西南科技大学

计算机实验报告

课程名称：数据挖掘应用技术与实践

实验名称： 先验算法

学 号：5120180269

学生姓名：李若昊

班 级： 计科1803

指导教师：吴珏

评 分：

实验日期： 2020年9月26日

实验目的：

1. 掌握Apriori算法在关联规则挖掘中频繁项集的产生以及关联规则集合的产生过程中的作用；
2. 根据算法描述编程实现算法，调试运行。并结合相关实验数据进行应用，分析实验结果。

实验内容：

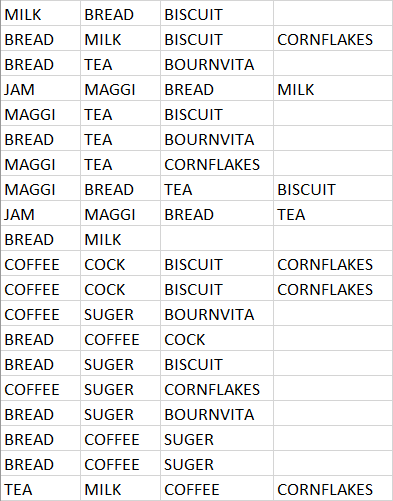
1. 频繁项集的生产与Apriori算法实现；
2. 关联规则的生成过程算法实现；
3. 结合实验数据对算法进行分析；

实验要求：

对于自己选取的数据集，找出数据间隐含的相关性，挖掘关联规则。

实验步骤：

首先介绍我选用的数据集：[来自Kaggle的杂货店数据集](https://www.kaggle.com/shazadudwadia/supermarket)，其包含一个超市的20条交易记录，结构如下：



本次实验中专有名词较多，先加以说明：

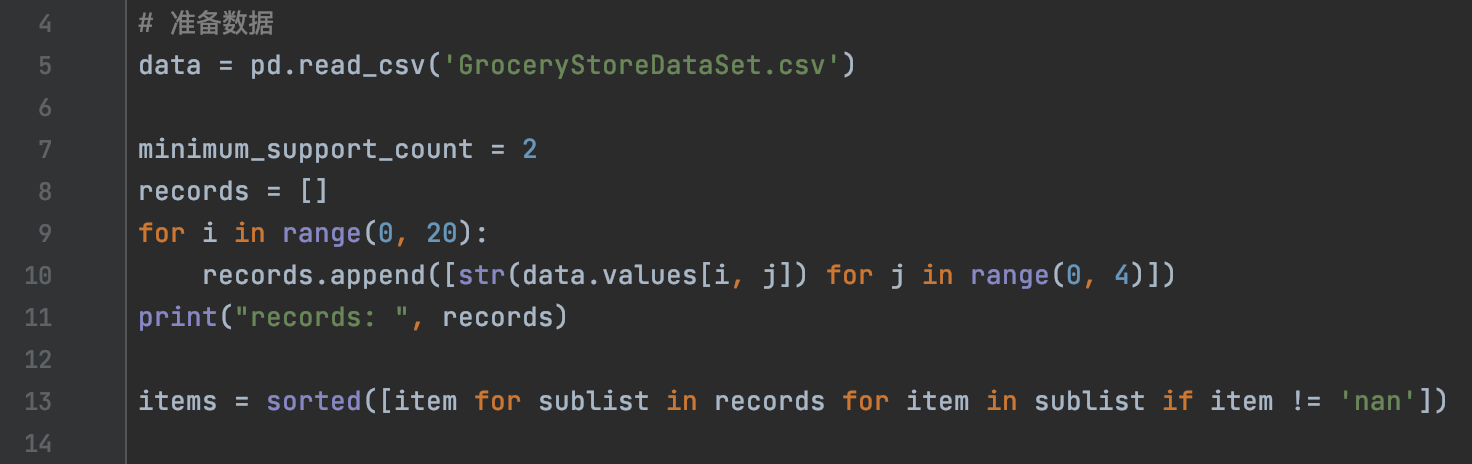
* 项集（Item Set）：本实验中指以商品作为元素的集合，形如{milk}, {milk, bread}
* k-项集：集合中有几个元素即称几项集，如上两例分别为1和2项集
* 支持度计数（Support Count）：对于某一（子）项集，计算该项集隶属的候选集中含有该（子）项集的交易记录的个数（可以理解为统计中的频数），这将作为该项集的一个属性
* 最小支持度计数（Minimum Support Count）：我们认可某项集在整个数据集中出现的最小频数，用它来定义怎样才算“频繁”。比如我们可以指定牛奶的最小购买次数为2，2即为此处的最小支持度计数
* 支持度（Support）：支持度计数对应统计中的频数，而支持度可理解为对应于频率
* 频繁项集L(k)（Frequent Item Set）：项集的支持度计数不小于我们所能接受的最小支持度计数的项集们，可形象地理解为“经常一起被购买的物品都有哪些”
* 候选集C(k)（Candidate Set）：频繁项集的子集们，初始令为整个数据集中的所有1项集
* 关联规则（Associational Rules）：暗示两种物品之间可能存在较强的关系，用A->B表示。这表示如果顾客买了A，那么接下来很可能还会继续买B
* 置信度（Confidence）：对于一条具体的关联规则，如A->B，将它的置信度定义为Support(A,B)/Support(A)，即对应于统计中的条件概率。可理解为“既买了A物品又买了B物品的顾客的占比”

