实验 2

背景：

在线上线下融合消费的时代，商家常通过发放优惠券吸引顾客，但如何让优惠券精准触达潜在用户并促进其消费，是当前面临的主要挑战。本次实验提供用户在 2016 年 1 月 1 日至 2016 年 6 月 30 日之间真实线上线下消费行为，可以用于预测用户在 2016 年 7 月领取优惠券后 15 天以内的使用情况。（注：为了保护用户和商家的隐私，所有数据均作匿名处理，同时采用了有偏采样和必要过滤。）

本次实验的数据包括用户和商家信息、优惠券折扣率、优惠券领取和使用情况、消费距离、消费日期，旨在通过数据分析优惠券使用规律。（注：本次作业不涉及预测任务，有兴趣的同学可以自行尝试。）

数据来源：

https://tianchi.aliyun.com/competition/entrance/231593/information

数据描述：

表格描述

本次实验共提供四个数据文件，ccf_offline_stage1_train.csv 和 ccf_online_stage1_train.csv 作为训练数据，分别为用户在 2016 年 1 月 1 日至 2016 年 6 月 30 日之间真实线下和线上消费行为，根据 ccf_offline_stage1_test_revised.csv 预测用户在 2016 年 7 月领取优惠券后 15 天以内的使用情况。sample_submission.csv 为预测结果样例。

注：sample_submission.csv 本次作业无需使用

文件名	数据描述
ccf_offline_stage1_test_revised.csv	用户 O2O 线下优惠券使用预测样本
ccf_offline_stage1_train.csv	用户线下消费和优惠券领取行为
ccf_online_stage1_train.csv	用户线上点击/消费和优惠券领取行为
sample_submission.csv	预测结果样例

字段描述

训练集和测试集中均包含用户信息、商家信息、优惠券信息等，训练集提供优惠券被使用信

息，测试集则不包含，online 表和 offline 表分别额外给出消费券动作和消费距离信息。

列名	含义	示例
User_id	用户标识符	2166529
Merchant_id	商家标识符	7113
Coupon_id	优惠券标识符[1]	6928
Discount_rate	折扣率[2]	200:20:00
Distance	User 与 Merchant 距离[3]	5
Date_received	优惠券领取日期	20160727
Date	优惠券使用日期[4]	NULL
Action	0 点击，1 购买，2 领取优惠券	2

注释：

[1] Coupen_id 为 NULL 表示无优惠券消费，此时 Discount_rate 和 Date_received 字段无意义，“fixed”表示该交易是限时低价活动。

[2] x∈[0,1]代表折扣率；x:y 表示满 x 减 y，fixed”表示低价限时优惠。单位是元，第二个冒号后 00 无意义。

[3] User 经常活动的地点离该 Merchant 的最近门店距离是 x*500 米（如果是连锁店，则取最近的一家门店），x∈[0,10]；null 表示无此信息，0 表示低于 500 米，10 表示大于 5 公里。

[4] 消费日期：如果 Date=null & Coupon_id != null，该记录表示领取优惠券但没有使用，即负样本；如果 Date!=null & Coupon_id = null，则表示普通消费日期；如果 Date!=null & Coupon_id != null，则表示用优惠券消费日期，即正样本。

实验任务

任务一：消费行为统计

根据 ccf_offline_stage1_train 和 ccf_online_stage1_train 表中数据，统计每个商家的优惠券使用情况，分别为领取优惠券未使用、未领取优惠券直接消费和领取优惠券并使用三种情况，线上和线下分开统计。

输出格式：

```
<Mechant_id> TAB <负样本数> TAB <普通消费数> TAB <正样本数>
```

任务二：商家周边活跃顾客数量统计

消费者与发券商家的距离很大程度影响优惠券是否被线下使用，距离越近可以被认为越活跃。根据 ccf_offline_stage1_train 表中数据，编写 MapReduce 程序，对每个商家与周边消费者的距离进行统计，给出不同距离的活跃消费者人数。注意表中 Distance 字段缺失为 NULL。

输出格式：

```
<Mechant_id>  TAB  <距离为 x 的消费者人数>
```

任务三：优惠券使用时间统计

根据 ccf_offline_stage1_train 表中数据，统计每一种优惠券的被使用次数，Coupon_id 缺失项不计入总使用次数，对于被使用次数大于总使用次数 1%的优惠券，给出它们从领取到被使用的平均间隔并排序。

输出格式：

```
<Coupon_id>  TAB  <平均消费间隔>
```

任务四：优惠券使用影响因素分析

优惠券使用行为（如优惠券是否被使用、使用时间）受到很多因素的影响。例如：商家种类，优惠券折扣率，商家距离等。

在上面的三个任务中，我们研究了商家距离和优惠券种类因素。现在，请你**自行选取可能影响优惠券使用行为的若干因素**作为研究对象，通过 MapReduce（或其他工具），根据统计结果（类似于上面三个任务的结果）阐述这些因素对优惠券使用行为的影响。

分析示例：

- 折扣率对消费行为的影响：可以根据 Discount_rate，计算不同优惠券的实际折扣率，或分析满减方案，统计不同方案下优惠券的使用情况，分析折扣率与使用概率和使用时间的关系。

即使你的结论是某一因素对用户的消费行为没有显著影响，这样的结果也是完全 OK 的。本次实验重点关注的是使用 MapReduce 进行统计的过程。

提交方式

提交 git 仓库地址或者相关文件的 zip 包。实验报告应包括设计思路、运行结果和可能的改进之处等。