# 实验报告

姓名：张嘉铭

学号：1829210092

## 实验名称：Coblab-1：单词计数

### 任务概述：

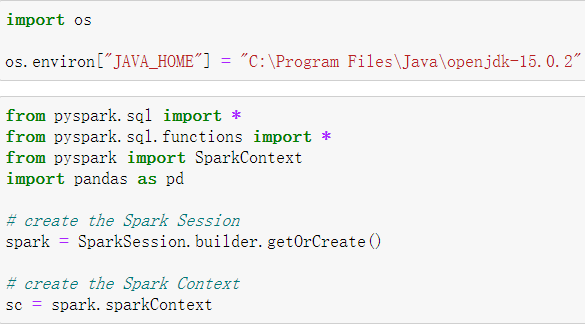
给定一个英文文本文件，使用spark的map-reduce相关方法统计文本中所有单词首字母的出现情况。

### 使用数据

pg100.txt: 一份关于莎士比亚作品全集的文本文件。

### **实**验过程及结果

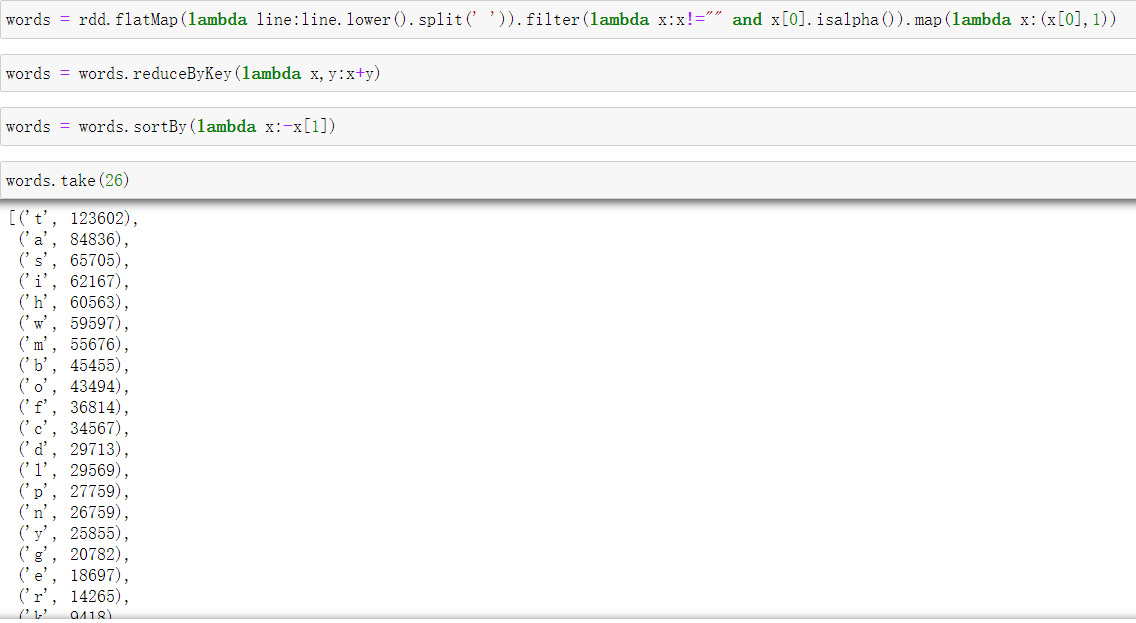
1. 环境配置



1. 读入文本数据



1. 用空格分开，过滤掉空字符串和首字母不为字母的情况，并将其映射为 (字母, 1)，再进行reduce操作，最后对结果进行排序。



### 实验中的问题及解决方法

会有空行出现，不然在取首字母时会出错，需要过滤掉。

只要首个字词为字母就可以统计，不需要整个都由字母组成，不然和答案对不上。