由于最终想要得到的关联规则需要通过频繁项集产生，所以得先算频繁项集有哪些：

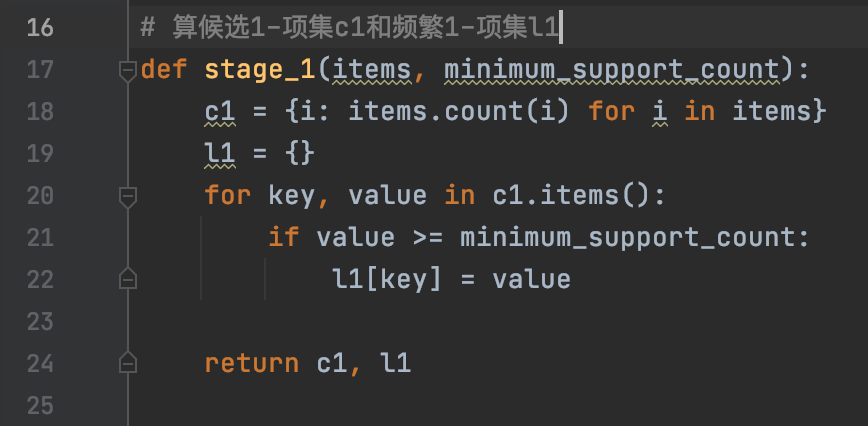
注：

Apriori算法的优点即体现于此处：假设我们需要找到支持度大于2的所有项集，有的时候物品太多，那么这样的话穷举所有的支持度计算量将会很大。为了减少运算量，Apriori原理告诉我们，如果一个项集是频繁项集，那么它的所有子集也是频繁项集；它的逆否命题，即，如果一个项集不是频繁项集，它的所有超集必定不是频繁项集。这样一来，假如已经算出{a,b}并不是频繁项集，那么以后遇到的它的所有超集都没有必要去算它们的支持度了，因为由上述的逆否命题知道，它们必定不会符合我们的要求。如此即可在可接受的时间范围内得到频繁项集都有哪些。

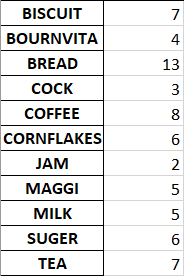
1. 数据准备：首先本次实验约定最小支持度是2，最小置信度是50%。用Pandas库导入上述数据集。



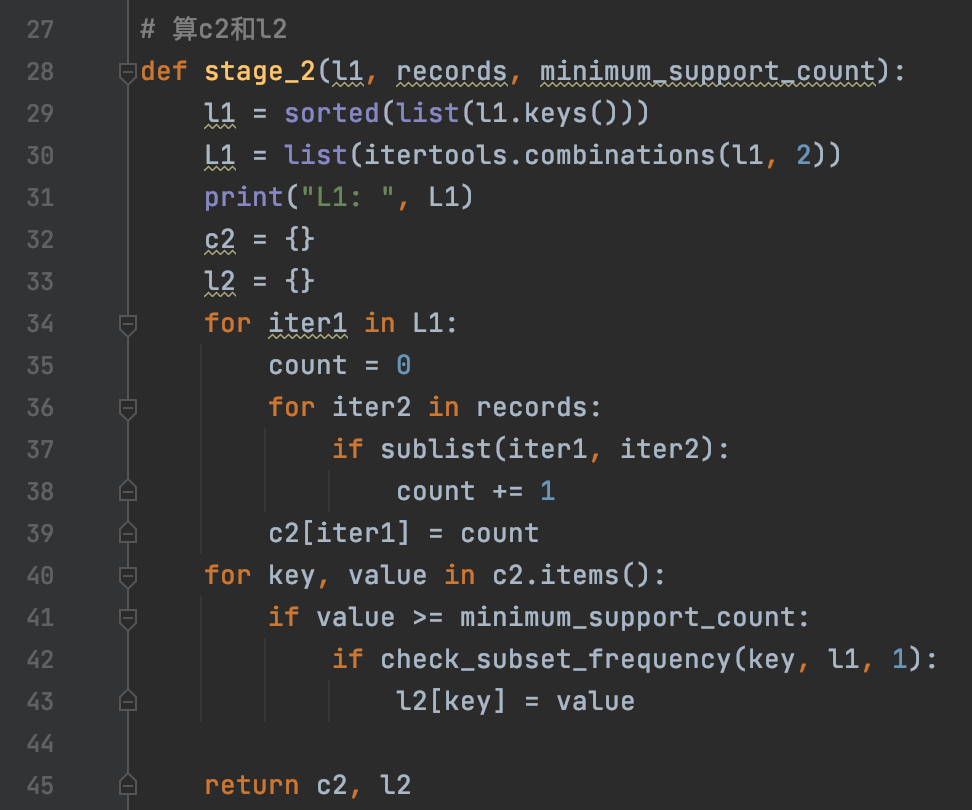
1. 先要算出单个物品购买的频数，其实每个单个物品在这一步中就是所谓的项集，它们组合起来称作候选1项集C (1)。拿刚才得到的这些频数逐个与我们约定的最小值支持度（最低频数）比较，剔除小于最小支持度的项集，剩下的合起来称作频繁1项集L(1)。

