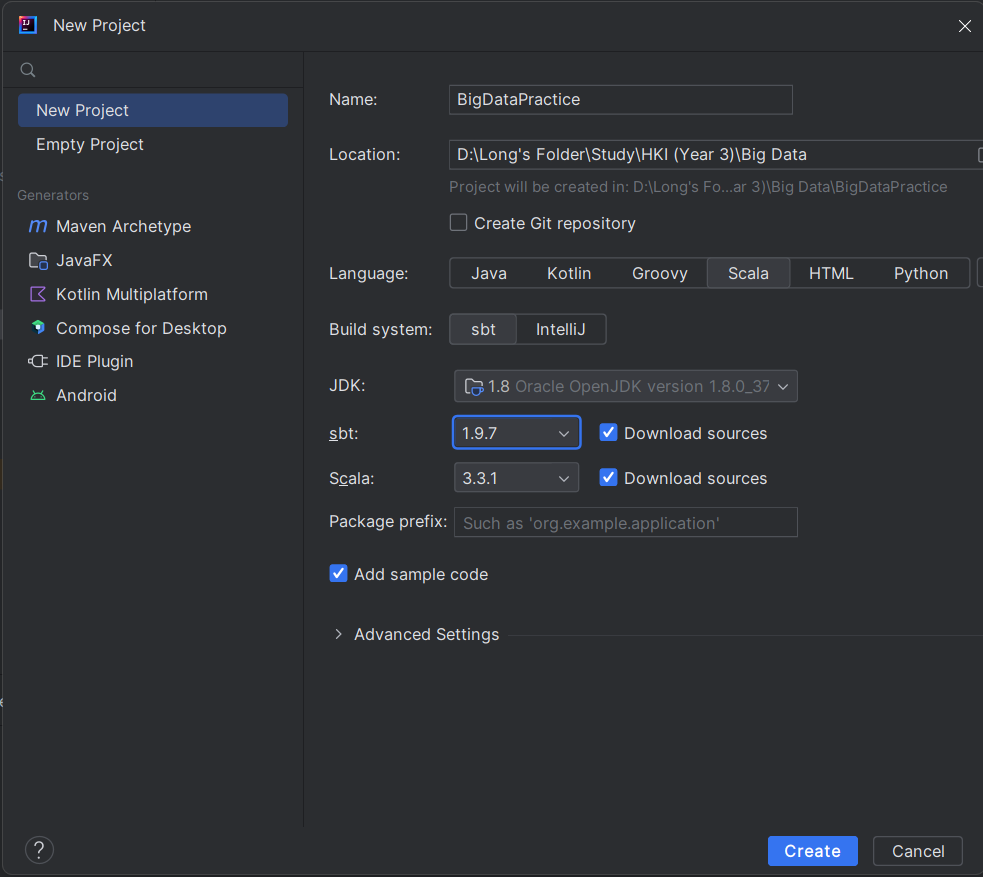
**Lab 3: WordCount**

**Họ và Tên**: Trần Nguyễn Gia Long

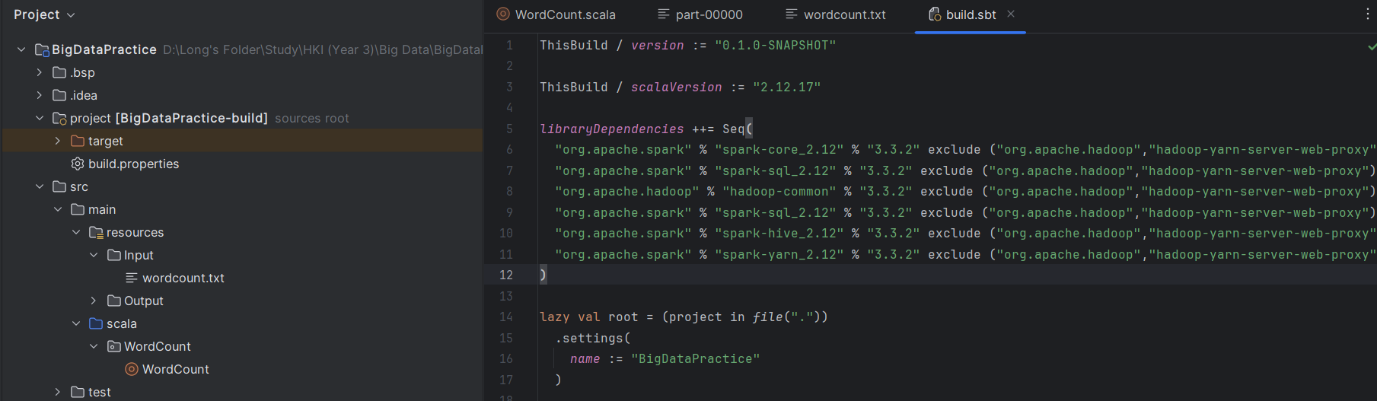
**MSSV**: 1050080059

# 1. Cài đặt