实验4：Spark 编程

实验任务

实验背景及数据集同实验2

任务一：Spark RDD编程

描述：

请使用**Spark RDD**编程的方式，完成下面的任务。

1. **统计优惠券发放数量：** 使用 `ccf_online_stage1_train` 统计每种优惠券的被使用次数，并按数量降序排列输出，完整结果以附件形式给出，实验报告中给出前十名优惠券的结果。

   输出格式：

```
<Coupon_id>  <总使用次数>
```

2. **查询指定商家优惠券使用情况：** 使用 `ccf_online_stage1_train` 统计每个商家的优惠券使用情况，分为负样本、普通消费和正样本三种，按照 `Mechant_id` 升序排序并将结果存储在新表 `online_consumption_table` 中，实验报告中给出前十行结果。

   注：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。

   输出格式：

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

任务二：SPARK SQL编程

描述

请使用**Spark SQL**编程的方式，完成下面的任务。

1. **优惠券使用时间分布统计：** 根据`ccf_offline_stage1_train`表中数据，统计每一种优惠券被使用时间位于一个月的上中下旬。给出每一种优惠券被使用时间的分布。

   输出格式：

```
<Coupon_id>  <上旬被使用概率>  <中旬被使用概率>  <下旬被使用概率>
```

2. **商家正样本比例统计：** 根据`online_consumption_table`表中数据，按正样本比例对商家排序，给出正样本比例最高的前十个商家。

   输出格式：

```
<Merchant_id>    <正样本比例>     <正样本数量>     <总样本数量>
```

任务三：Spark MLlib 编程

描述：

请完成以下预测任务：天池新人实战赛o2o优惠券使用预测

要求

1. 使用 **Spark MLlib** 提供的机器学习模型，预测用户在2016年7月领取优惠券后15天以内的使用情况。
2. 【可选】在阿里云天池平台注册并报名参赛，按赛题要求提交预测结果，并在实验报告中记录预测结果并提交平台所得的评分，附上评分截图作为证明（“我的成绩”页面）。

   建议和参考：

- 可以尝试决策树模型或者Logistic回归
- 推荐挖掘基础特征外的多种特征，如使用时间间隔、折扣方式、使用日期等

   注意事项：

- 本实验仅限使用Spark MLlib编程，得分高低不是重点。
- 如果愿意，可以在完成上述任务的基础上，进一步探索，提升评分和排名。

实验报告要求

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和运行结果，**完整代码和结果可放附录**。 需要在实验报告中给出解答每一个问题的**核心代码** 并给出对应的**部分结果截图**。在此基础上，可以写本次实验的感受以及收获。

注意：实验报告无字数要求，可以选择提交github仓库地址，但**请注意实验报告的撰写格式**，建议使用**MarkDown**编写，给出清晰的实验报告大纲，清楚描述实验过程和实验结果即可，**避免大段截图堆砌和日志内容**。如果提交word或pdf，请尽量不要超过20页。