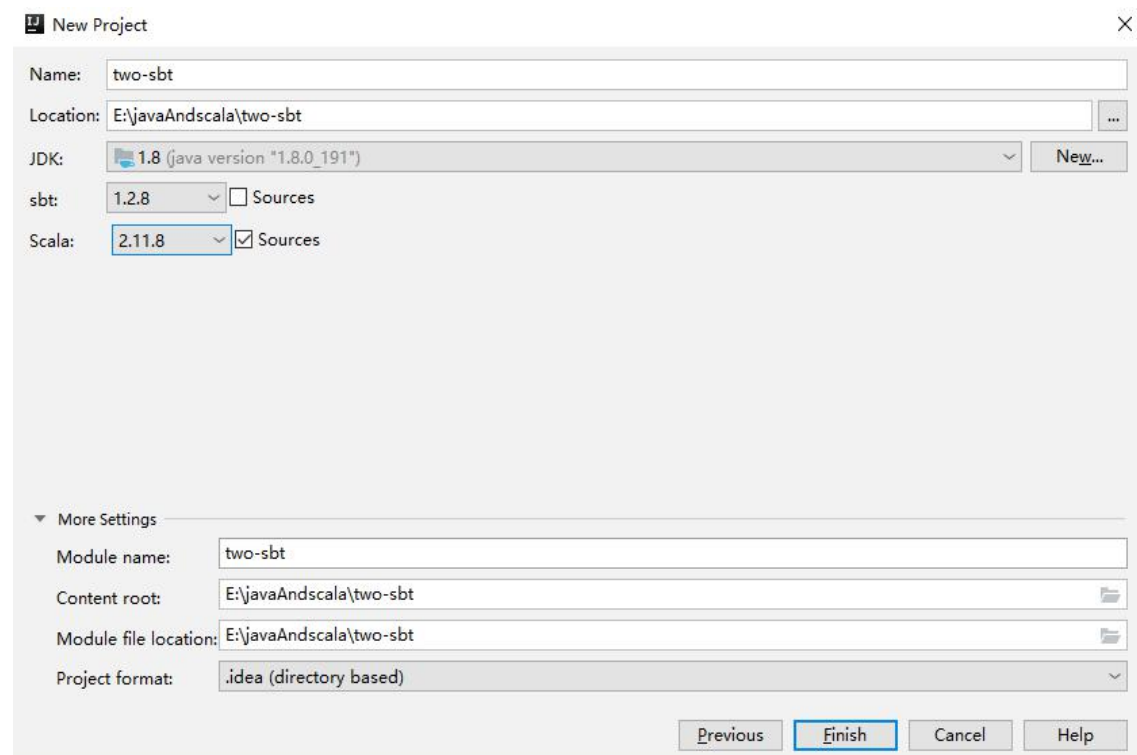
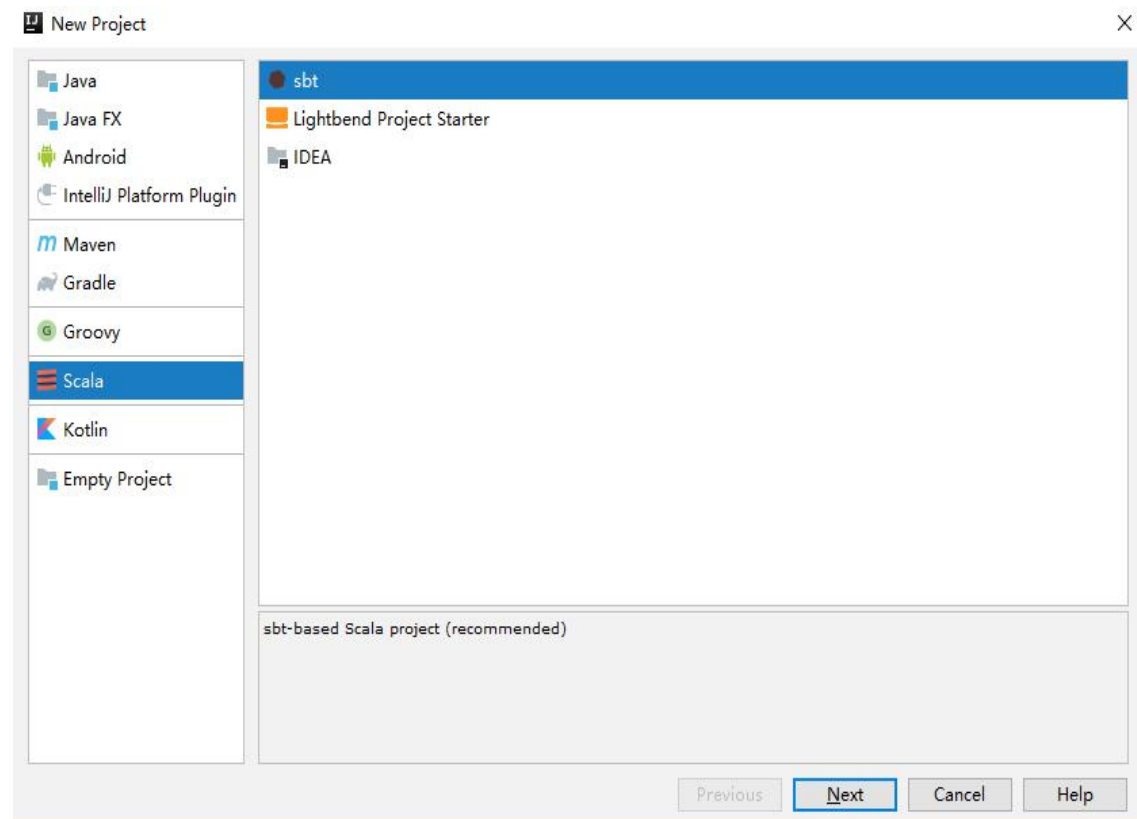


