# 基于Window和IntelliJ构建Spark本地开发调试环境

## 引言

用于在window下调试spark代码。

环境：window7 + java1.8 + scala10.6

安装：IntelliJ，<https://www.jetbrains.com/idea/download/#section=windows>

选window版本，可以选择是社区版还是付费版，付费版只能用一个月，之后每次只能用半小时，建议社区版。

下载：spark，<http://spark.apache.org/downloads.html>

Spark1.6.0，下载了两个版本一个hadoop-2.4，一个source code。

## 搭建环境

Step1.建立文件，并建立output 的文件夹，通常是project父目录下建立一个out文件夹，作为本项目的输出文件夹。

Step2.建立project，设定jdk和scala，以及output path



备注：这里output path如果不设定，可能后面导致。Cannot start compilation: the output path is not specified for module的错误。

Step3.加入scala到 library



Step4.将spark的jar加入到library



在spark也可以放入libararies里，这里导入spark-assembly的jar即可。另外在右边的红圈的“+”可以导入soure code，以便后面可以查看spark的源码。此处source code导入的文件夹为引言中下载的source code并解压的文件夹。

建议：可以在project文件夹下建立一个lib文件夹，将需要的jar等文件放进去，将解压的source code放进去。

Step5.建立module





Step6.建立Package和Class或Object





Step7.测试编译环境

在Object中输入代码

packagetestCore

importorg.apache.spark.{SparkConf, SparkContext}

importorg.apache.spark.sql.SQLContext

importorg.apache.spark.streaming.{Seconds, StreamingContext}

/\*\*

\* Created by DataShoe on 2018/1/5.

\*/

objecttestCoreCompile {

def main(args: Array[String]): Unit = {

// Test compile.

println("Hello world!")

// Test SparkContext

valconf = new SparkConf().setMaster("local").setAppName("DataShoe-Learning")

valsc = new SparkContext(conf)

valsqlc = new SQLContext(sc)

valrdd = sc.parallelize((0 until 100).toList)

valrdd\_sum = rdd.reduce(\_ + \_)

println(rdd\_sum)

}

}

查看，编译是否成功

备注1：如果测试的object能够编译成功，但测试SparkContext时不成功，报错



则是scala和spark的版本不匹配，该版本spark可能要求2.10版本的scala。

备注2：如果出现找不到spark对应类的情况



请在module中加入libray



## 建立一个自动设定SparkContext的抽象类

### 背景

如果每次都设定master和appName太麻烦，可以建立一个基本的module，里面建立一个抽象类myApp，使得其他module可以引用该，不用每次都设定SparkContext。

### 步骤

Step1.建立新的module，名称为base



Step2.将spark加入到该module

将需要的spark的lib add到该module中来,方法同前面一样。Libraries/global libray => add to modules

Step3.建立对应package和测试package，并在其中建立抽象类的和对应的测试object

1）建立两个package：com.self.myApp和com.self.testMyApp，分别用于建立抽象类和测试抽象类

2）在新建的module base中建立com.self.myApp的Package，在其中建立类myAPP

packagecom.self.myApp

importorg.apache.spark.{SparkConf, SparkContext}

importorg.apache.spark.sql.SQLContext

importorg.apache.spark.streaming.{Seconds, StreamingContext}

/\*\*

\* Created by DataShoe on 2018/1/5.

\*/

abstract class myAPP {

lazyvalconf = new SparkConf().setMaster("local").setAppName("DataShoe-Learning")

lazyvalsc = new SparkContext(conf)

lazyvalsqlc = new SQLContext(sc)

lazy valsmc : StreamingContext = new StreamingContext(conf, Seconds(1))

def run(): Unit

def main(args: Array[String]): Unit = {

run()

}

}

3）在另一个Package，com.self.testMyApp中建立Object——testMyAPP测试该类是否成功

packagecom.self.testMyApp

importcom.self.myApp.myAPP

/\*\*

\* Created by DataShoe on 2018/1/5.