project

- Đầu tiên, tạo một project với ngôn ngữ Scala với sbt là 1.9.7.



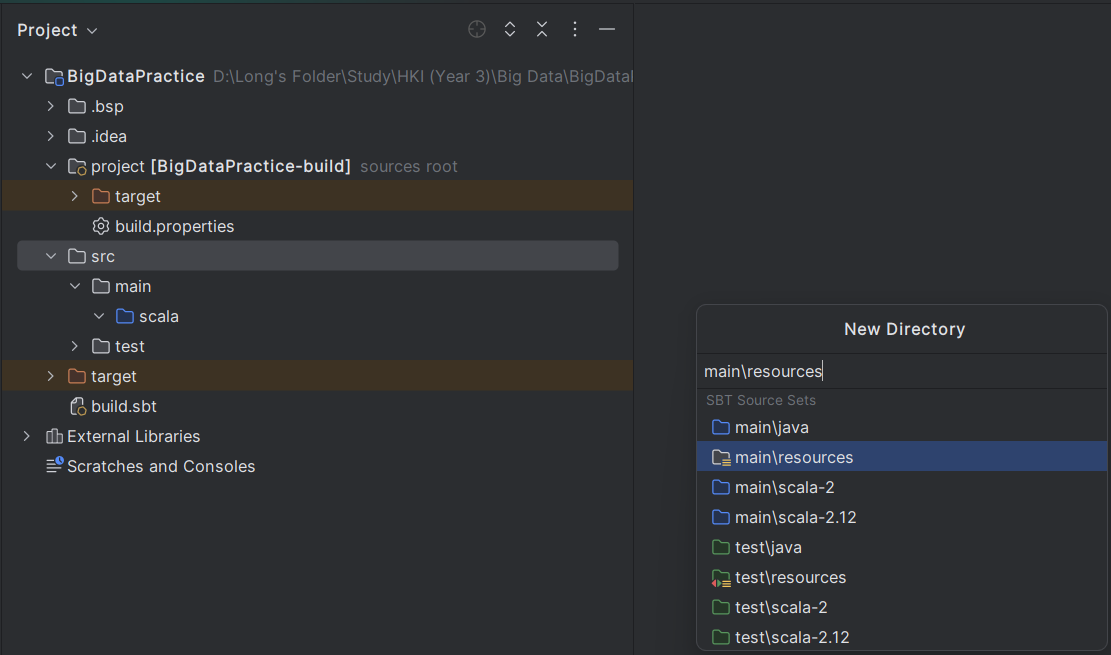
- Tiếp theo ta sẽ tạo thư viện trong file build.sbt.



