1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：P22提到一种常见的解决方法是将数据分成几个比较小的块，每个子集都包含几个频率相近的项目。对这个分类方法（依据）不是很理解

讨论后的理解： 这个是将不同频率区间的项目分到一个子集里面，而这个频率区间是人为定的，比如0-0.2，0.2-0.4，0.4-0.6，0.6-0.8，0.8-1.0，根据这个频率区间分为5个子集，然后每个子集给定一个最小项目支持度，来解决无意义关联规则和组合爆炸的问题。

1. 提出的问题2：避免频繁项和稀有项出现在同一个项目集的意义是什么？

讨论后的理解：这样可以避免许多无意义的关联规则，并且可以顺利地挖掘稀有项。书中经常出现的诸如{面包，鸡蛋，牛奶，潘}这样的项目集是没有意义的，因为在每个食品商店里，面包、牛奶、鸡蛋等商品都是每个顾客在每次购买时购买的。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：为什么设置最大支持度可以使频繁项目集仍具有向下封闭属性？

自己的理解：其实设置最大支持度也不能完全让频繁项目集保有向下封闭属性，但是可以减小频繁项和稀有项同时出现带来的影响。向下封闭属性应该还是要同一个最小支持度才能保证向下封闭属性。

1. 问题4为什么ms-apriori算法中需要计算f-a的支持计数？

自己的理解：ms-apriori算法在计算关联规则的时候，仅仅记录每个频繁项目集是不够的。有可能一个频繁项目集的子集不是频繁项目集，导致它的支持计数没有保留下来，置信度无法计算。

1. （必填）读书计划

1、本周完成的内容章节：如2.1-2.1

2、下周计划：2.5-2.7

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1、读书摘要及理解（选做）

2.1关联规则的基本概念

关联规则挖掘的目的是找出数据项中所有的并发关系。目标是找出满足事务几何体t中最小支持度（minup）和最小置信度（minconf）的所有关联规则。支持度是指t中包含的事务的百分比。规则支持度的计算如下：

支持=

置信度是指包含X和Y的所有交易中包含X和Y的交易数量的百分比。置信水平计算如下：

信心=

2.2 Apriori算法

Apriori算法是一种针对关联规则的频繁项集挖掘算法。它是挖掘布尔关联规则频繁项集的最有影响力的算法之一。该算法的基本思想是：首先找到所有的频率集，并且这些项集的频率至少与预先定义的最小支持度相同。然后根据频率集生成满足最小支持度和最小置信度的强关联规则。然后，使用步骤1中找到的频率集生成期望的规则，并生成只包含该集合项的所有规则。其中，每条规则的右侧只有一项。这里采用了中介规则的定义。一旦这些规则被生成，只有那些大于用户给出的最小置信度的规则被留下。为了生成所有的频率集，使用递归方法。

2.3关联规则挖掘的数据格式

关联规则挖掘的主要数据类型是事务数据。纯文本文件也可以看作是事务数据。每一份文件都是一笔交易，文字被视为项目。关联规则也可以在关系数据表上实现，只需要将表数据转换为事务数据，

2.4多最小支持度关联规则挖掘

MS-Apriori算法采用了另一种方法。由于统一的支持值不能考虑所有情况，所以可以设置多个支持值。每个类别项都有一个最小支持阈值，然后将频繁项的最小支持值作为项集元素的最小支持值。这样，如果一个稀有项集出现在频繁项集中，则该项集的最小支持值将被拉低。如果另一个项目由高频项目组成，则支持阈值将再次提高。当然，如果出现更困难的异常情况，即在频繁项中出现稀有项和普通项，我们可以通过设置SDC支持差异限制来限制这种情况的发生，从而使挖掘的频繁项更加合理。