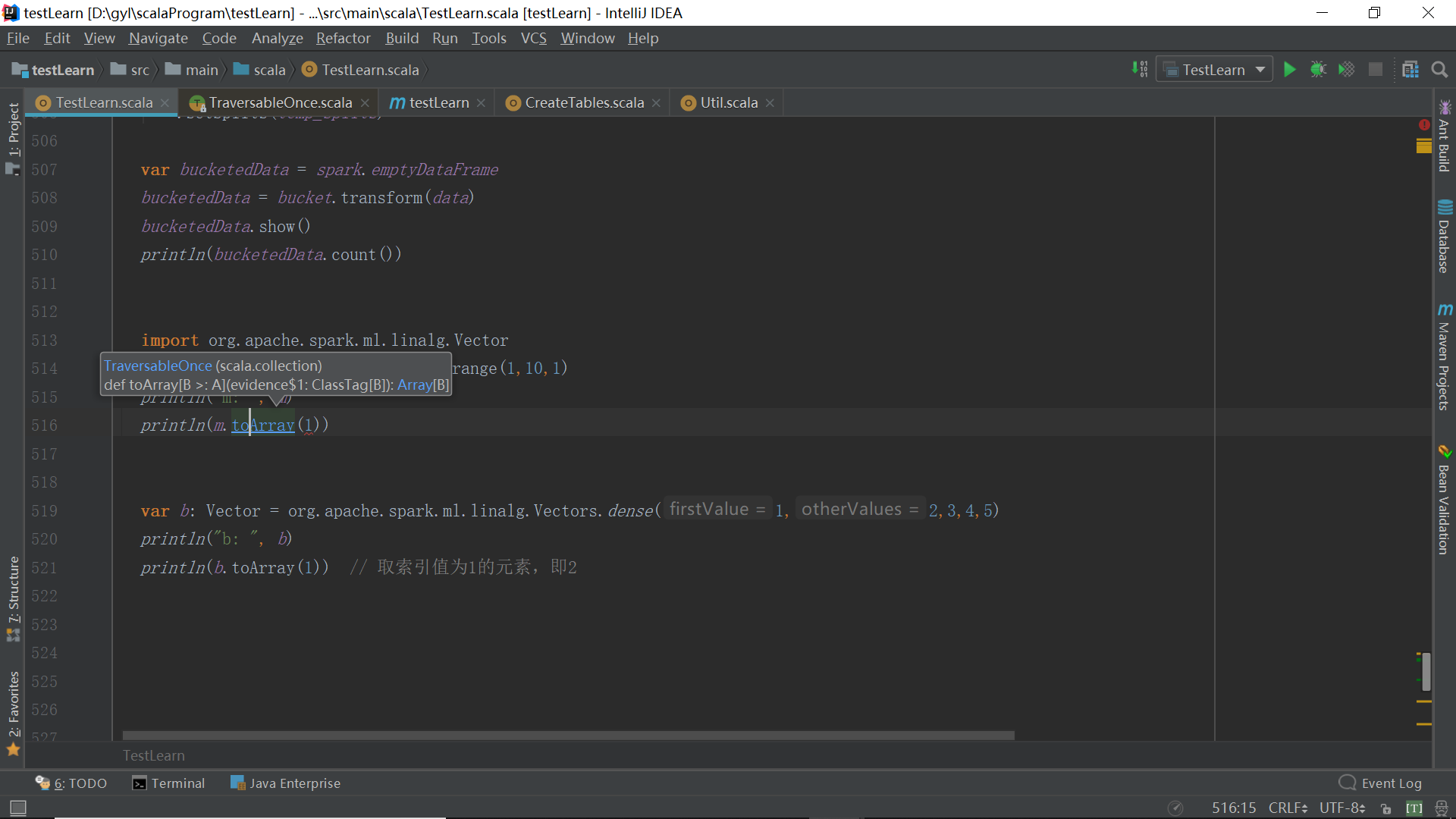
然后，点击Vector发现该Vector并不是import进来的Vector，如下图