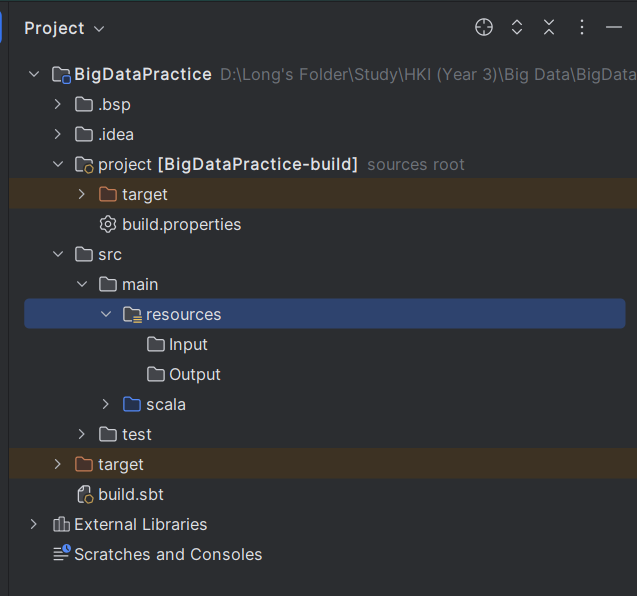
*libraryDependencies* ++= Seq(  
 "org.apache.spark" % "spark-core\_2.12" % "3.3.2" exclude ("org.apache.hadoop","hadoop-yarn-server-web-proxy"),  
 "org.apache.spark" % "spark-sql\_2.12" % "3.3.2" exclude ("org.apache.hadoop","hadoop-yarn-server-web-proxy"),  
 "org.apache.hadoop" % "hadoop-common" % "3.3.2" exclude ("org.apache.hadoop","hadoop-yarn-server-web-proxy"),  
 "org.apache.spark" % "spark-sql\_2.12" % "3.3.2" exclude ("org.apache.hadoop","hadoop-yarn-server-web-proxy"),  
 "org.apache.spark" % "spark-hive\_2.12" % "3.3.2" exclude ("org.apache.hadoop","hadoop-yarn-server-web-proxy"),  
 "org.apache.spark" % "spark-yarn\_2.12" % "3.3.2" exclude ("org.apache.hadoop","hadoop-yarn-server-web-proxy")  
)

# 2. Cài đặt file đầu vào và đầu ra

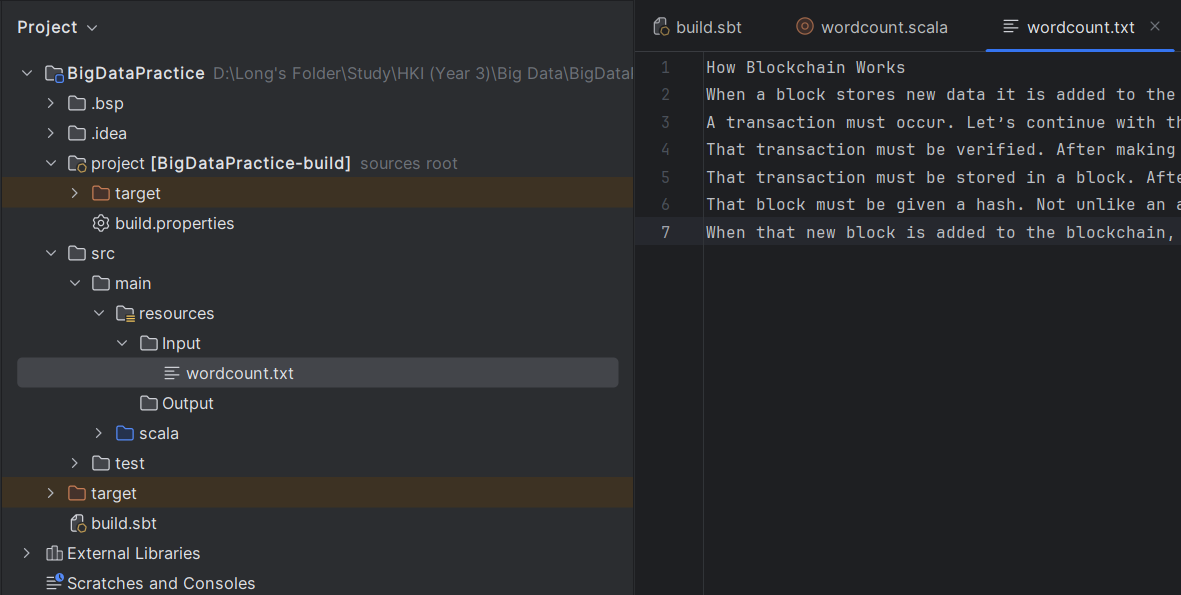
- Tạo một thư mục **resources** trong thư mục **src** để chứa file đầu vào và nơi chứa file xuất.



