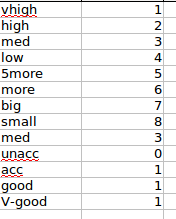
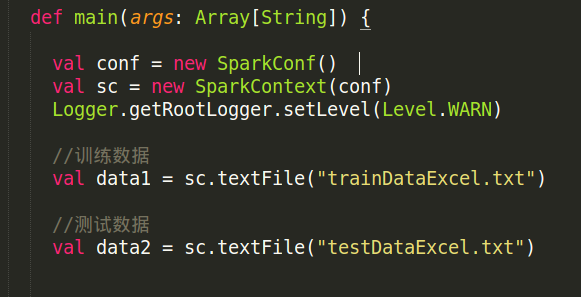
总结：提交的代码有两个：TreeDemo.scala（决策树）和RandomForest.scala（随机雨林），其中本来打算用随机雨林做4分类，但是当我把树的棵数调到18000的时候，提示内存不够，然后我把树的棵数降到5000，但遗憾的是，还是没有完成4分类的任务。我再次看了一下数据，在数据量为1800左右的TrainData中，good和vgood的数据实在太少，可能这是主要没有能完成4分类任务的主要影响因素。

在学习MLlib的时候，发现其中的向量元素只支持Double和Int类型，但是本次实验给出的训练数据和测试数据中含有String类型，没办法，时间有限，找到了一个StringIndexer的方法，但是没有尝试成功。

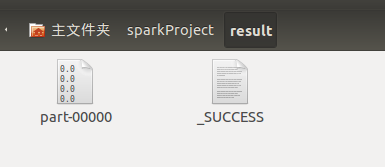
所以选择了一种较笨的方式来解决，使用excel手动把String类型的数据映射成Int类型，映射关系如下：



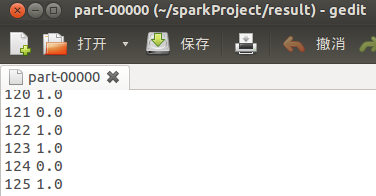
然后导出txt为：trainDataExcel.txt(提交的文件中有,如果您需要提交到集群上运行，请使用这个txt)，作为训练集。使用同样的方法生成testDataExcel.txt。然后把这两个txt扔到算法中进行训练和预测：



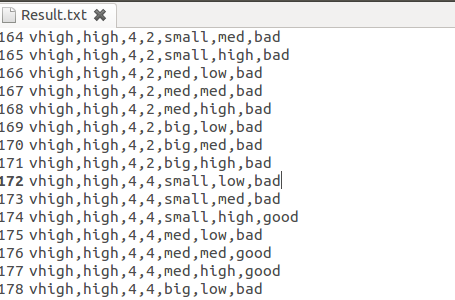
最后得到的结果是Double型的预测值：（使用repartition把分区设置为了1）



得到的预测结果是下面的形式。然后再按照之前的映射方式，把这个double型的数据映射回原来的good和bad。



最后使用excel把映射回去的结果（一列），copy到实验原始数据testData的后面，最后以csv的格式导出，另存为Result.txt：



附：sbt配置如下：使用的spark平台是2.0.2,scala版本是2.11.8,java版本是1.8

