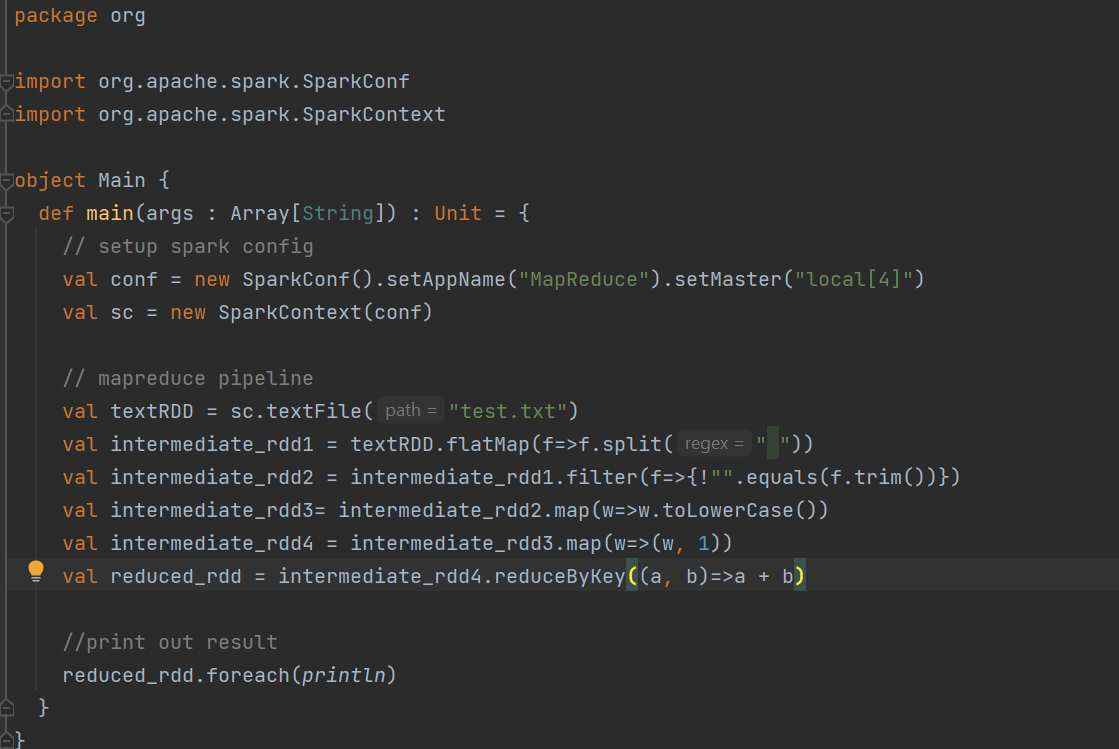
# HW1 WordCount Spark 实现

## 实验环境

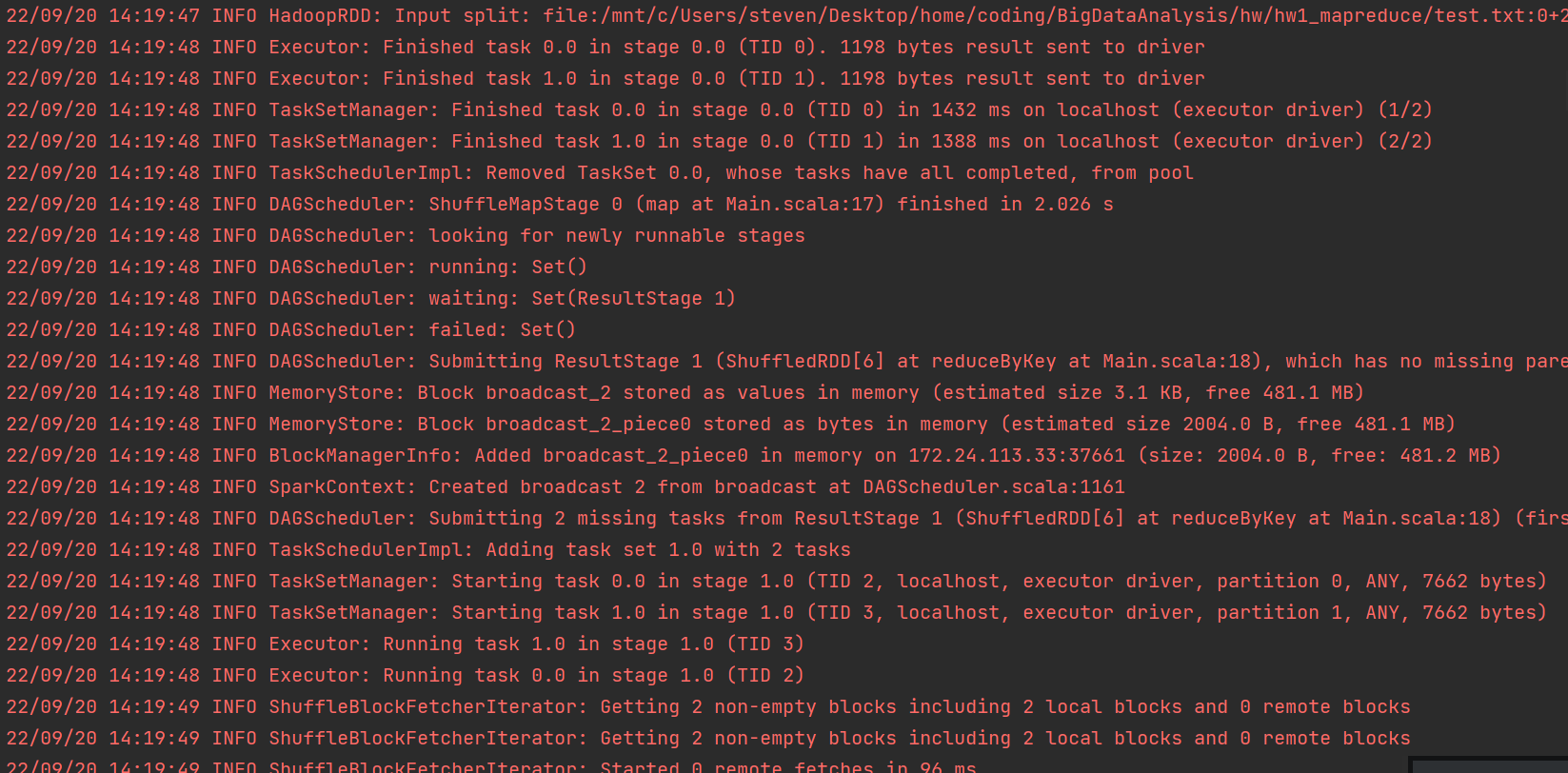
使用的是 Scala 2.11.7, Spark 2.4.4, 在 WSL 下运行