- Sau đó tạo mục Input và Output

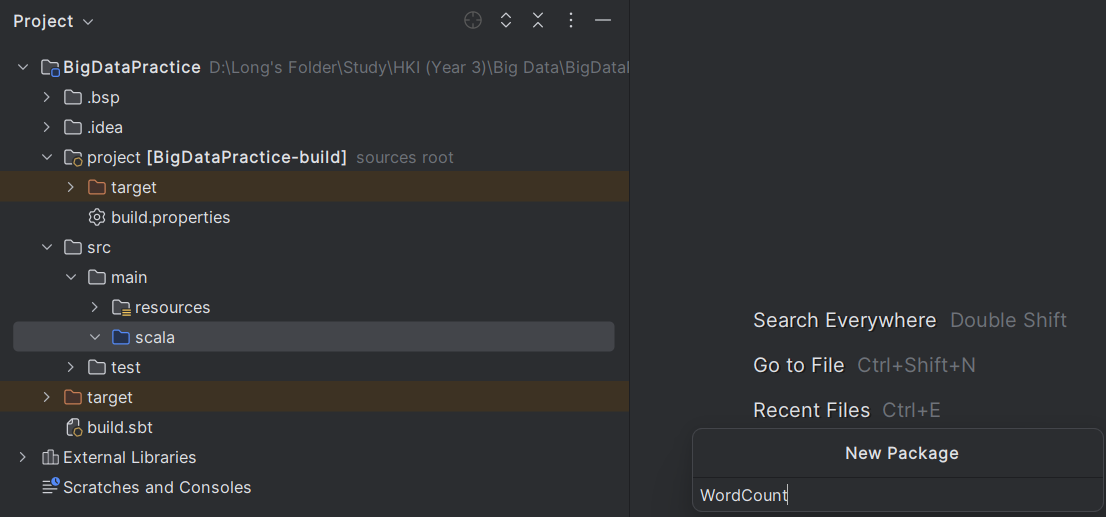


- Thêm dữ liệu để đếm từ trong wordcount thầy đã đưa.