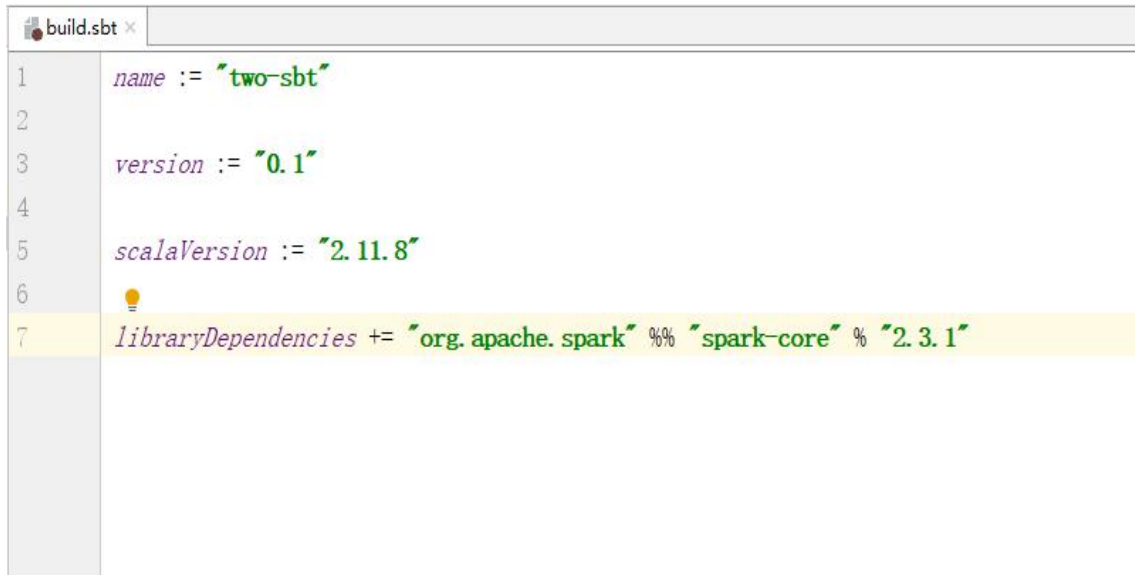
windows 下编写 spark 程序并运行

1. 打开 IDEA
2. 新建一个 scala 工程



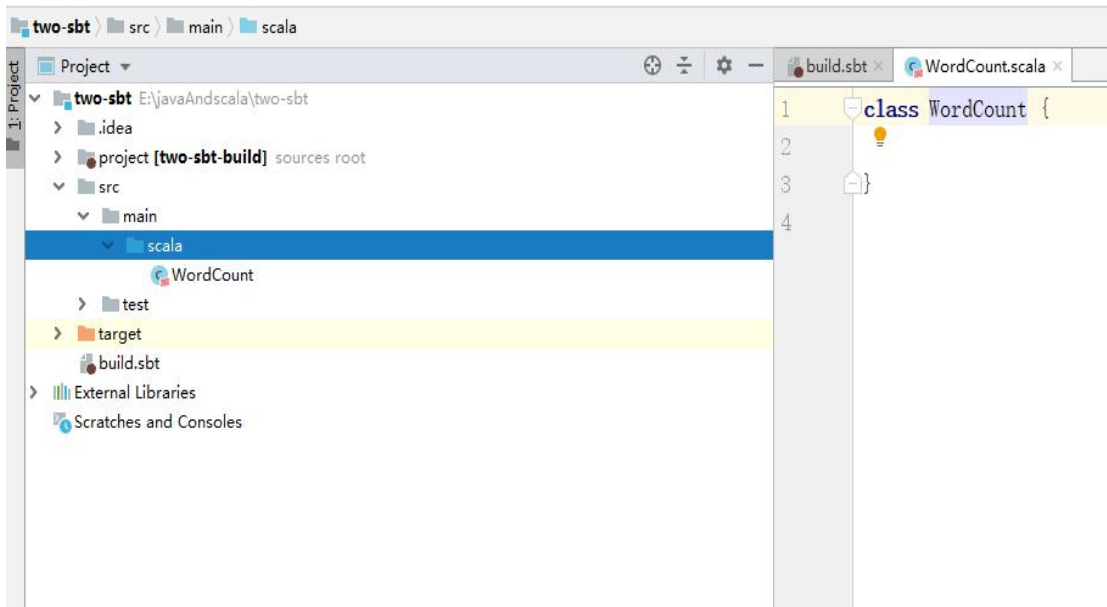
注意：1. scala sources 哪里选择对应的版本，我选择的时 2.11.8

3. 编辑 build.sbt



```
1  name := "two-sbt"
2
3  version := "0.1"
4
5  scalaVersion := "2.11.8"
6
7  libraryDependencies += "org.apache.spark" %% "spark-core" % "2.3.1"
```

4. 在 src/main/scala 目录下，新建一个 WordCount.scala 文件



```
1  class WordCount {
2
3  }
4
```

