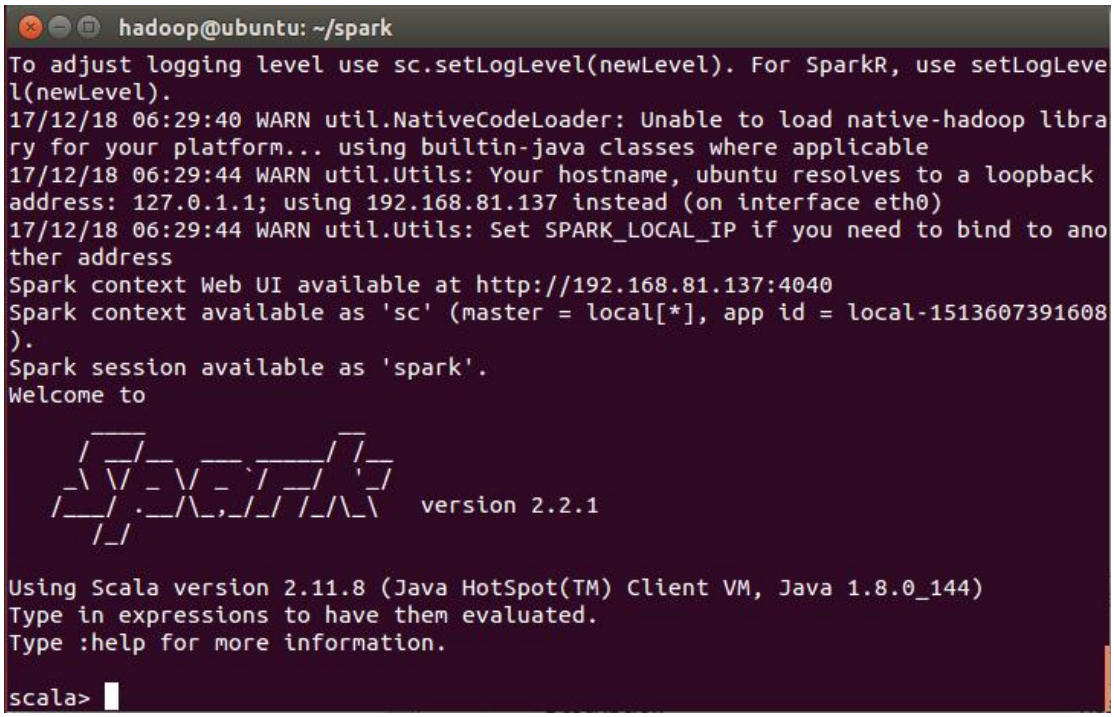
金融大数据 作业 9

151220006 陈安宇

1.启动 spark



2. 实现词频统计并排序

说明：先用 Wordseg.java 和 prepareText.java 去除停词和分词，再用 ScaWordCount.scala 进行词频的统计，并排序，最后输出到文件中。

提交 scala 代码截图：

```
hadoop@ubuntu:~/hadoop$ /home/hadoop/spark/bin/spark-submit --class "ScaWordCount" /home/hadoop/spark/ScaWordCount.jar
```

部分运行结果截图：

(公告,807)	(股票,117)	(卫星,5)
(板块,317)	(资产,116)	(令人,5)
(提示,245)	(突破,109)	(应收,5)
(涨停,241)	(拉升,104)	(变化,5)
(投资,220)	(股权,98)	(正式,5)
(股东,205)	(净利,97)	(国务院,5)
(业绩,191)	(股价,95)	(优质,5)
(今日,171)	(利好,94)	(增值,5)
(银行,156)	(增持,87)	(综合,5)
(中国,152)	(发展,85)	(严重,5)
(行业,144)	(揭秘,83)	(募投,5)
(晚间,137)	(报告,81)	(举行,5)
(控股,134)	(年度,77)	(显示,5)
(资金,128)	(持续,76)	(工程机械,5)
(交易,124)	(市场,70)	(水面,5)
(增长,119)	(国企改革,67)	(首家,5)
(年报,118)	(国际,65)	(制造业,5)
(电力,117)	(龙头,64)	(直接,5)
(股票,117)	(企业,64)	(缩水,5)
(资产,116)	(持股,63)	(危机,5)
(突破,109)	(产业,62)	(工行,5)