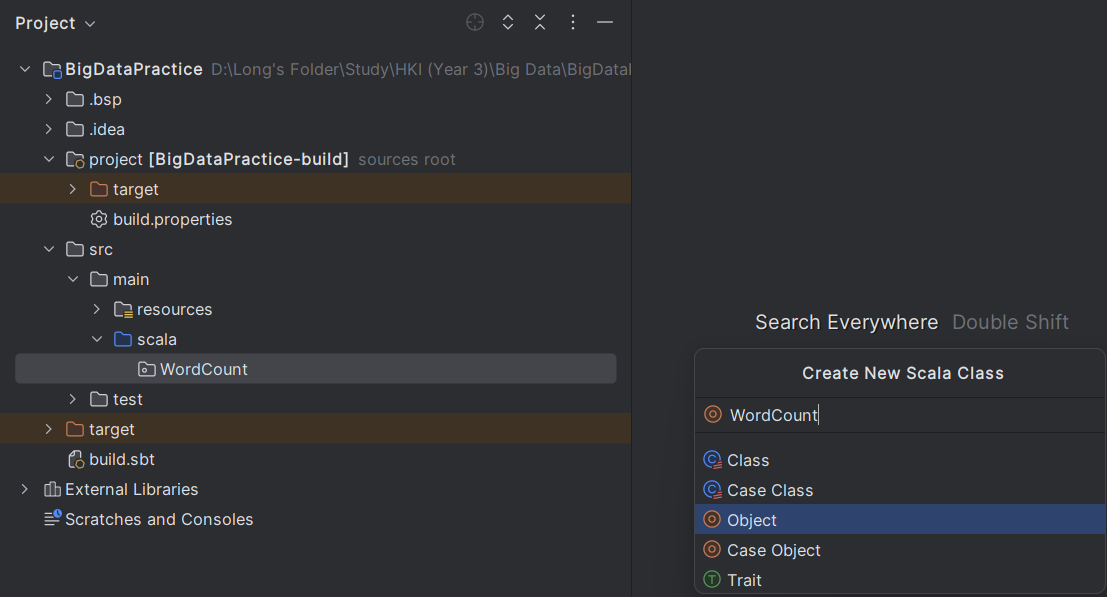


# 3. Cài đặt lệnh

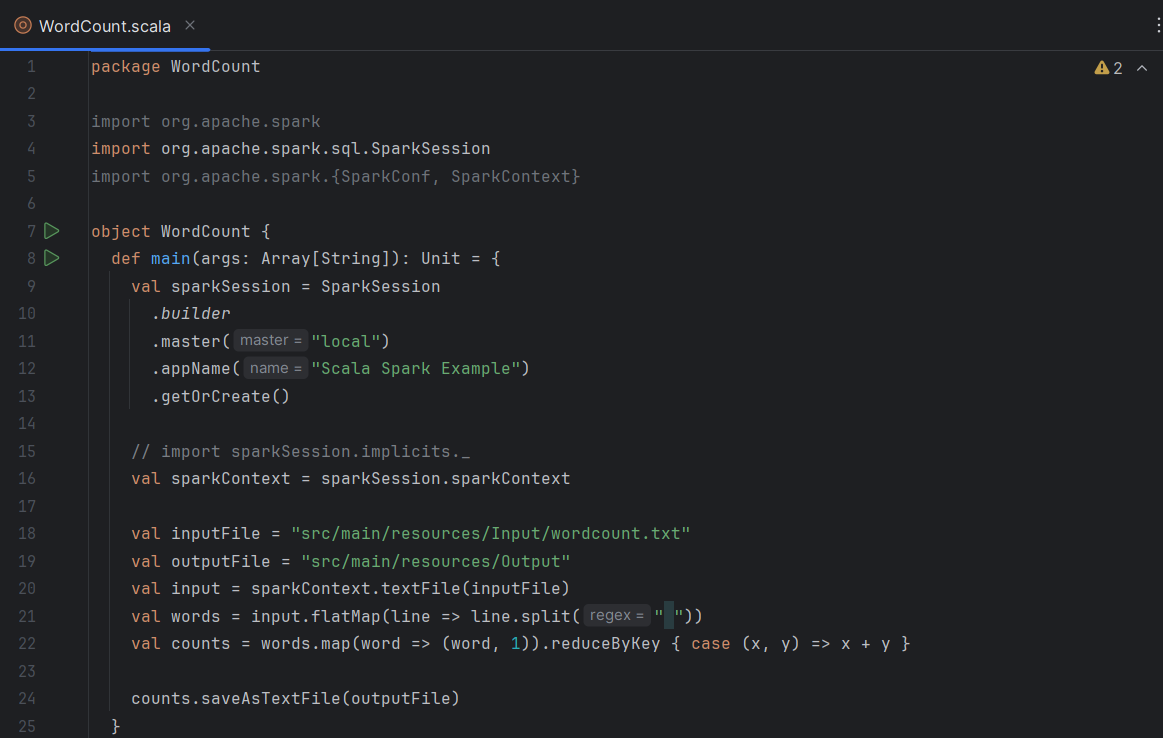
- Trong đường dẫn *src/main/scala* tạo một **Gói** tên là **WordCount**.



Sau đó là **object** từ **scala class**. Ở đây ta đặt tên là WordCount.