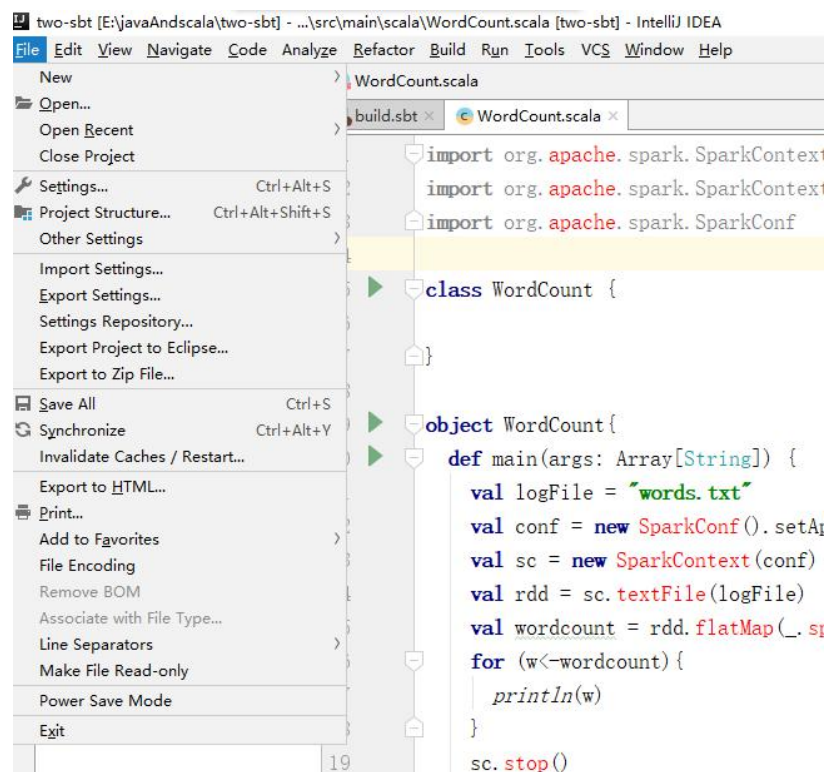
5. 输入以下代码

```
build.sbt x WordCount.scala x
1 import org.apache.spark.SparkContext
2 import org.apache.spark.SparkContext._
3 import org.apache.spark.SparkConf
4
5 class WordCount {
6
7 }
8
9 object WordCount {
10 def main(args: Array[String]) {
11     val logFile = "words.txt"
12     val conf = new SparkConf().setAppName("Simple Application").setMaster("local")
13     val sc = new SparkContext(conf)
14     val rdd = sc.textFile(logFile)
15     val wordcount = rdd.flatMap(_.split(" ")).map((_, 1)).reduceByKey(_ + _).map(x => (x._2, x._1)).sortByKey(false).map(x => (x._2, x._1))
16     for (w<-wordcount) {
17         println(w)
18     }
19     sc.stop()
20 }
21 }
```

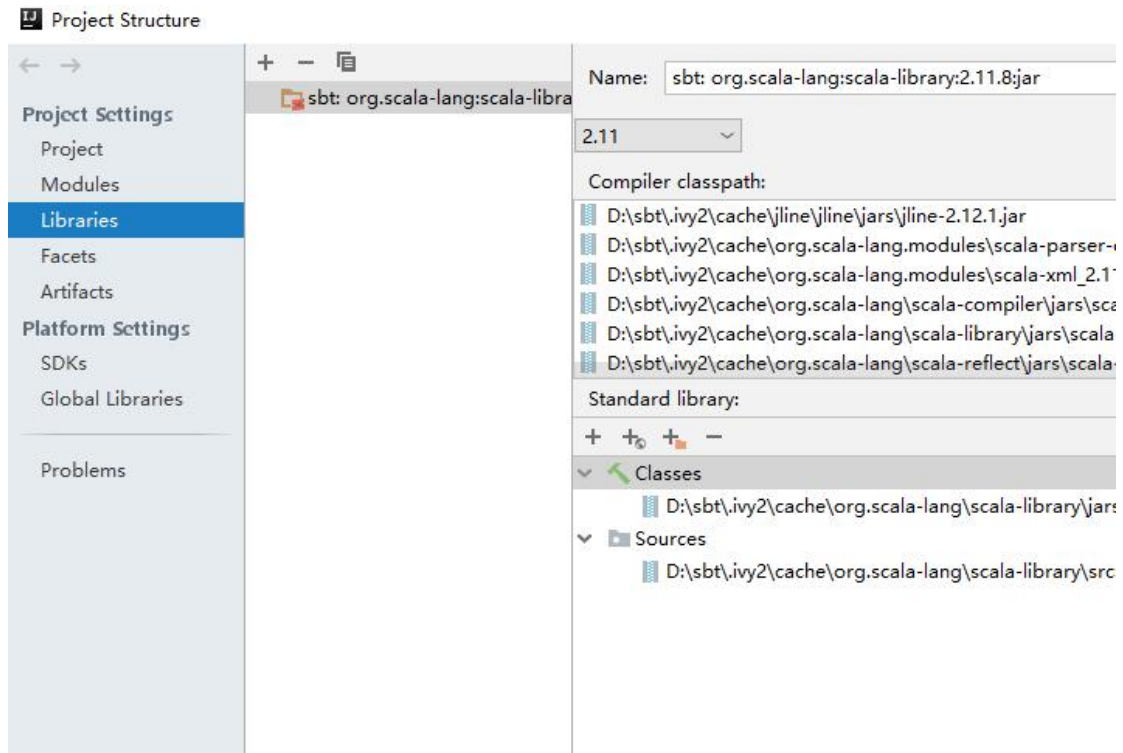
注意：1.可以看到 SparkConf, SparkContext 等呈现的是红色，那是因为我们没用到相应的 jar 包导入