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

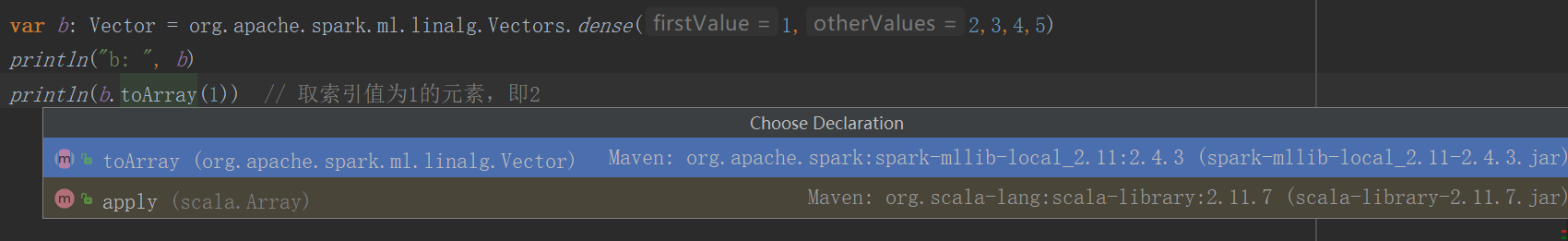


为什么直接把包名带上定义的变量b是Vector型，而变量m是Seq[Int]或scala.Vector[Int]。是因为变量m里的Vector实际是scala中的容器，并不是spark中的vector。在m.toArray(1)上按Ctrl+鼠标点击toArray也可以看出该toArray函数是容器中的函数，如下图



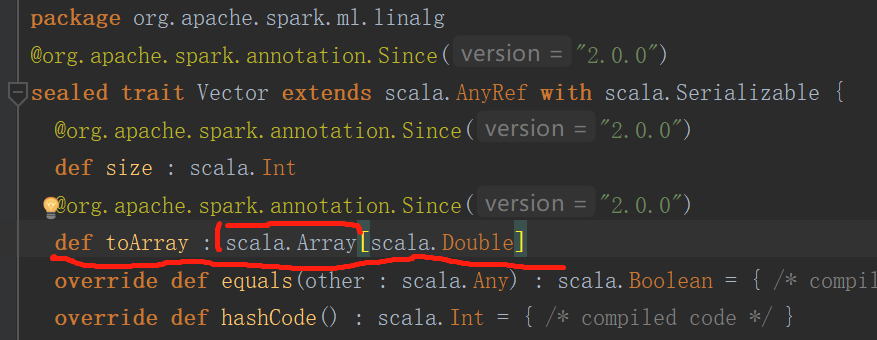
从上面可以看出，即使将spark中Vector给import进来，但是直接写vector还是调用的scala中的容器vector。

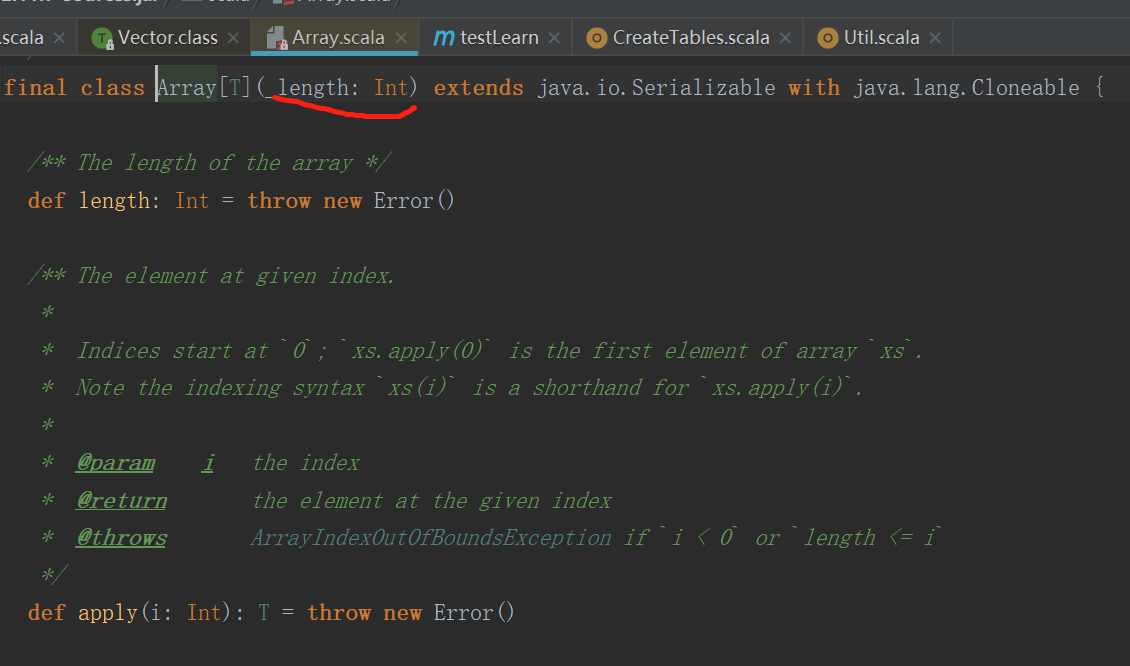
到这里，问题还没解决，就是实际spark的API中Vector.toArray是不带参数的，这里带参数为啥不报错？