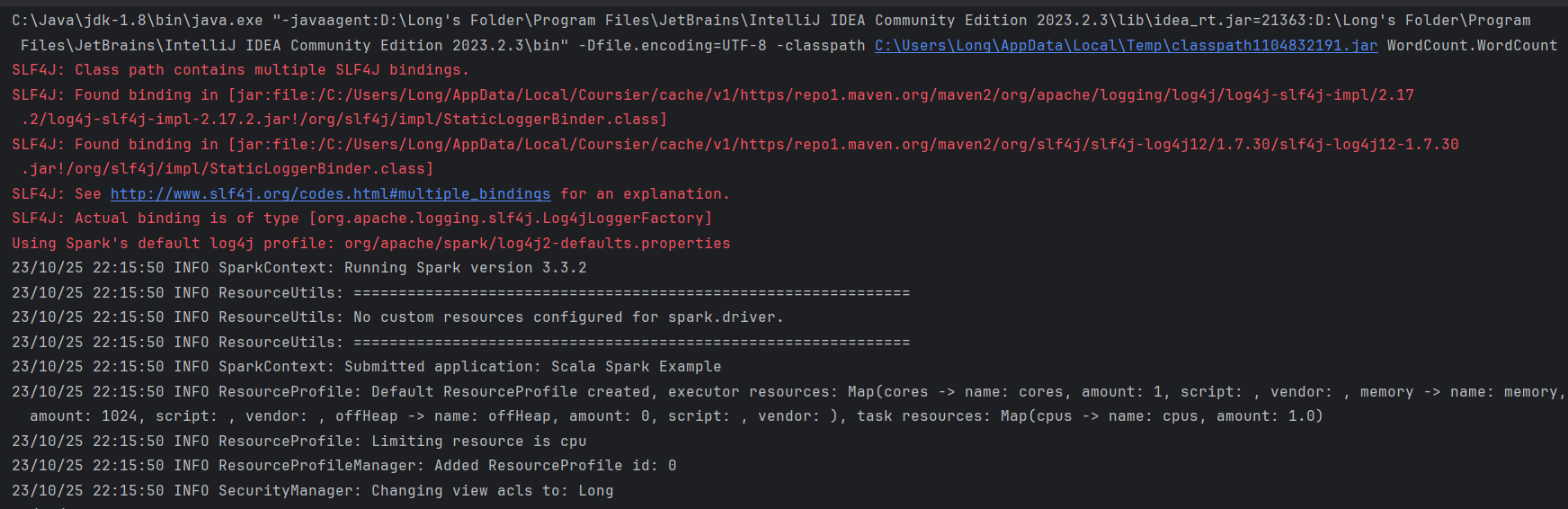
- Tiếp theo là ta nhập code đếm từ bằng ngôn ngữ scala.