6. 导入 jar 包

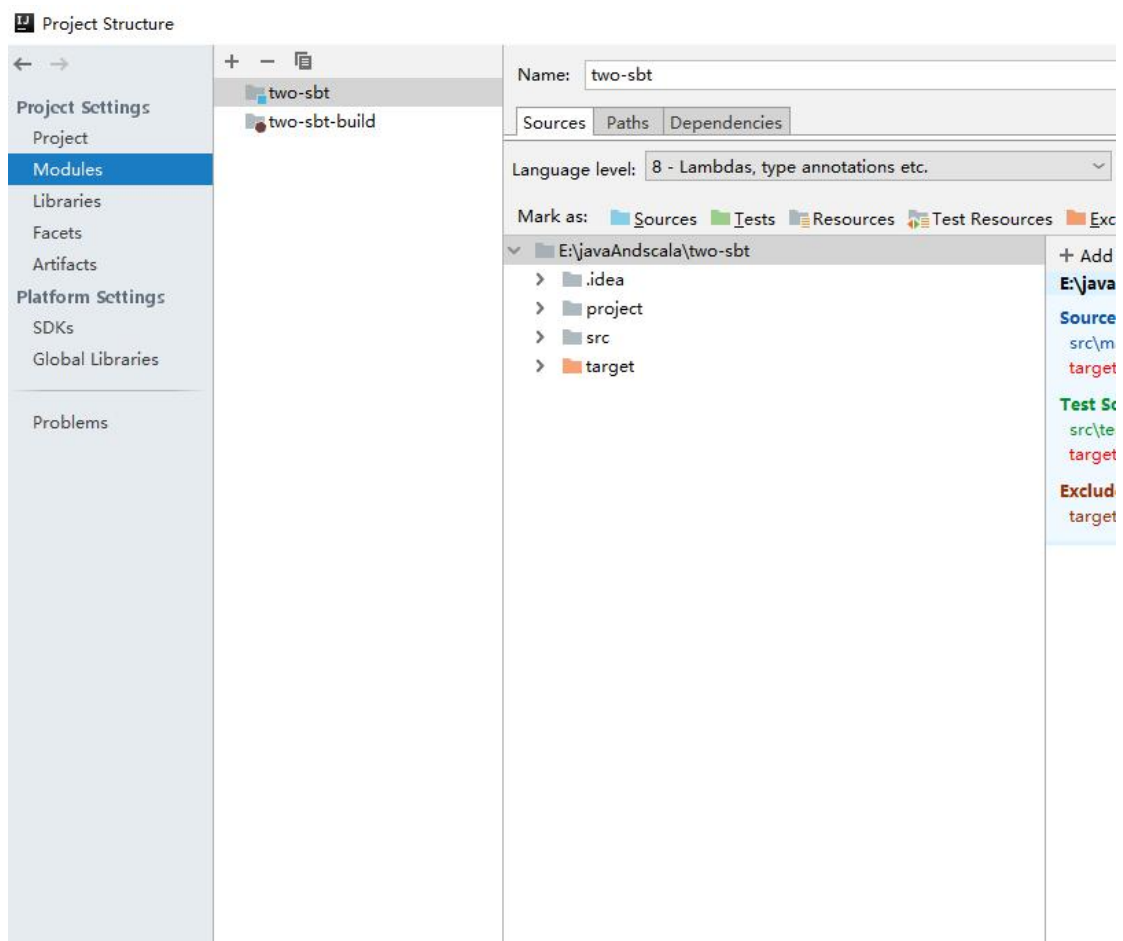
6.1 点击左上角的 file



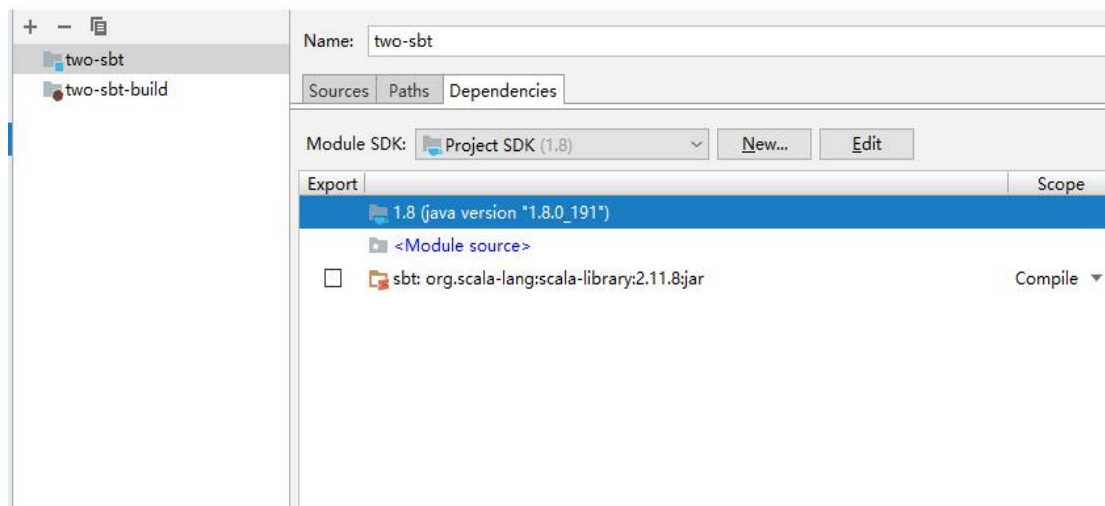
6.2 点击 project structure



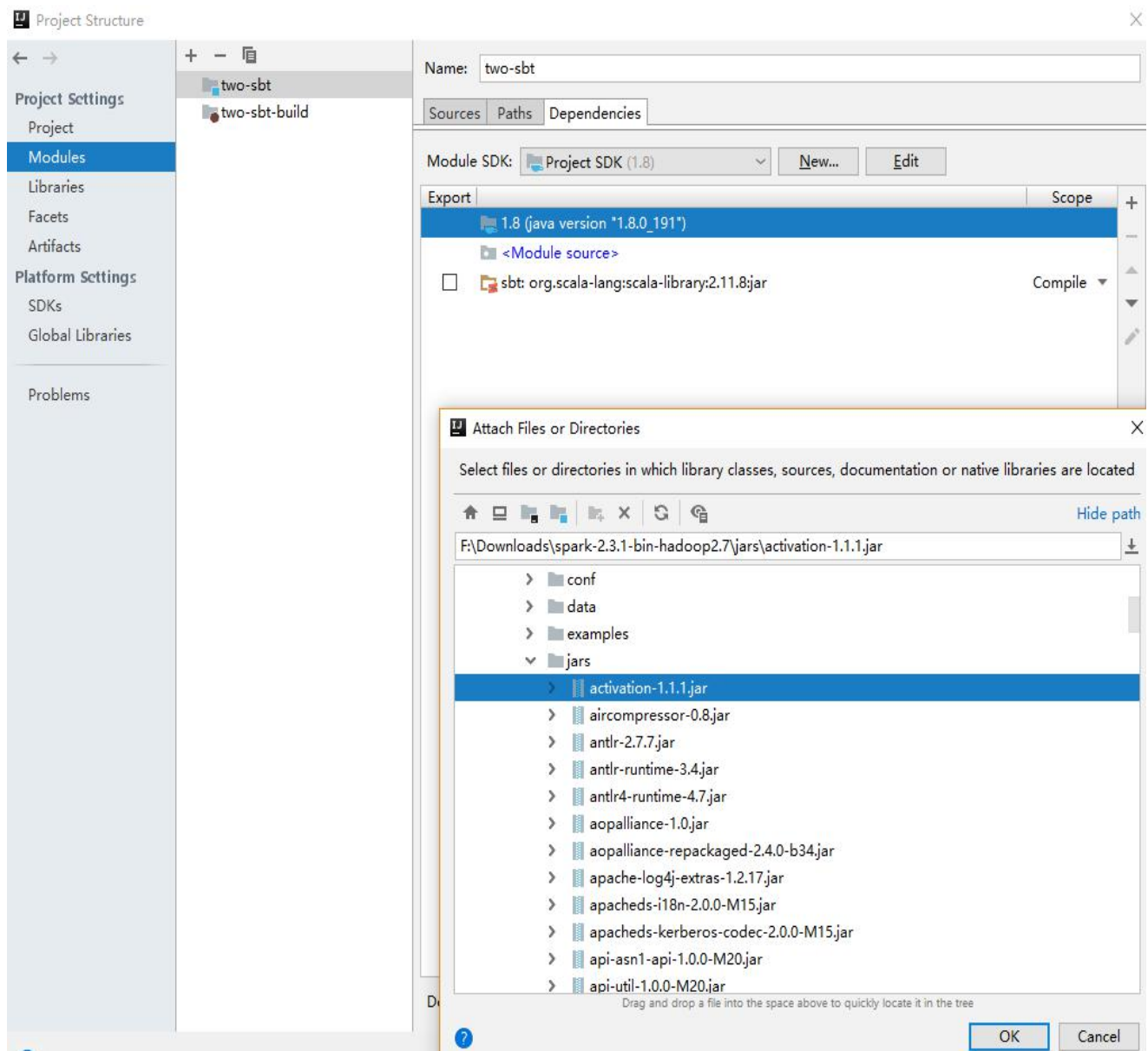
6.3 点击 Modules



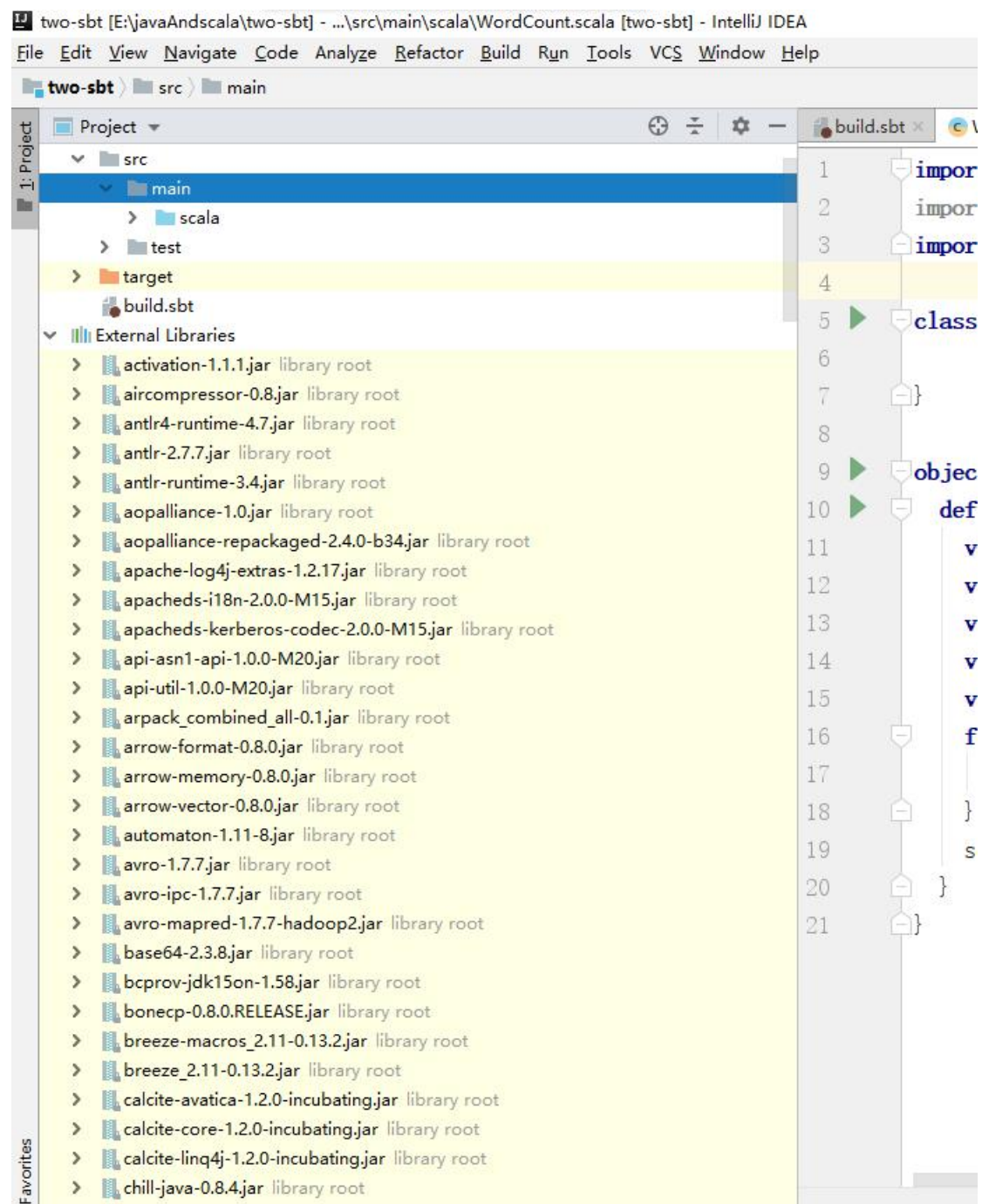
6.4 点击右侧的 dependencies



6.5 点击右侧的“+”，导入相应的 jar 包



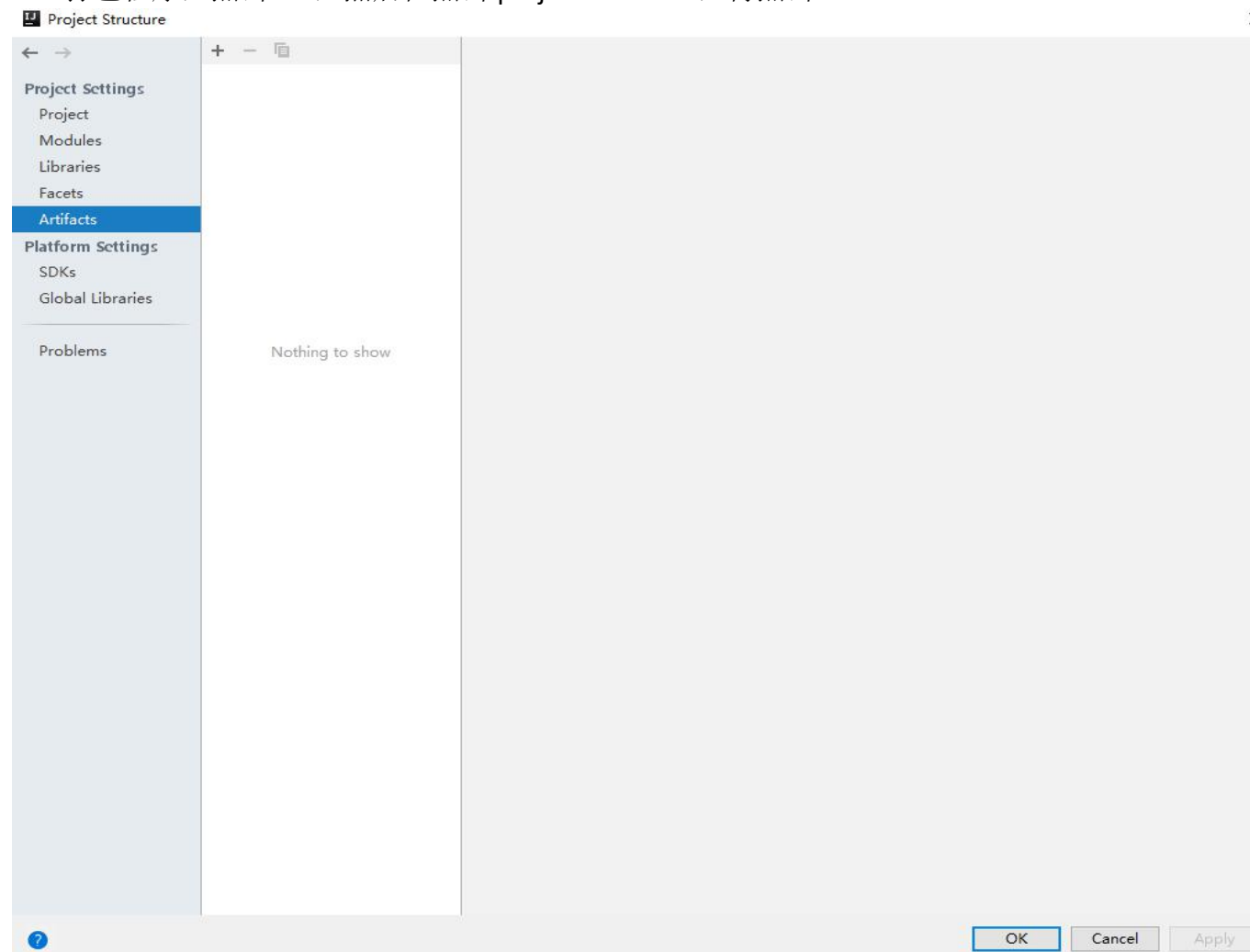
6.5 将 jars 中的包全部导入