## 1.1源代码截图

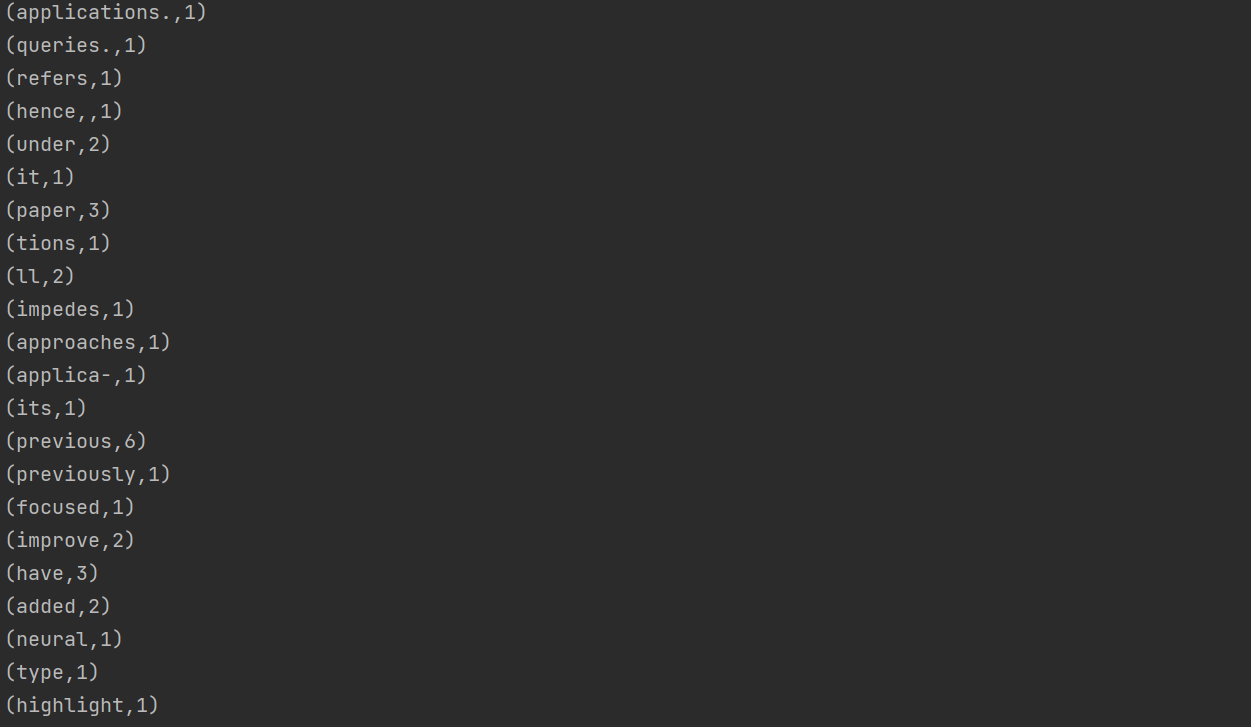


## 1.2运行截图

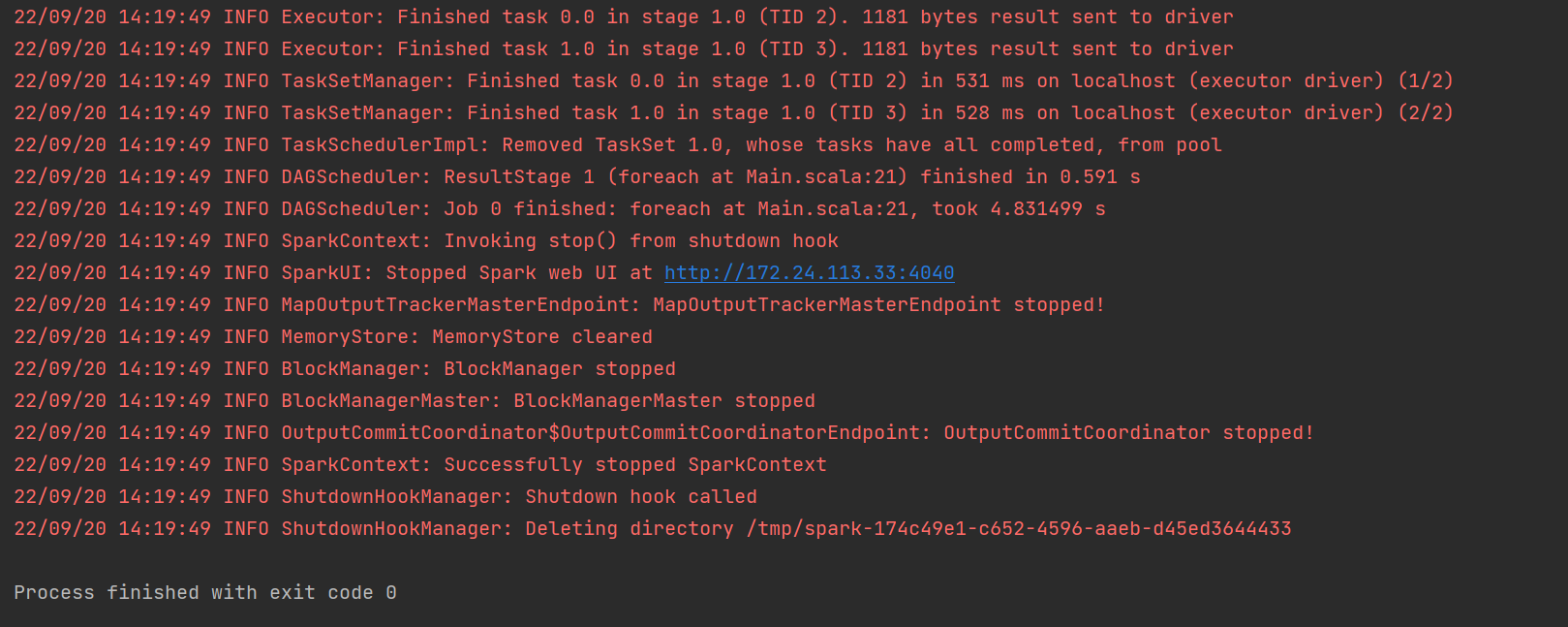
启动部分日志



输出结果(部分截图)



任务完成，退出



## 实现思路

将整个任务分成几个阶段，数据读取 => 文本分词 => 预处理(单词转换成小写) => map(把 word 映射为 (word, 1), 便于下一步reduce ) => reduce(按照 key 来reduce, 累加word出现的次数) => print result(输出到标准输出或者文件里)

## 实验心得

实验的代码实现比较简单，但是前期搭建环境花了比较多的时间。在 IDEA 中使用 Scala 需要安装插件，要用 spark 需要在 pom.xml 中引入依赖, 且需要注意 Scala 和 spark 的版本匹配, 此外还有 guava 的版本问题。最后选择在 WSL 下运行，因为 Hadoop 在 Windows 下直接运行会出现一些问题。