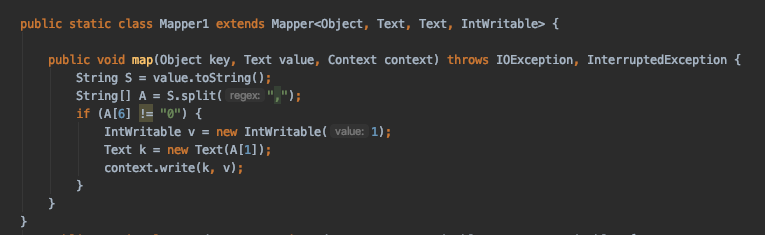
实验四实验报告

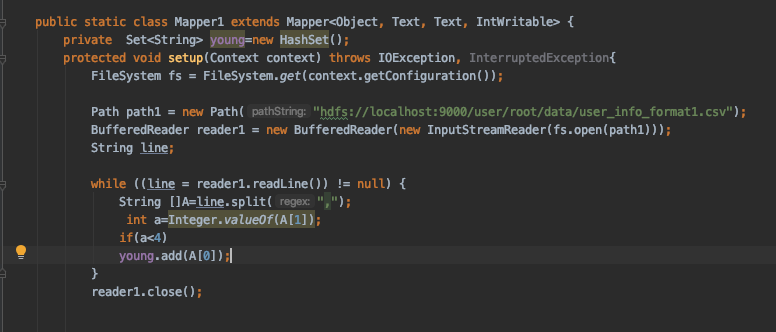
1. MapReduce统计双十一最热门商品与年轻人最关注的商家
2. 统计最热门商品：map 阶段与 wordcount 类似，需要注意的是需要添加一个判断条件，使得action为“0”，即仅仅点击商品而没有购买，添加购物车或收藏的数据被筛选掉



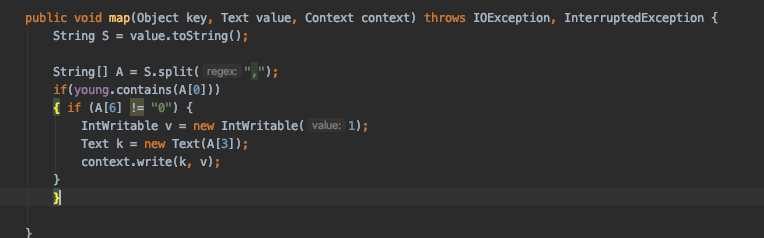
1. 统计年轻人最关注的商家：

具体思路与 1.（1）类似，reduce 部分的代码完全一致，需要改变的是在map 阶段之前需要先从 user\_info.csv 中读取年龄小于 30 的购买者的 user\_id,并存放于一个数组中，然后在 map 阶段进行判断，如果这一条数据对应的 user\_id 没有存在于这个数组中，即说明这个购买者年龄大于30，因此需要筛除