\*/

objecttestMyAPP extends myAPP{

override def run(): Unit = {

// Test compile.

println("Hello World")

// Test SparkContext.

valrdd = sc.parallelize((0 until 100).toList)

valrdd\_sum = rdd.reduce(\_ + \_)

println(rdd\_sum)

}

}

Step4.将base加入core的依赖中。



Step5.在core中建立一个object测试是否成功

在core中建立一个常用的package：com.self.core，在其中建立子package，testBaseApp，此时目录并在一块，不用担心，IntelliJ会自动合并只有单个子目录或子文件的目录，下次如果想继续在com.self.core下建子目录，比如newPackage，直接在core下面新建一个com.self.core. newPackage即可。

将代码贴进来



发现没有import myApp，Alt+Enter即可。

编译之，成功。

Spark2.1.0安装

1.如何导入jar：下载一个非source code版本解压，另外也下载source code（方便调试源码）解压，在global library或library中选择一个非source code中的jars文件夹（或者也可以只导入spark-assembly），将之导入进来。

2.错误：java.lang.NoSuchMethodError:scala.Predef$.$conforms()Lscala/Predef$$less$colon$less;

解决：需要scala2.12.2

## 异常处理

### 某些类不在Package中,但实际上却在

1. 一般的可能性

1)可能性1,缺乏正确了SDK或scala编译器

2)可能性2,缺乏正确的library

3)可能性3,缺乏正确的dependency

1. 如果以上都排除

有还有以下情况:

1）可能性1，缓存的index信息不对,可以file => invalid cache/restart

2）可能性2，.iml新建立，还没有把index信息加入进来。这种情况我出现在svn中clean up => update后把其他同事的iml更新进来了。

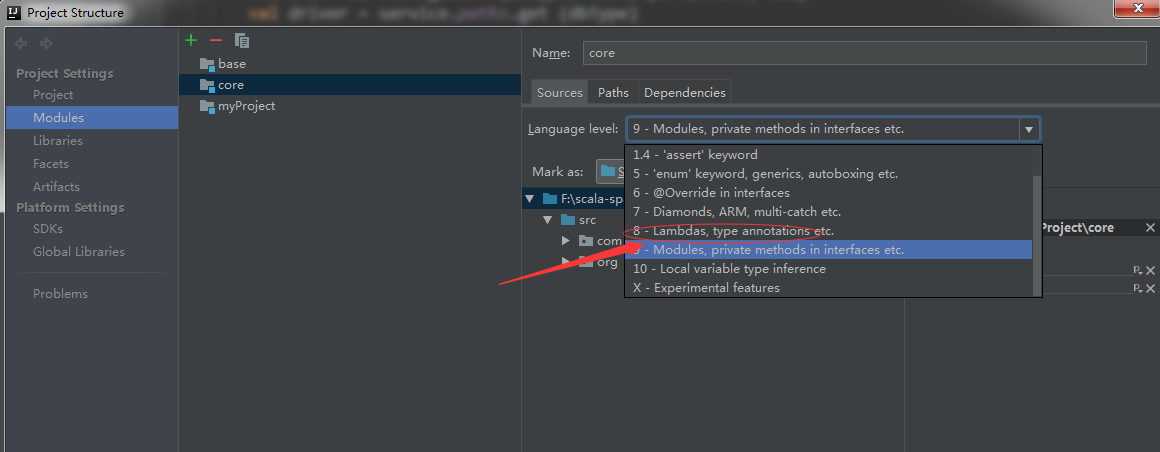
解决方式：将整个项目作为dependency加入到找不到package文件所在的module中（有时还需要将顺序调到最上面），编译一下，通过，然后在dependency中删除整个项目，重新编译。

解决方式2：rebuild整个项目

### Error:java: 无效的源发行版: 9

一种可能是JDK的版本不一致，如果这个排除，另一个可能是IDE的language level不一致。

解决：module => source => language level 改一下其他的试试



分析：

Ide有自动的语法解析设定，在编译是会对一些新出现的语法特性进行解读。

## IntelliJ的一些常用查询快捷键及其应用场景

1. Ctrl + N 搜索类名

Ctrl + H 查看类的继承关系

Ctrl + Alt + B 查看子类中方法的覆盖情况

Ctrl + r 文本替换

Ctrl + f 文本查找，查找的是字符串

Ctrl + F12 快速找到某类在项目中的引用情况，查找的是文件名

Alt + F7 查看类或者方法的引用情况，查找的是类和方法的引用

Ctrl + Alt + L 代码格式化

Ctrl + shift + f全局文本搜索

## IDE每行代码的字符数

当我们使用IDE写代码时，为了保证代码的可阅读性和优雅性，通常会借助IDE的代码风格设置功能，令IDE智能完成的代码部分或者格式化输入的代码，可以按照预期的格式输出。其中有一项设置就是限制一行代码的宽度，以IDEA举例，默认限制为120