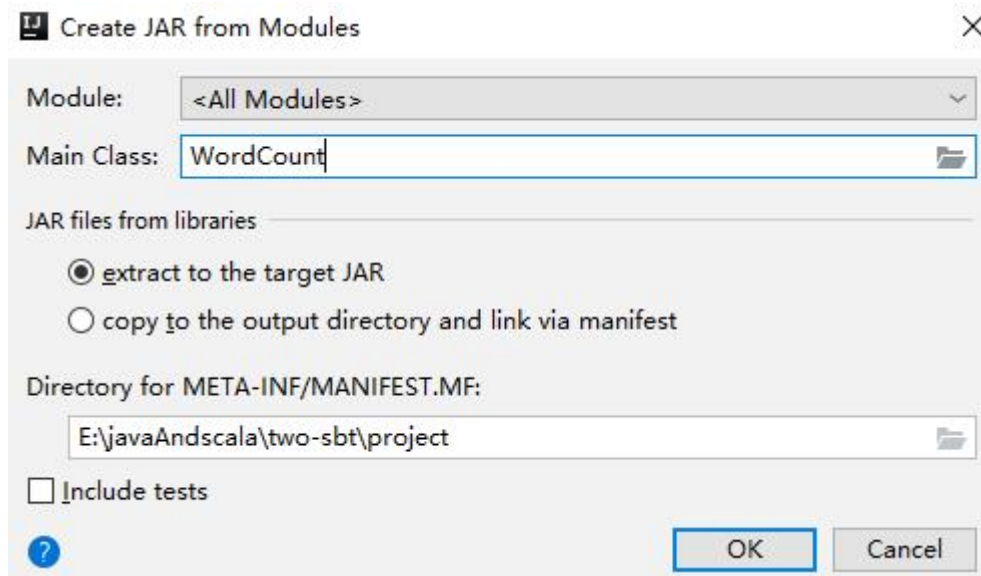
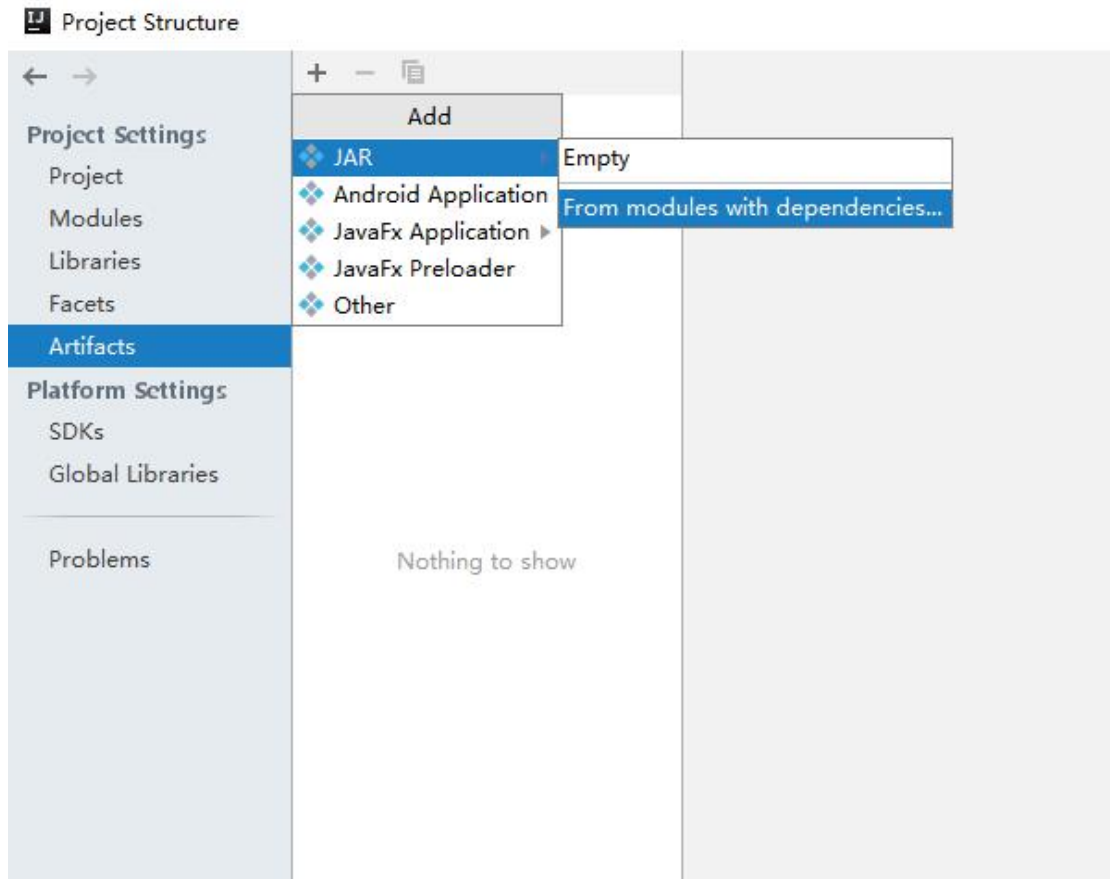
7. 点击 run，运行成功。


```
2019-01-14 16:53:42 INFO TaskSetManager:54 - Starting task 0.0 in stage 2.0 (TID 2, local
2019-01-14 16:53:42 INFO Executor:54 - Running task 0.0 in stage 2.0 (TID 2)
2019-01-14 16:53:42 INFO ShuffleBlockFetcherIterator:54 - Getting 1 non-empty blocks out