setup 函数，用于读取 user\_Info 表

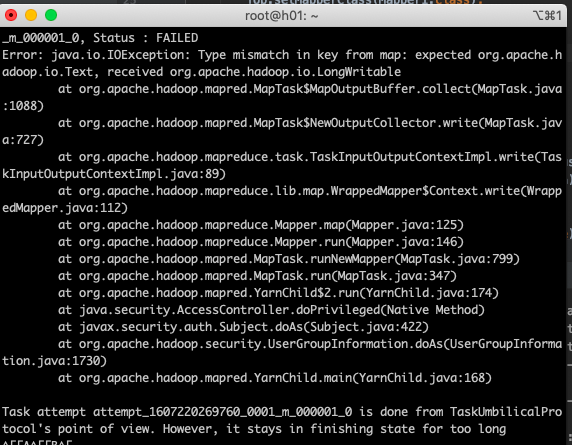


map阶段的代码：

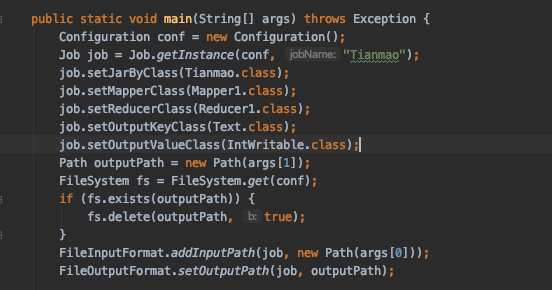


1. 这一步出现的问题：

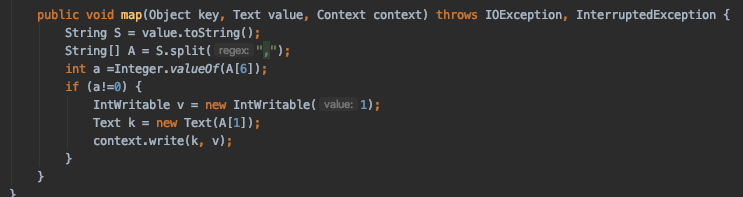
（a).



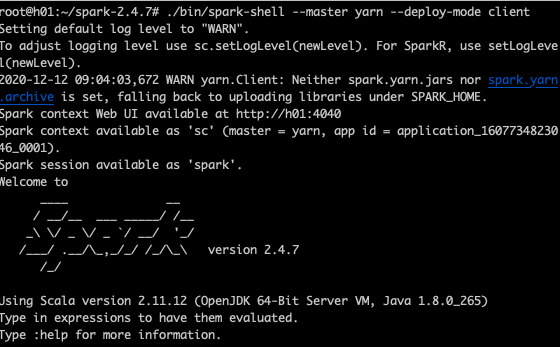
这里的第五行和第六行需要写成 Mapper1和 Reducer1而不是 Mapper 和 Reducer，否则会出现上面的报错



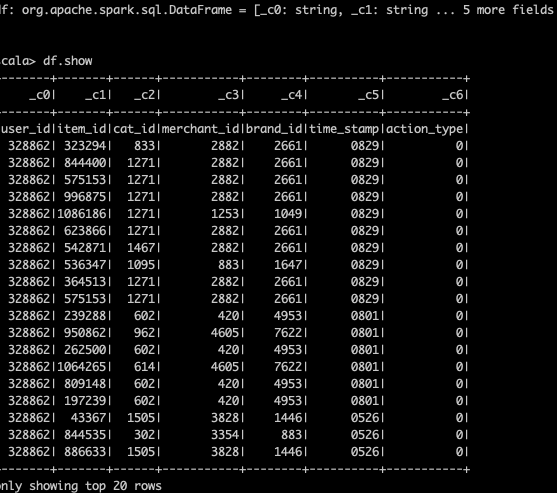
(b).这里在进行判断之前需要先将 A[6]转换成int 型的数，而非直接当做字符串与“0”判断，（1）的截图中的代码就是这样的错误，这样可以运行，但是会导致最终的结果不对，可能与数据集文件的格式有关