1. 设定分隔线

在File->settings->Code Style->General中，修改“Right margin (columns)”的值即可改变代码行宽度的限制。

1. 设定自动换行

第一种，在上述的“Right margin (columns)”的下方，有“Wrap when typing reaches right margin”选项；

第二种，在File->settings->Code Style->Java中，选中“Wrapping and Braces”选项卡，在“Keep when reformatting”中有一个“Ensure rigth margin is not exceeded”，选中它。

## IntelliJ在某个目录下非常卡

描述：IntelliJ在某个目录下突然非常卡，对一些新建项目的目录不卡，而且之前也不卡。

原因分析：基本排除电脑问题，很可能是某个目录结构非常繁杂的对象被引入项目，这种情况很可能是由于版本管理的文件（svn和git的隐藏文件或者svn和git设置没有关闭）

## IntelliJ在第三方包的库中引入一个jar包但在intellij中找不到对应的类文件

描述：IntelliJ在第三方包的库中引入一个jar包但在intellij中找不到对应的类文件

解决：intellij的类文件没有刷出来。将第三方的library再引入一次

## IntelliJ将文件添加进来但是编译显示仍然不在

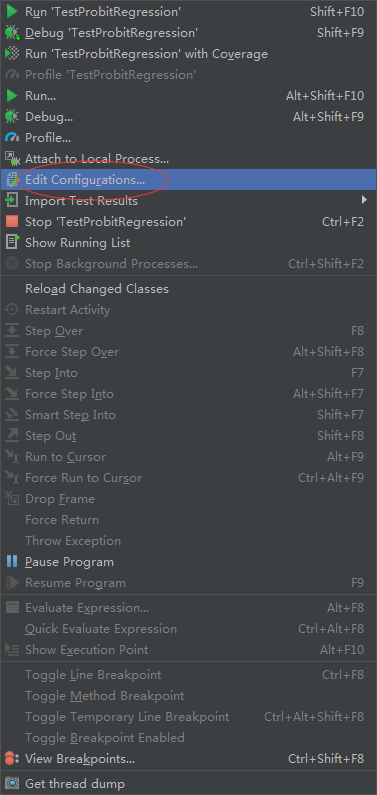
和4.1的原理一致，这里我碰到了隐士转换引入之后发现没有该方法，编译不过，但脚本中名明星有该方法。