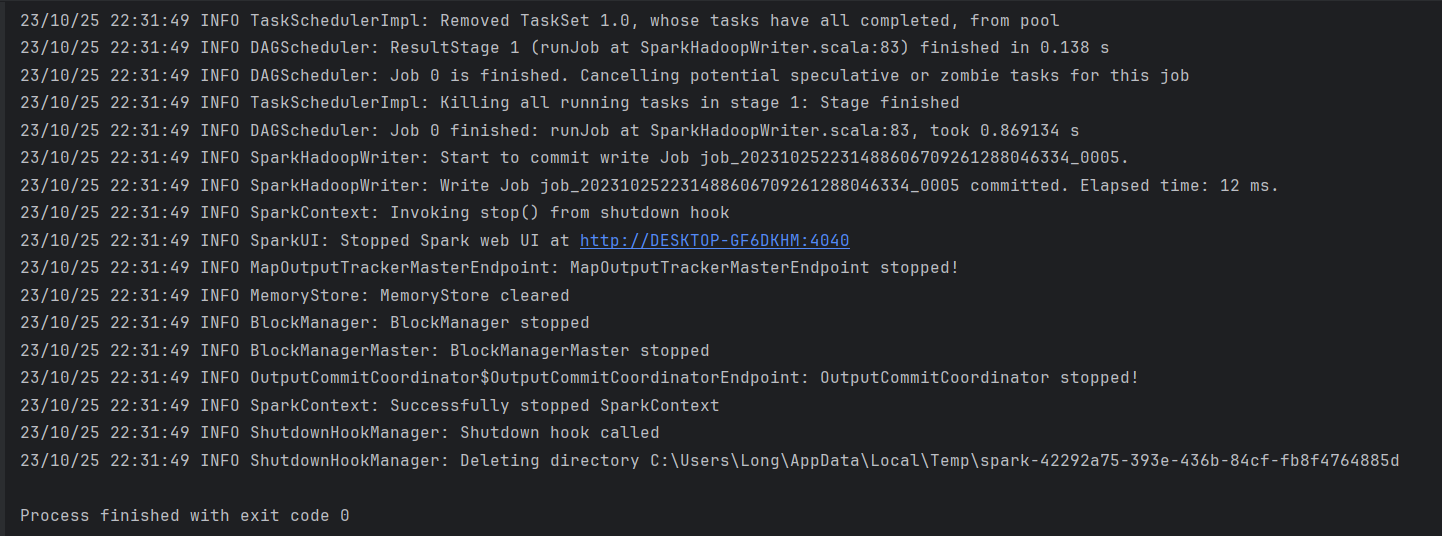
package WordCount  
  
import org.apache.spark  
import org.apache.spark.sql.SparkSession  
import org.apache.spark.{SparkConf, SparkContext}  
  
object WordCount {  
 def main(args: Array[String]): Unit = {  
 val sparkSession = SparkSession  
 .*builder* .master("local")  
 .appName("Scala Spark Example")  
 .getOrCreate()  
  
 // import sparkSession.implicits.\_  
 val sparkContext = sparkSession.sparkContext  
  
 val inputFile = "src/main/resources/Input/wordcount.txt"  
 val outputFile = "src/main/resources/Output"  
 val input = sparkContext.textFile(inputFile)  
 val words = input.flatMap(line => line.split(" "))  
 val counts = words.map(word => (word, 1)).reduceByKey { case (x, y) => x + y }  
  
 counts.saveAsTextFile(outputFile)  
 }  
}

# 4. Chạy chương trình

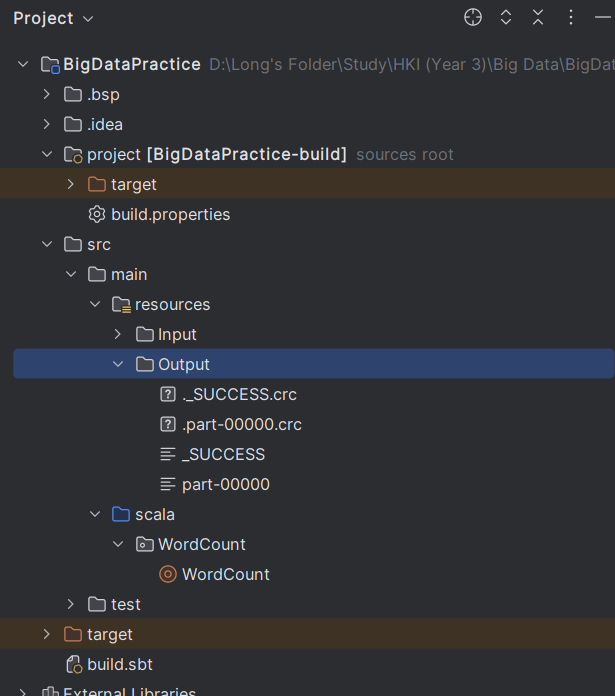
- Để chương trình chạy đếm từ



- Chương trình đã chạy xong.

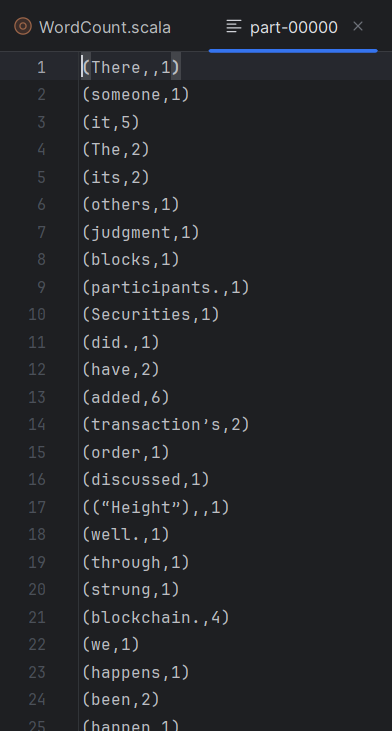


- File xuất sẽ được tạo trong mục Output.



# 5. Dữ liệu đầu ra

- Kết quả đếm từ.



