数据科学基础第二次实践

考查内容：频繁模式与关联规则挖掘，考察对 Apriori&&FP-Growth 算法的掌握程度。

一．任务描述：

1. 在给定数据集上（10,000 量级的数据）上使用关联规则挖掘算法。
2. 通过改变不同等级的支持度和置信度，比较 Apriori，FP-Growth 和 baseline 算法（说明见）的性能。
3. 试图根据 Apriori 或 FP-Growth 算法揭露出的物品集相关性，发现一些有趣的关联规则。

二．数据集说明

此次实验，我们采用 GroceryStore 数据集，该数据集包含一个商店一个月的交易记录。数据集内容见 Groceries.csv 文件，其中，每一行包含一条交易记录，购买物品的列表在打括号“{}”里面，物品之间用逗号“,”分割；需要注意的是，有些物品名用空格 space 分割，有些用斜杠“/”分割。整体来看，共包含 9835 条交易记录，物品总数为 169。

三．Baseline 算法--dummy Apriori 算法

我们实验的 baseline 算法称为 dummy Apriori 算法。相比于正常的 Apriori 算法（下面称为 advanced Apriori），dummy Apriori 没有进行剪枝 trick，即用暴力搜索的方法根据 k 项频繁项集生成候选 k+1 项集，对事务表不做任何处理。

四．Apriori 算法的三种剪枝策略

Advanced Apriori 可以有如下三种剪枝策略：

1. Advanced Apriori_1: 减小生成候选项集规模。
   如果一个 k+1 候选项是频繁的，那么生成它的 k 频繁项集中必包含其 k+1 个子集。例如，如果{t1, t2, t3, t5}是频繁的，那么它将由{t1, t2, t3}和{t1, t2, t5}生成，否则它就不会被生成。这么做要求生成候选项集的频繁项集的项内是有序的，各项之间也是有序的。
2. Advanced Apriori_2: 减小事务表（数据记录表）规模。
   如果一个 k+1 项候选项能与事务表中一条数据记录匹配，那么其 k+1 个子集也必能与事务表中该条记录匹配。因此在事务表中匹配 k 候选项集的时候，统计每条记录被匹配到的次数，如果少于 k+1 次，那么将该条记录从事务表中移除，因为它绝不可能与下一轮的任一 k+1 项候选项匹配。这样每次迭代都减小了事务表的规模，从而减小扫描事务表的时间消耗。
   例如，对一条数据记录{t1, t2, t4, t5}，对于一个与该数据记录匹配的 3-项候选集{t1, t2, t4}时，其 2-项候选集{t1, t2}, {t1, t4}, {t2, t4}必然都与该记录匹配，因此在匹配 2 候选集时，该记录至少会被匹配 3 次。因此少于 3 次产生匹配的数据记录（事务）可以不予考虑，从而不断降低事务表的规模。
3. Advanced Apriori_3: 减少事务表中元组的项。
   如果事务表中元组的某一项能包含在一个 k+1 频繁项中，那么该项必出现在这个 k+1 频繁项的 k 个 k 项子集中。所以在扫描事务表统计 k 项候选集的出现频次时，如果事务表中任一元组的某一项未被匹配中 k 次,那么该项将在筛选出 k 频繁项集后从元组中除去，从而减少统计 k+1 项候选集频次时与事务表中元组的匹配次数。具体实现在减少事务表规模实现的基础上稍作修改即可。

例如，对事务表中的一条数据记录{t1, t2, t4, t5}, t2 项目能出现在 3-频繁项{t1, t2, t4}中，因此其一定也会出现在其对应的 2-频繁子集{t1, t2}, {t4, t5}中，即在扫描 2 项集出现的频次时，t2 会出现 2 次。若 t2 没有出现 2 次，其也不会在 3 项集中出现，此条数据记录元组可以缩减为{t1, t4, t5}。

五．算法分析模块

在算法实现之后，要对算法做若干分析：

1. 在给定支持度的情况下，挖掘频繁项集。
2. 在给定支持度和置信度的情况下，挖掘关联规则。
3. 寻找一些置信度高的规则，进行该规则产生的逻辑内涵。
   例如为什么啤酒尿布为什么会一起出现。

六．实验基本要求（85 分）

1. 数据处理（15 分）
   能够 load 并对数据做适当的预处理。
2. 算法实现（50 分）
   第三部分，手动实现 dummy Apriori 算法-20 分
   第四部分，手动实现 advanced Apriori 算法的前两种剪枝策略-10*2=20 分调用库 pyfpgrowth 实现 pyfpgrowth 算法-10 分
3. 算法应用（20 分）
   使用 Apriori 算法在 GroceryStore 数据集上挖掘频繁 3-项集（支持度 0.01）-5 分
   使用 Apriori 算法在 GroceryStore 上挖掘关联规则（支持度 0.01，置信度 0.5）-5 分
   使用 FP-Growth 算法在 GroceryStore 数据集上挖掘的一些关联规则（置信度 0.5）-5 分
   算法效率分析，对比 dummy Apriori、仅使用第一种剪枝策略 Advanced Apriori_1、同时使用第一种和第二种剪枝策略 Advanced Apriori_12 的时间损耗，画图对比在支持度为 0.01 时，随着频繁项集项的增加，总耗时的变化情况。-5 分

七．实验额外要求（共 30 分，作业给分上限 100 分）

使用命令行接收参数（数据集、支持率、挖掘频繁项集的大小）-5 分

寻找 2 个关联规则，比较强的关联规则，进行市场分析。-5 分

实现第四部分第三种剪枝策略，并将三种剪枝策略一起使用算法运行时间一起画在算法时间对比图表中。-10 分

实验比较不同算法（Dummy Apriori, Advanced Apriori_1, Advanced Apriori_12, Advanced Apriori_123, FP-Growth）的内存使用情况，并画图分析。-10 分

八．实验要求

独立完成，并在 2021 年 1 月 24 日之前提交实验报告给助教xxx@mail.ustc.edu.cn。提交作业的内容为 zip 压缩包，里面包括：PDF 实验报告、源码，PDF 和压缩包的文件名以及邮箱名格式为 “学号+姓名+第几次实践”。

实验报告包括：实验说明、算法原理与分析、实验结果与分析、完成了那些要求（需要给出每一项的分数以及总分），实验报告控制在 3 页以内，需要展示的频繁项集或者关联规则如果太多展示十行左右即可。