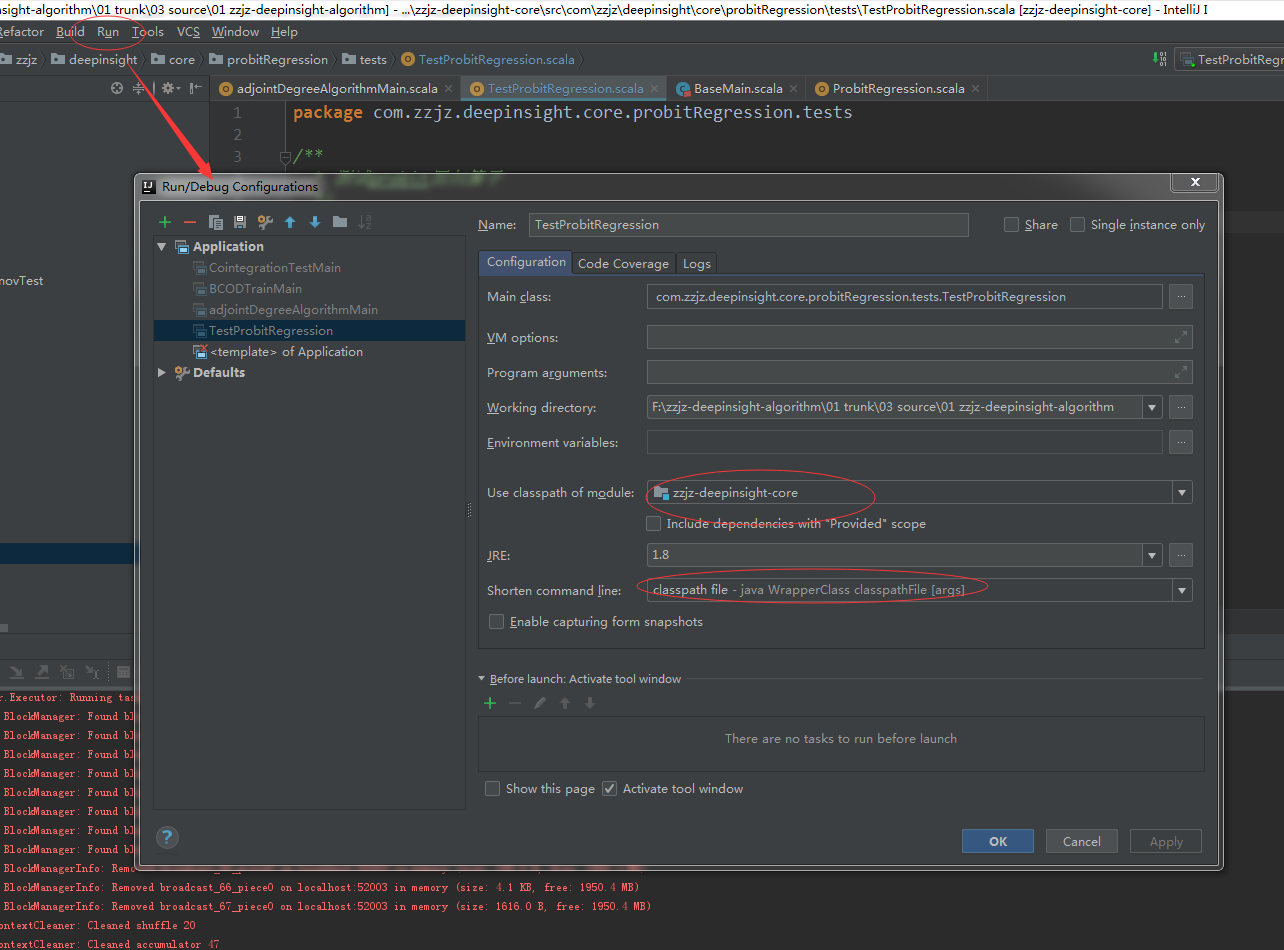
解决：重新编译该脚本，在编译隐士转换出错的脚本。

另外：还有一种可能是包的名字是不对的package A.B.C和目录不一致。

## Command is too long

解决方式：在run => edit configuration 将shotten command line 改为基于classpath





## 一个较low的解决显示不了目录下文件的方法

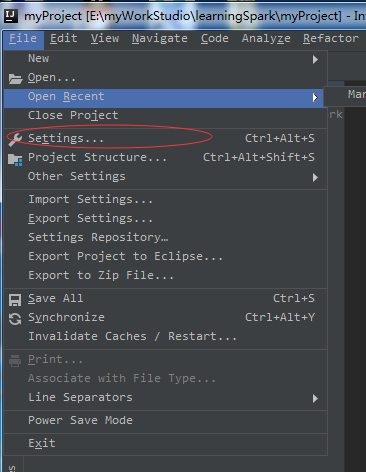
描述：明明有文件，但是点不开文件夹，比如曾经有段时间test就找不到。

分析：肯定做了某些设置，导致的，但不知道什么搞回去

解决：将.idea文件删了，重新导入项目。。。。。。

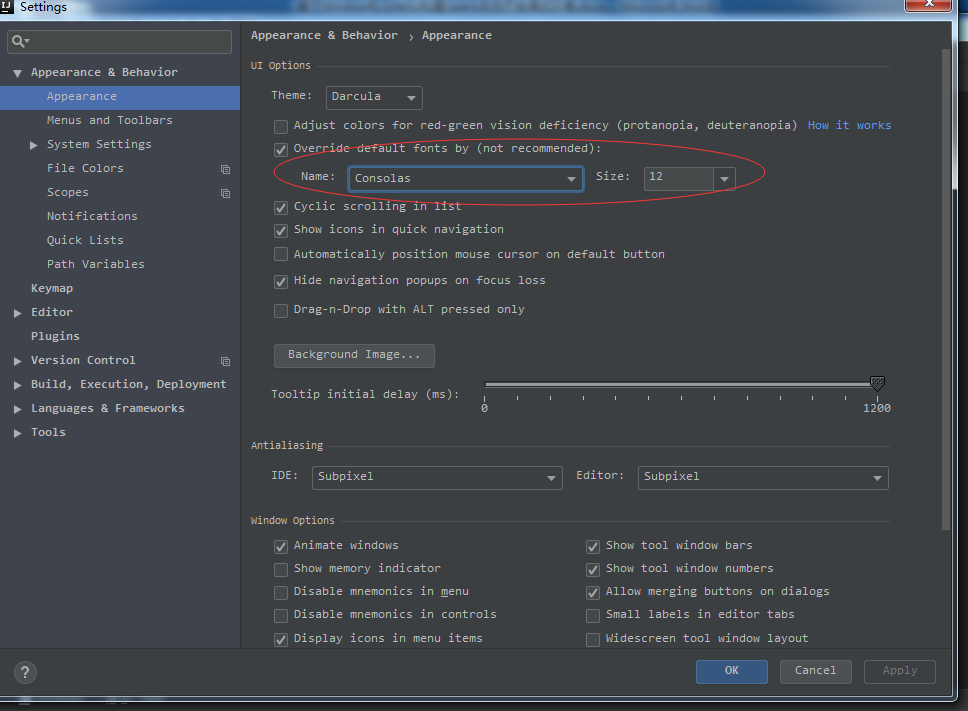
## 解决intellij菜单栏显示不能使中文的方法

下图中原文应该是一些某国文字：



分析原因：是appearance中的字体发生了变更，变成了某个国家的文字（这个问题常出现于windows更新后）