Dữ liệu được đếm:

(There,,1)  
(someone,1)  
(it,5)  
(The,2)  
(its,2)  
(others,1)  
(judgment,1)  
(blocks,1)  
(participants.,1)  
(Securities,1)  
(did.,1)  
(have,2)  
(added,6)  
(transaction’s,2)  
(order,1)  
(discussed,1)  
((“Height”),,1)  
(well.,1)  
(through,1)  
(strung,1)  
(blockchain.,4)  
(we,1)  
(happens,1)  
(been,2)  
(happen,1)  
(look,1)  
(who,1)  
(clicking,1)  
(Commission,,1)  
(With,2)  
(along,2)  
(Once,1)  
(block.,2)  
(make,2)  
(Blockchain,,1)  
(users',1)  
(thousands,1)  
(called,1)  
(earning,1)  
(the,19)  
(however,,2)  
(are,1)  
(records,1)  
(consists,1)  
(most,1)  
(impulsive,1)  
(given,3)  
(when,1)  
(be,8)  
(all,2)  
(including,1)  
(as,3)  
(library,,1)  
(block’s,1)  
(multiple,2)  
(purchase,2)  
(together,1)  
(better,1)  
(Let’s,1)  
(hashed,,1)  
((More,1)  
(stored,2)  
(is,4)  
(about,1)  
(charge,1)  
(After,3)  
(check,1)  
(confirm,1)  
(on,1)  
(light.,1)  
(against,1)  
(Amazon’s,1)  
(at,1)  
(they,1)  
(new,3)  
(dollar,2)  
(likely,1)  
(Not,1)  
(cases,1)  
(information,,1)  
(details,1)  
(has,1)  
(four,1)  
(entries.,1)  
(Amazon,2)  
(thousands,,1)  
(Wikipedia,,1)  
(hastily,1)  
(or,2)  
(green,1)  
(data,,1)  
(By”),1)  
(anyone,1)  
(As,1)  
(available,1)  
(code,1)  
(Exchange,1)  
(see,1)  
(of,11)  
(unique,,1)  
(verified,1)  
(must,7)  
(making,1)  
(potentially,1)  
(occur.,1)  
(view—even,1)  
(checkout,1)  
(gets,1)  
(identifying,1)  
(accurate,,1)  
(this,1)  
(job,1)  
(vetting,1)  
(wings,,1)  
(rushes,1)  
(A,1)  
(Works,1)  
(once,1)  
(also,1)  
((“Time”),,1)  
(for,2)  
(network,2)  
(go,1)  
(transactions,,1)  
(your,9)  
(purchase,,2)  
(example,1)  
(up,1)  
(hash,1)  
(can,1)  
(was,1)  
(verified,,1)  
(signature,1)  
(packaged,1)  
(Bitcoin’s,1)  
(how,1)  
(hash.,2)  
(When,3)  
(time,,1)  
(said,1)  
(there’s,1)  
(by,1)  
(How,1)  
(like,2)  
(digital,2)  
(If,1)  
(an,1)  
(things,1)  
(That,4)  
(signature,,1)  
(happened,1)  
(recent,1)  
(Amazon,,1)  
(above,,1)  
(it.,1)  
(continue,1)  
(with,3)  
(hundreds,,1)  
(data,2)  
(Blockchain,1)  
(in,7)  
(transaction,9)  
(computers.,1)  
(angel,1)  
(In,1)  
(is,,1)  
(take,1)  
(block,10)  
(from,1)  
(public,1)  
(blockchain,,4)  
(join,1)  
(way,1)  
(verified.,2)  
(left,1)  
(publicly,1)  
(other,2)  
(you,6)  
(becomes,1)  
(computers,1)  
(name,1)  
(that,6)  
(a,13)  
(many,1)  
(stores,1)  
(local,1)  
(will,4)  
(information,2)  
(to,11)  
(unlike,1)  
(prompt,,1)  
(you.,1)  
(group,1)  
(access,1)  
(together.,1)  
(where,1)  
(amount,,2)  
(so,1)  
(suggests,,1)  
(and,4)  
((“Relayed,1)  
(transactions,1)  
(purchase.,2)  
(second.),1)