关键词分类——运行说明文档

参赛id：stdcoutzyx

首先，将源代码打包为jar包，不指定主类，在提交结果中已经包含了一个已经打好的jar包了。

需要运行四个命令得到最终结果。在实验中，KWSeg与GroupSVMClassify类都遇到过运行到中间job异常结束的情况，再次运行即可运行成功，为了确保所有流程都可运行通过，这里将总流程拆分为四个，如果其中某个运行失败，重复执行该处的命令即可，避免了每次都从头开始。

四个命令如表1所示：

表 1 运行所有流程所需要的四个命令

|  |  |
| --- | --- |
| 编号 | 命令 |
| 1 | hadoop jar kwclassify.jar com.zyx.exfeature.KWSeg your\_path/keyword\_class.txt your\_path/kwseg false |
| 2 | hadoop jar kwclassify.jar com.zyx.exfeature.ExFeature your\_path/kwseg you\_path/ex\_features/ |
| 3 | hadoop jar kwclassify.jar com.zyx.svm.GroupSVMClassify your\_path/ex\_features/test\_data/ your\_path/ex\_features/train\_data/part-r-00000 your\_path/group\_out 0 0 |
| 4 | hadoop jar kwclassify.jar com.zyx.svm.VoteResult your\_path/group\_out your\_path/vote\_result |

上面的命令显示的都不是很直观，空格不明显，可参考同目录下的文件command\_example。

对于每个命令，参数各不相同，下面对参数进行介绍。

对于命令1，它调用KWSeg完成分词。它有三个参数，前两个为输入文件与输出文件夹，第三个参数必须是true或者false，它代表IKAnalyzer不同的分词模式，true代表智能分词，false代表细粒度分词。实验表明，使用细粒度分词效果较好，所以命令中选择了false。

对于命令2，它调用Count和IndexData两个类中的Map-Reduce过程（见《代码说明文档》），在输出文件夹下得到三个文件夹——count、test\_data、train\_data。其中test\_data、train\_data下均有输出标志着该命令成功运行。

对于命令3，调用GroupSVMClassify类完成分组的svm分类（详细见《程序说明文档》）。它有五个参数，第一个参数为测试集输入，第二个参数为训练集输入，该参数所代表的文件将会在setup中读入，因而必须具体到文件名part-r-00000。第三个参数为输出文件夹，由于程序会运行(组数-1)\*组数/2个job，因而在该文件夹下会产生(组数-1)\*组数/2个子输出文件夹，在现有的jar包中组数=8。每个输出文件夹的名称为（groups组数1-组数2），如groups0-1、groups6-7等，其中组数1的值小于组数2的值。如图1所示。

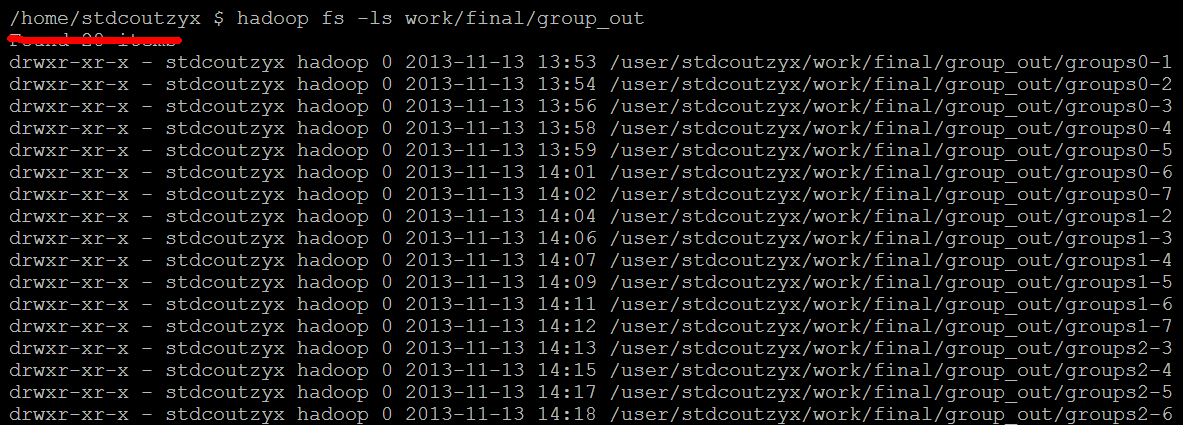


图 1 输出文件夹下的子文件夹命名示例

第四五个参数用于按照进度重启任务，用于解决任务异常终止需要重启时不重复执行已经成功的job以节省时间，在开始时，这两个参数为0、0，表示job从组号0、1开始执行，假如遇到程序异常终止，查看结果文件夹发现已经生成到了多个文件夹，找到最后生成的文件夹groupsi-j——即失败前最后成功执行的job所生成的文件夹，最后的groupsi-j的识别在于i最大且在i最大的情况下j最大。比如在图1的情况下最后生成的文件夹为groups2-6。在重启任务时将最后两个参数设为最后生成文件夹的i、j，以图1为例，i=2，j=6。这样就避免了已经成功的job的重新执行。该命令执行成功的标志是在输出文件夹下生成28个子文件夹。

对于第四个命令，有两个参数，第一个参数为上一步的输出文件夹，该值必须与上个命令的第三个参数保持一致。第二个参数为输出文件夹。

四条命令的大致运行时间如表2所示：

表 2 各项命令运行时间

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 命令编号 | 1 | 2 | 3 | 4 | 合计 |
| 运行时长(min) | 3.5 | 3.5 | 50 | 15 | 约70 |

需要注意的是：本程序的执行利用到了当前平台的一个性质，即所有文件的输出都是在输出文件夹下的part-r-00000文件里，程序反复应用到了这一性质，如果评委的测试平台与现有平台里该特性不一致，则不能得到正确结果。