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：





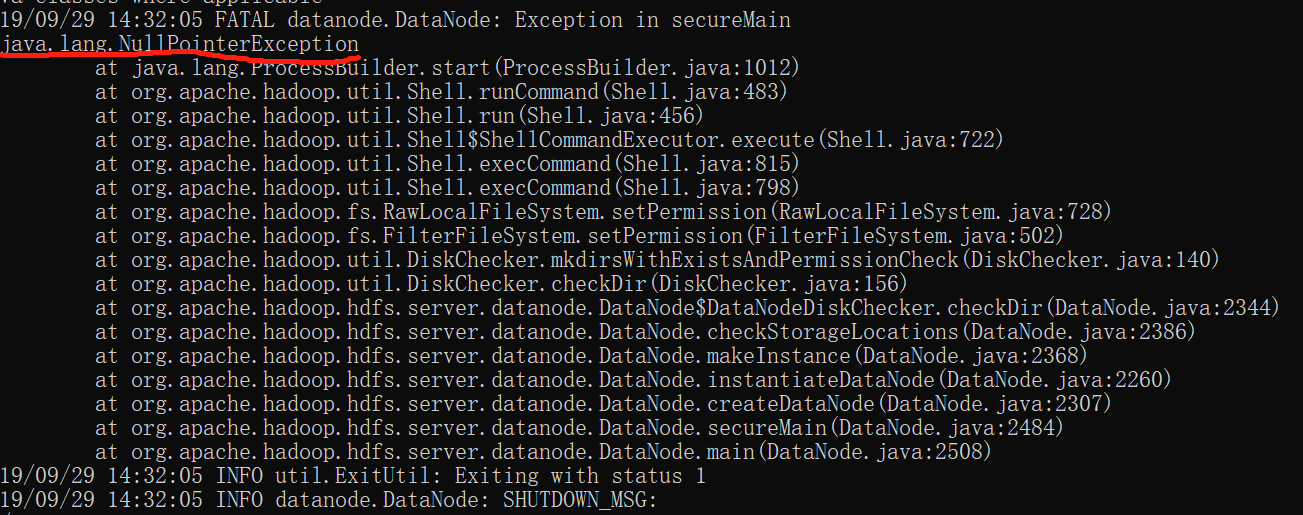
不过，说明一个问题，就是spark中vector的toArray方法是可以带参数的，参数表示元素的索引，scala中vector容器的toArray方法不能带参数。

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:



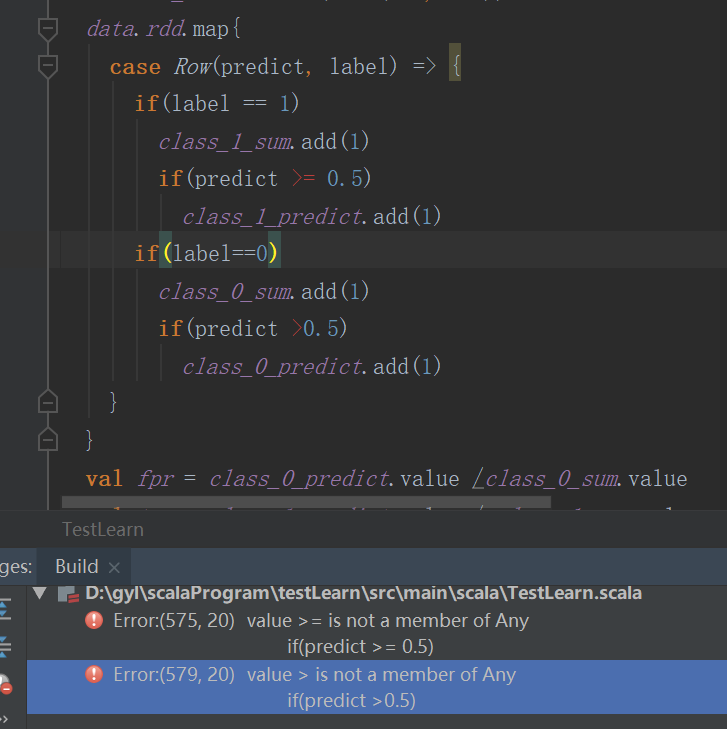
（2）idea中运行程序报错：



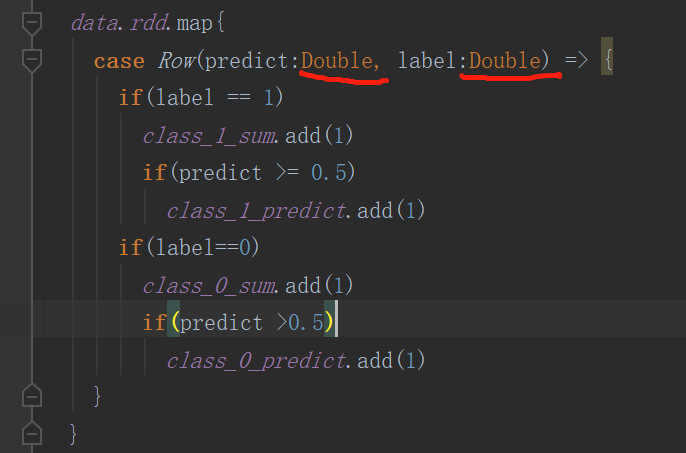


出现以上问题的原因是hadoop缺乏在windows上运行的必要组件，像winutils.exe等，这些组件的下载地址：<https://github.com/steveloughran/winutils>，从winutils-master\hadoop-2.7.1目录中直接拷贝bin文件夹，覆盖掉hadoop-2.7.1/文件夹中的bin目录即可。

16、出现如下报错：比较符号无法解析

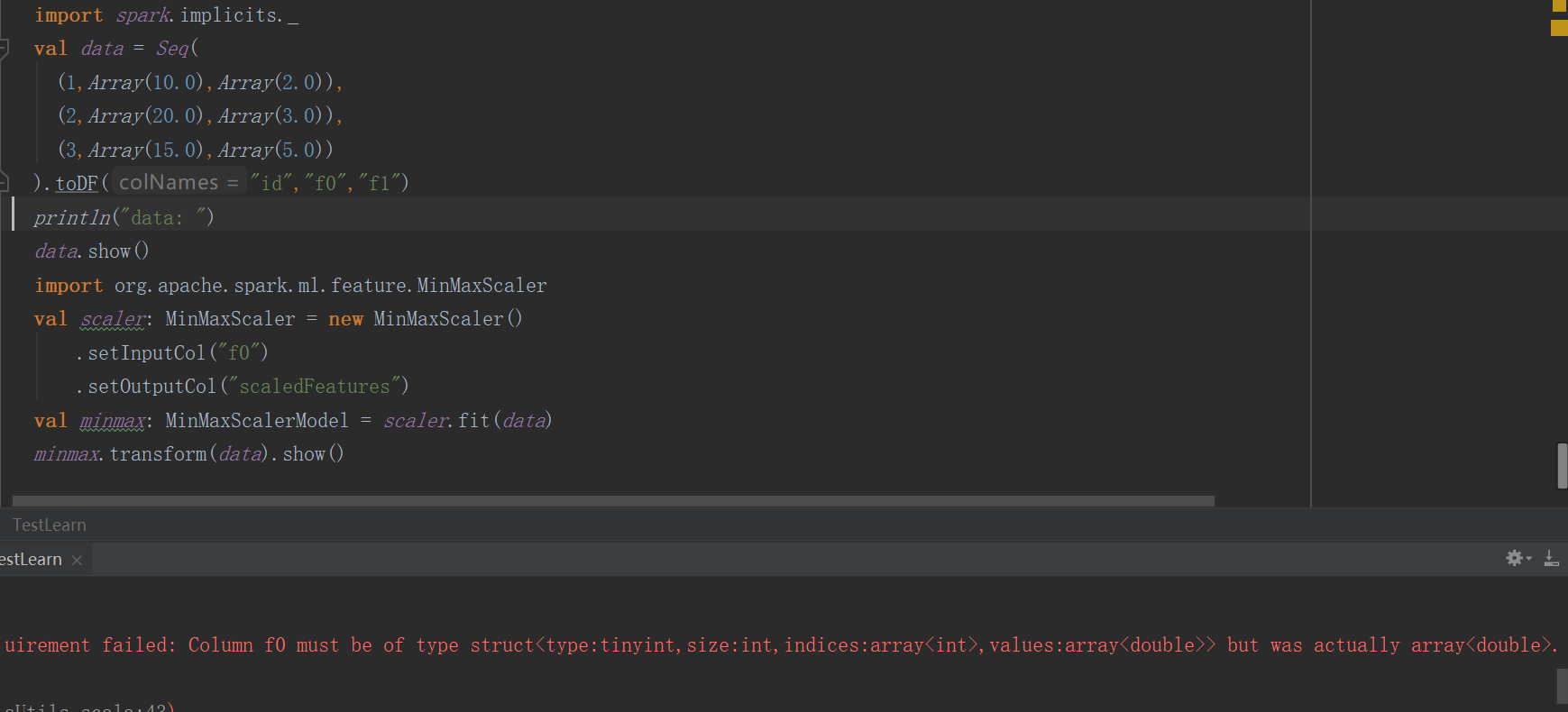


解决方法：根据提示，比较符两边的值是Any类型**(对于一个未定义类型的变量，scala默认为Any型)**，解决办法如下：



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。