1. 代码和结果文件在tianmao（mapreduce 实现）文件夹中
2. 在Spark和 SparkSql下进行的实验（代码在 Spark 文件夹中）



docker 下启动 spark，并读取csv 文件

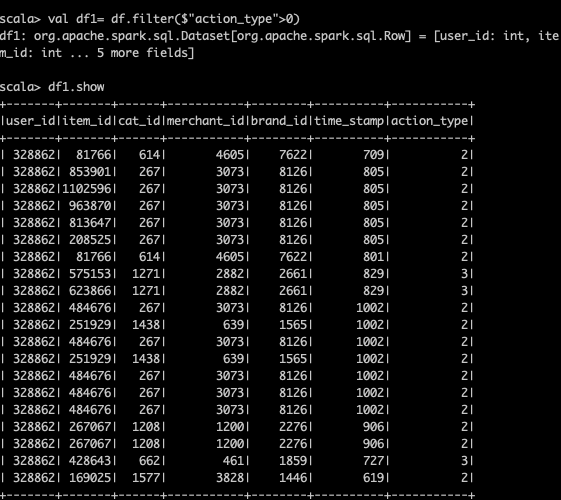


（一）.Spark统计最热门商品

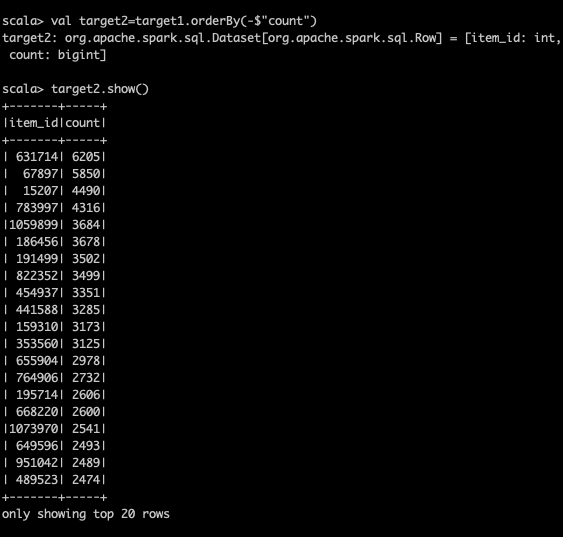
将读取到的数据类型从 strting 转换成 int，方便之后的filter,groupBy和 orderBy操作