2019-01-14 16:53:42 INFO ShuffleBlockFetcherIterator:54 - Started 0 remote fetches in 0 m
(hello, 6)
(scala, 1)
(tom, 1)
(jerry, 1)
(tomcat, 1)
(kitty, 1)
(world, 1)
2019-01-14 16:53:42 INFO Executor:54 - Finished task 0.0 in stage 2.0 (TID 2). 1052 bytes
2019-01-14 16:53:42 INFO TaskSetManager:54 - Finished task 0.0 in stage 2.0 (TID 2) in 46
2019-01-14 16:53:42 INFO TaskSchedulerImpl:54 - Removed TaskSet 2.0, whose tasks have all
2019-01-14 16:53:42 INFO DAGScheduler:54 - ResultStage 2 (foreach at WordCount.scala:16)
2019-01-14 16:53:42 INFO DAGScheduler:54 - Job 0 finished: foreach at WordCount.scala:16,
```

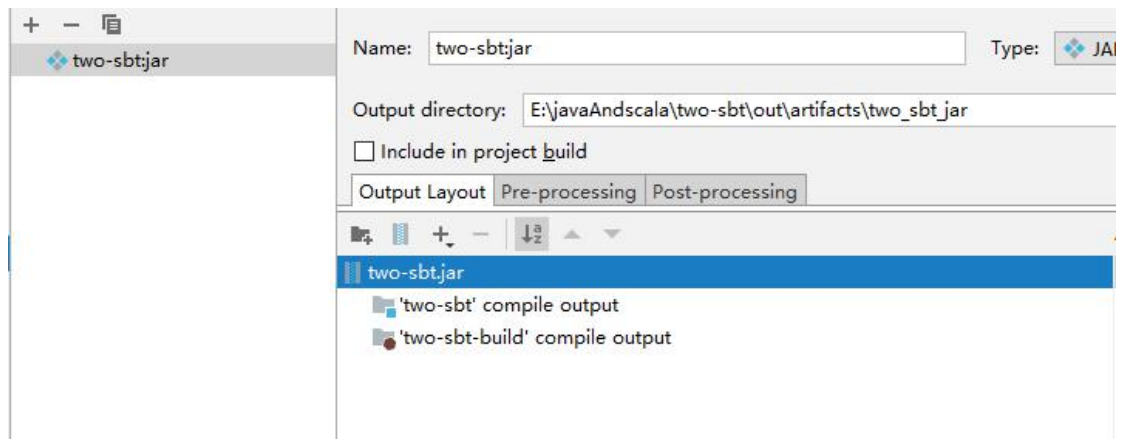
8. 打包程序，点击 file，然后在点击 project structure，再点击 artifacts