解决：设置setting中appearance的override def font，随便设置一中字体，比如第一条consolas。



此时文字变为英文，再按自己的喜好设置。

## main package not found \*\*\*\*

运行的主类找不到。

原因：还是在编译中的问题，本例子中之前可以编译通过，但后面抛出这个错误，后来解决后发现本来这个就编译不过。这个是同学出现的，他描述他的做法是：编译一个脚本，成功，改下改吗，失败，删除编译文件，报错

解决：试了rebuid失败，试了删除out目录下的文件失败。最终，改了下主类名称，成功定位到代码问题，改了下代码问题，编译成功。

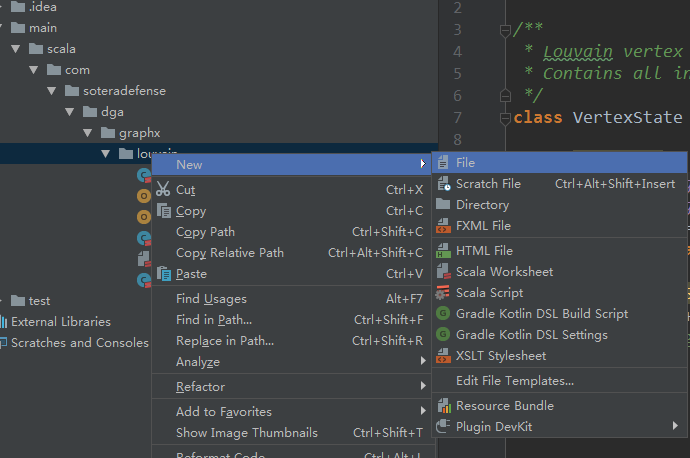
## SparkSQL在本地创建DataFrame碰到空指针或者，ralativeURI的情况

是conf配置的问题：注释的是出问题的，为注释的是可以执行的

// System.setProperty("hadoop.home.dir", "file:/F:/zzjz-deepinsight-algorithm2.0/01spark2.0/03src/spark-warehouse/")  
  