通过 filter(action\_type) 筛选掉“仅仅进行点击而没有进行其他操作”的记录

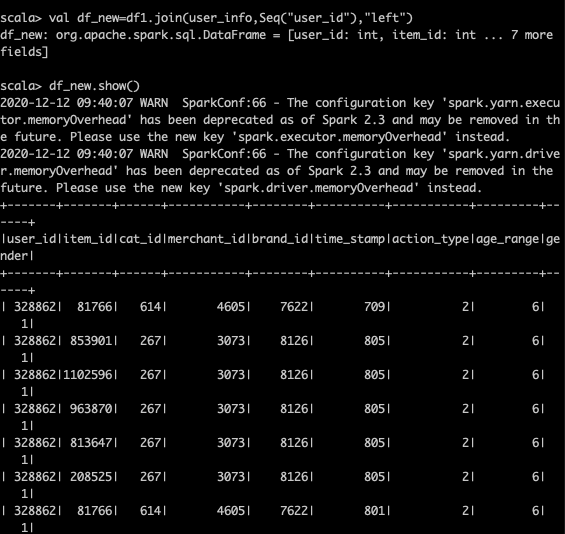


分组并进行排序

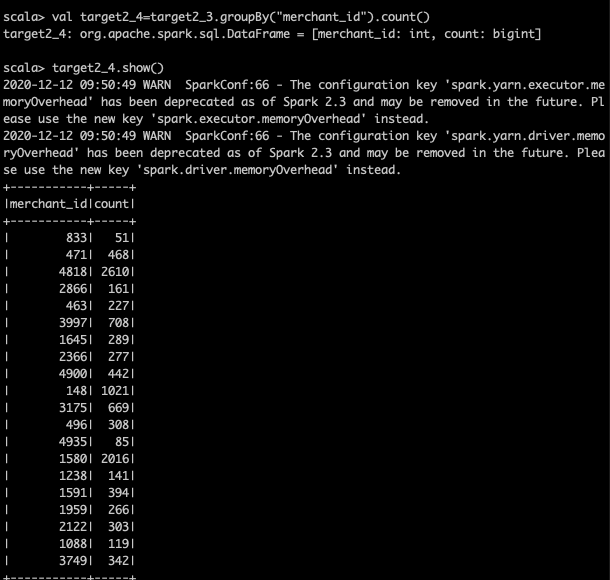


（二）.Spark 统计最受年轻人关注商家

将 user\_log 和 user\_info 两张表通过 join 合并在一起



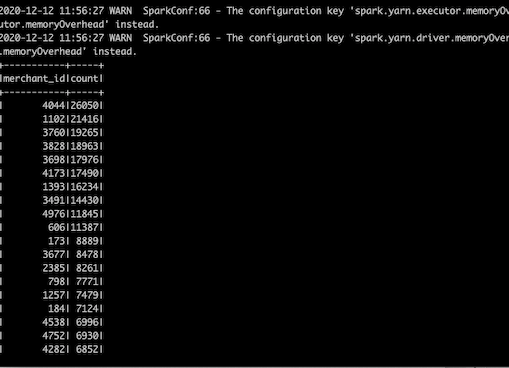
根据 merchantise\_id进行分组



排序并取前 100



运行结果：

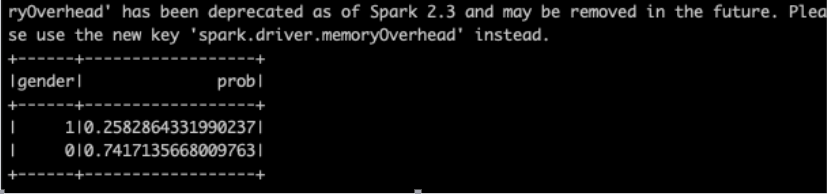


（三）.Spark 统计购买商品的买家中男女比例

在（一）（二）得到的中间表的基础上统计 action\_type==2（即发生了购买操作）的买家中男女各自的 count，并进行求和得到 sum，并分别除以这个 sum 得到比例；

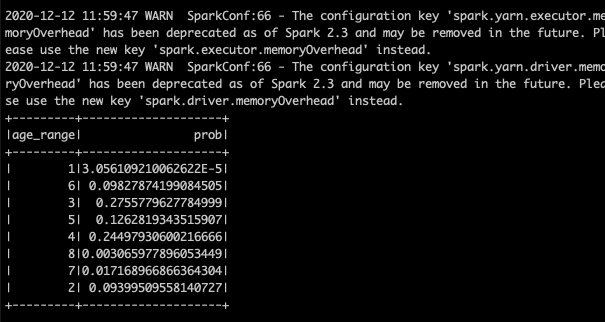


注意这里需要 filter 掉gender 为2 和 null 的情况



（四）.Spark 统计购买商品的买家中年龄比例

在（一）（二）得到的中间表的基础上统计 action\_type==2（即发生了购买操作）的买家中各个 age\_range的 count，并进行求和得到 sum，并分别除以这个 sum 得到比例；