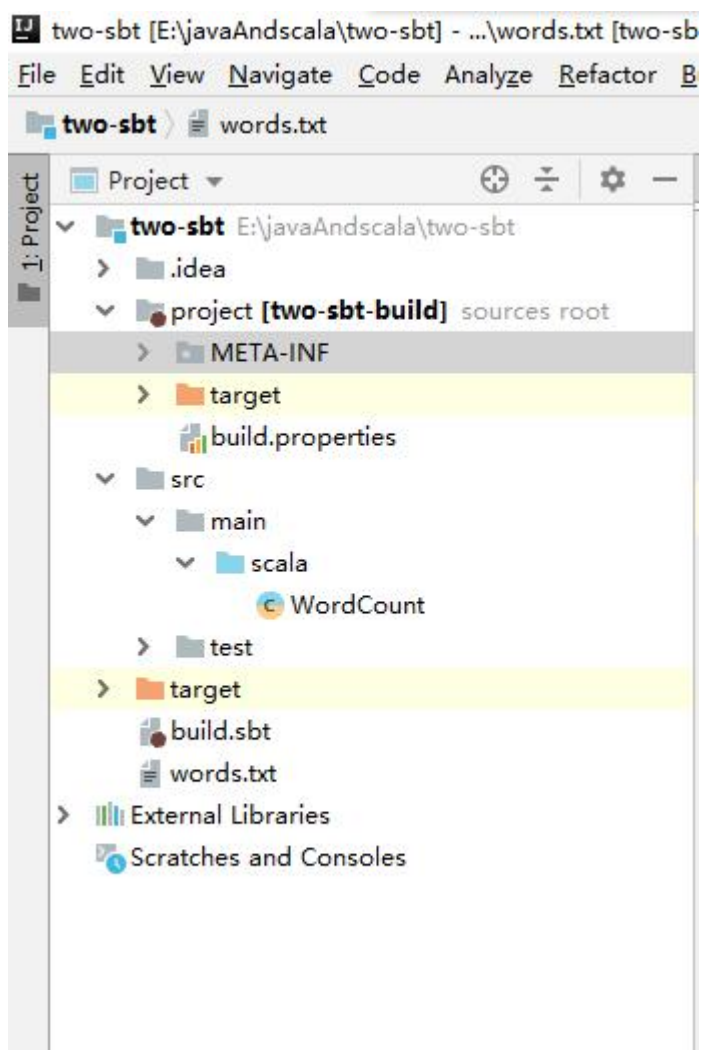
9. 点击“+”，选择如下：



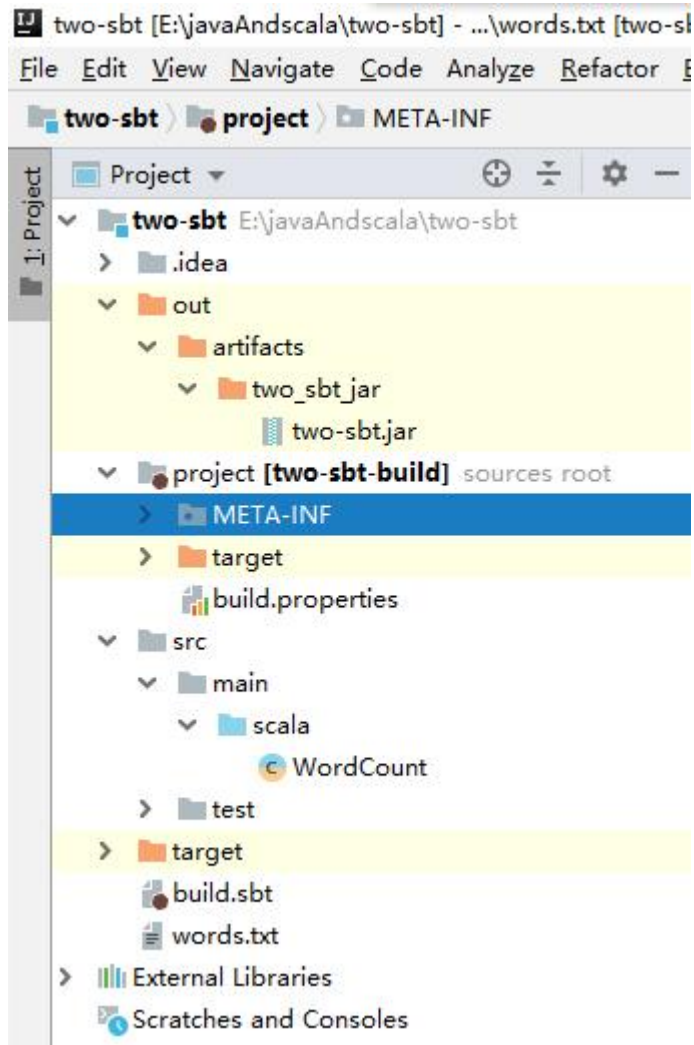
10. 删除多余的 jar 包，只需上下如下内容



11. 右侧会出现 META-INF，然后用 build artifacts 进行导出



12. 结果如下



13. 将 two-sbt.jar 和 words.txt 放到 centos7 的 root 目录下

14. 输入以下命令运行，并运行成功！

```
[root@HadoopMaster bin]# ./spark-submit --class "WordCount" /root/two-sbt.jar
```

```
19/01/14 17:10:47 INFO MemoryStore: Block broadcast_3 stored as
19/01/14 17:10:47 INFO MemoryStore: Block broadcast_3_piece0 st
19/01/14 17:10:47 INFO BlockManagerInfo: Added broadcast_3_piec
19/01/14 17:10:47 INFO SparkContext: Created broadcast 3 from b
19/01/14 17:10:47 INFO DAGScheduler: Submitting 1 missing tasks
19/01/14 17:10:47 INFO TaskSchedulerImpl: Adding task set 2.0 w
19/01/14 17:10:47 INFO TaskSetManager: Starting task 0.0 in sta
19/01/14 17:10:47 INFO Executor: Running task 0.0 in stage 2.0
19/01/14 17:10:47 INFO ShuffleBlockFetcherIterator: Getting 1 n
19/01/14 17:10:47 INFO ShuffleBlockFetcherIterator: Started 0 r
(hello,6)
(scala,1)
(tom,1)
(jerry,1)
(tomcat,1)
(kitty,1)
(world,1)
19/01/14 17:10:47 INFO Executor: Finished task 0.0 in stage 2.0
19/01/14 17:10:47 INFO TaskSetManager: Finished task 0.0 in sta
19/01/14 17:10:47 INFO TaskSchedulerImpl: Removed TaskSet 2.0, 1
19/01/14 17:10:47 INFO DAGScheduler: ResultStage 2 (foreach at 1
19/01/14 17:10:47 INFO DAGScheduler: Job 0 finished: foreach at 1
19/01/14 17:10:47 INFO SparkUI: Stopped Spark web UI at http://1
```