// lazy val cfg: SparkConf =new SparkConf().setAppName("zzjz-deepinsight-algorithm").setMaster("local[2]")  
// lazy val sparks: SparkSession =SparkSession.builder().config(cfg).enableHiveSupport().getOrCreate()  
// lazy val sc: SparkContext =sparks.sparkContext  
// lazy val sqlc: SQLContext =sparks.sqlContext  
// lazy val hqlc: SparkSession =sparks  
// lazy val ssc : StreamingContext = new StreamingContext(cfg, Seconds(1))  
  
 */\*\* the config set for myself --xuhao not commit \*/* **lazy val** *cfg*: SparkConf =**new** SparkConf()  
 .setAppName("DataShoe\_Learning")  
 .setMaster("local[4]")  
 .set("spark.sql.warehouse.dir", "file:///")  
  
 **lazy val** *sparks*: SparkSession =SparkSession.*builder*().config(*cfg*).getOrCreate()  
  
 **lazy val** *sc*: SparkContext =*sparks*.sparkContext  
 **lazy val** *sqlc*: SQLContext =*sparks*.*sqlContext* **lazy val** *hqlc* = *sqlc* **lazy val** *ssc* : StreamingContext = **new** StreamingContext(*cfg*, *Seconds*(1))

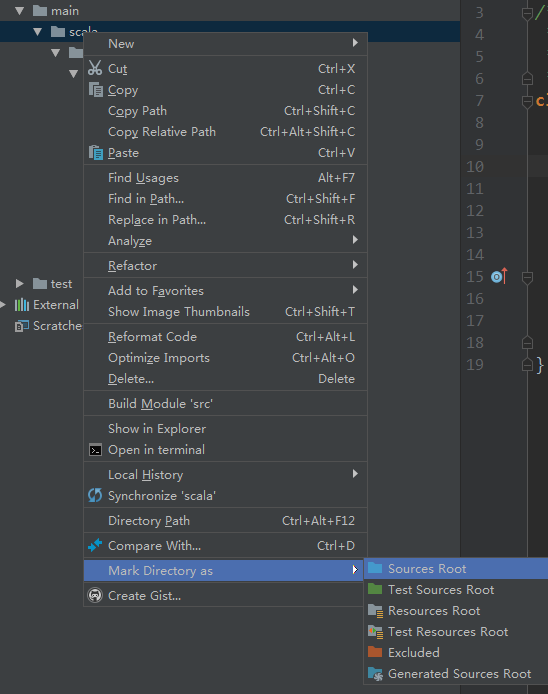
## Intellij中无法新建scala对象的问题

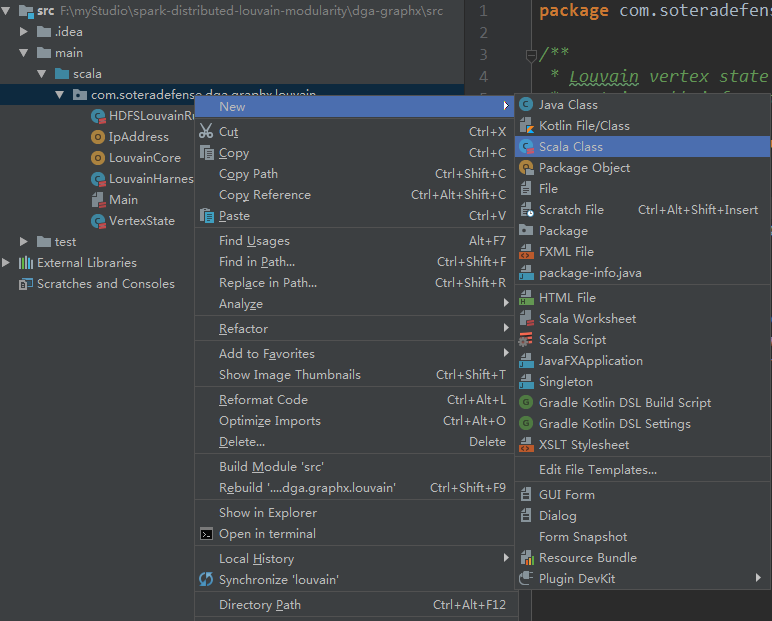
描述：



分析：是因为不知道是source文件夹

解决：





开森^^

## Intellij快捷键查看和设置

Set => keyMap可以查看快捷键