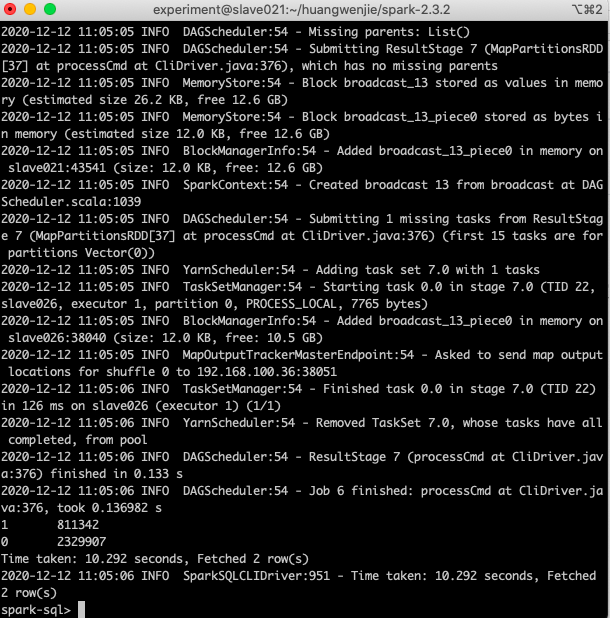
（五）SparkSQL查询买家中男女比例

具体思想与（三）类似，这里不再赘述

截屏2020-12-12 上午11.14.25

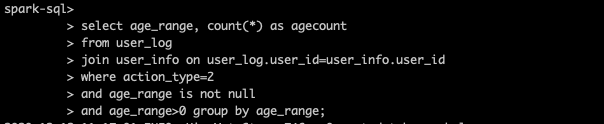
截屏2020-12-12 上午11.07.40

运行结果：

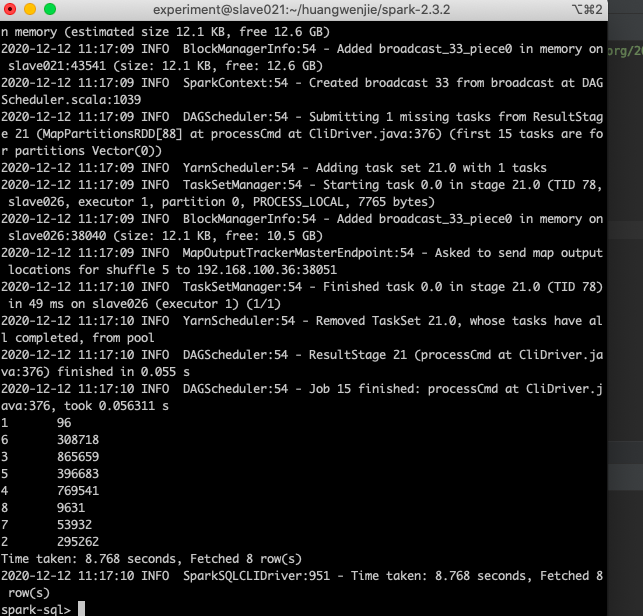


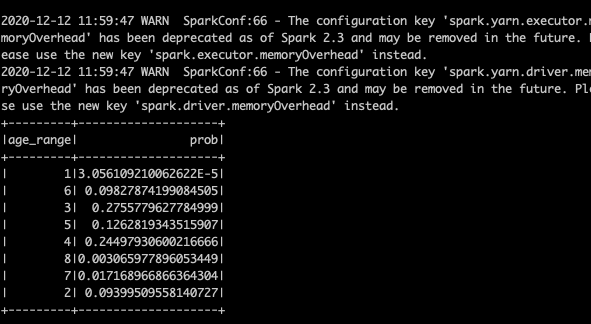
（六）SparkSQL查询买家中年龄比例

具体思想与（四）类似



运行结果：



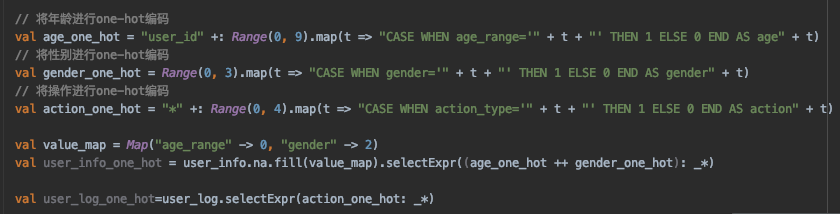


1. 基于 Spark MLlib进行的回头客预测（代码和最终结果也在 Spark 文件夹中
2. 训练集数据的准备：

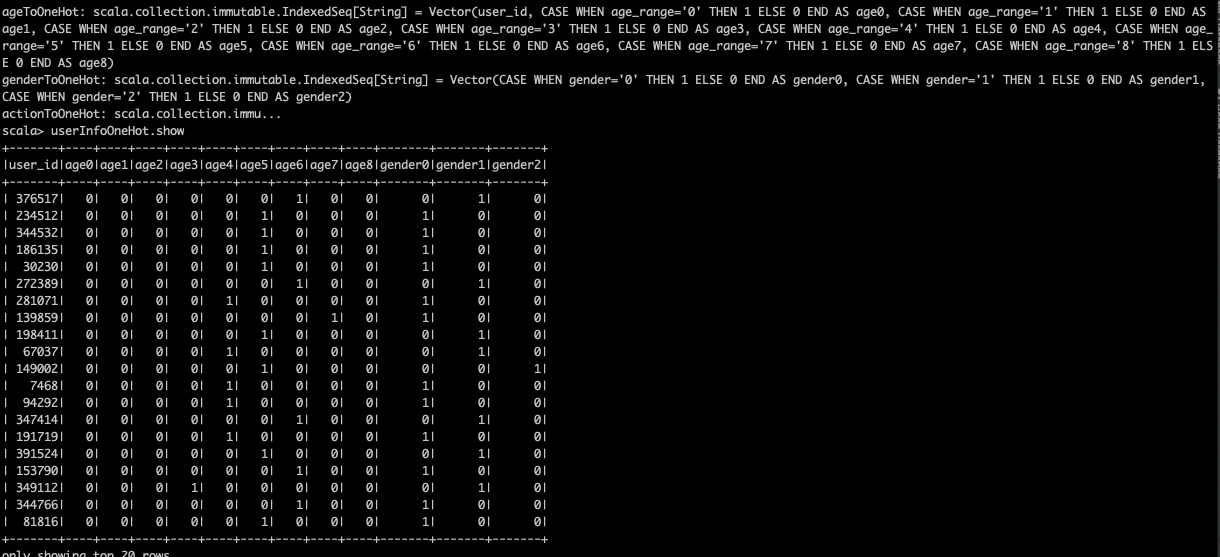
（1）读取数据，将两张表合在一起，并添加一列 Source,表示该数据来自训练集还是测试集



