1. 对给出的莎士比亚文集Shakespere.txt进行wordcount（注：文件中包含特殊字符，请先进行过滤操作仅留下英文字符）

读入文本文件：

val lines =sc.textFile("file:///home/hadoop/Shakespeare.txt")

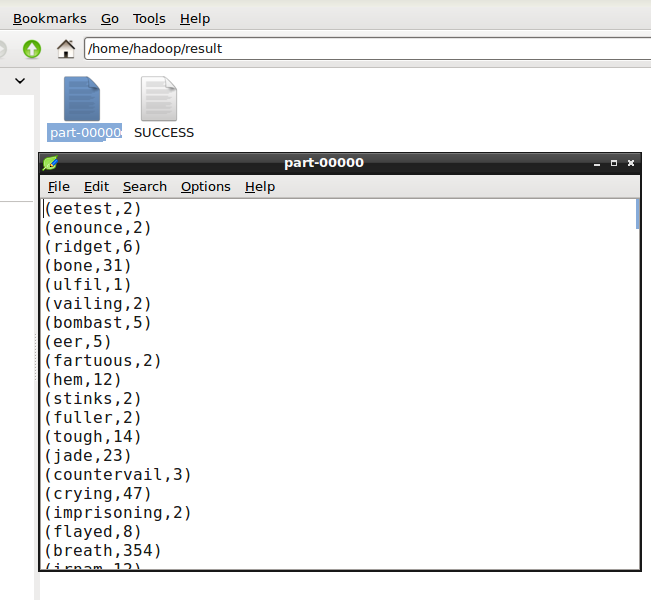
因为最终只统计英文单词的次数，需要去除其它字符，只留下英文字符，所以将每行按非英文字母分割即可，split函数支持正则表达式，然后把空字符串过滤掉，然后每个单词映射为(单词, 1)键值对，然后用.reduceByKey把相同键的值相加得到每类单词和其出现的次数：

val wordCount=lines.flatMap(line => line.split("[^a-z]")).filter(word => !word.isEmpty()).map(word=>(word,1)).reduceByKey((a, b) => a + b)

保存结果：

wordCount.saveAsTextFile("file:///home/hadoop/result")

结果截图：



可以按照单词出现的次数从大到小排序：

val lines =sc.textFile("file:///home/hadoop/Shakespeare.txt")

val wordCount=lines.flatMap(line => line.split("[^a-z]")).filter(word => !word.isEmpty()).map(word=>(word,1)).reduceByKey((a, b) => a + b).sortBy(\_.\_2,false)

wordCount.saveAsTextFile("file:///home/hadoop/result\_sort")

结果截图：