（2）将数据转变成0-1编码，方便之后对特征值的计算



转换之后的结果

（二）.计算特征值：

1. 计算用户的特征值：



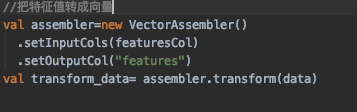
（2）计算商家的特征值：



1. 计算商家和用户的联合特征值：

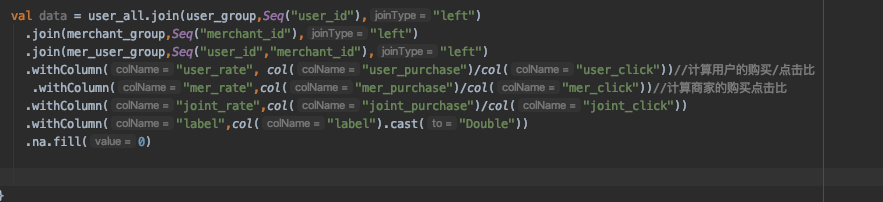


1. 将特征值转换成向量

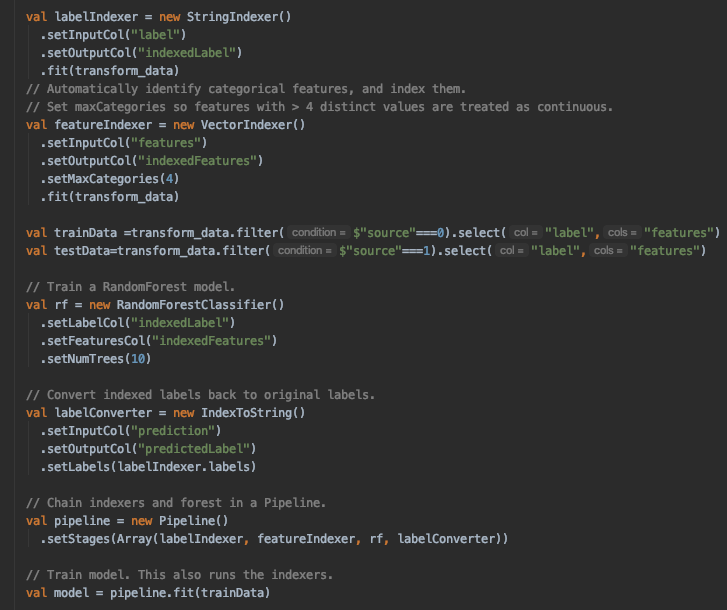


1. 利用 SparkMLlib 模型进行预测

准备训练数据，计算用户和商家的”购买点击比率“



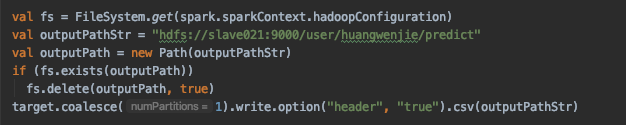
机器学习模型，



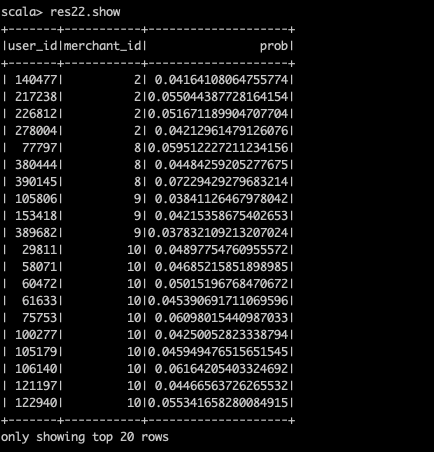
获取运算结果，并计算概率



输出结果



1. 最终结果和网页排名



提交至网站后所获得的分数与排名