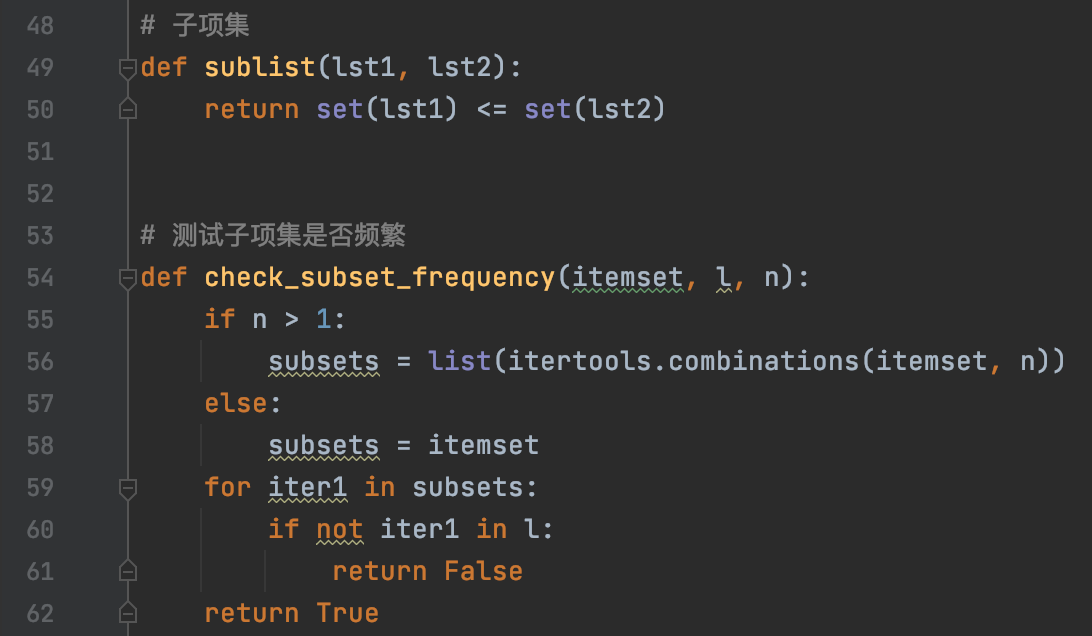


L(1)用直观的方式表示如下：

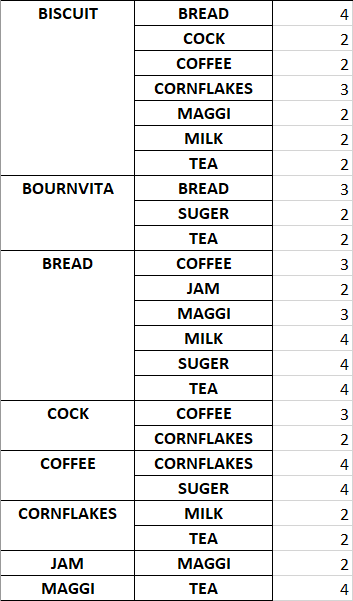


1. 接下来由L(1)产生候选2项集C(2)，并对C(2)计算它里面的各个子项集的频数（支持度），并逐个与预先约定的最小频数（最小支持度）做比较，同样剔除小于最小频数的项集，留下的合起来生成频繁2项集L(2)：

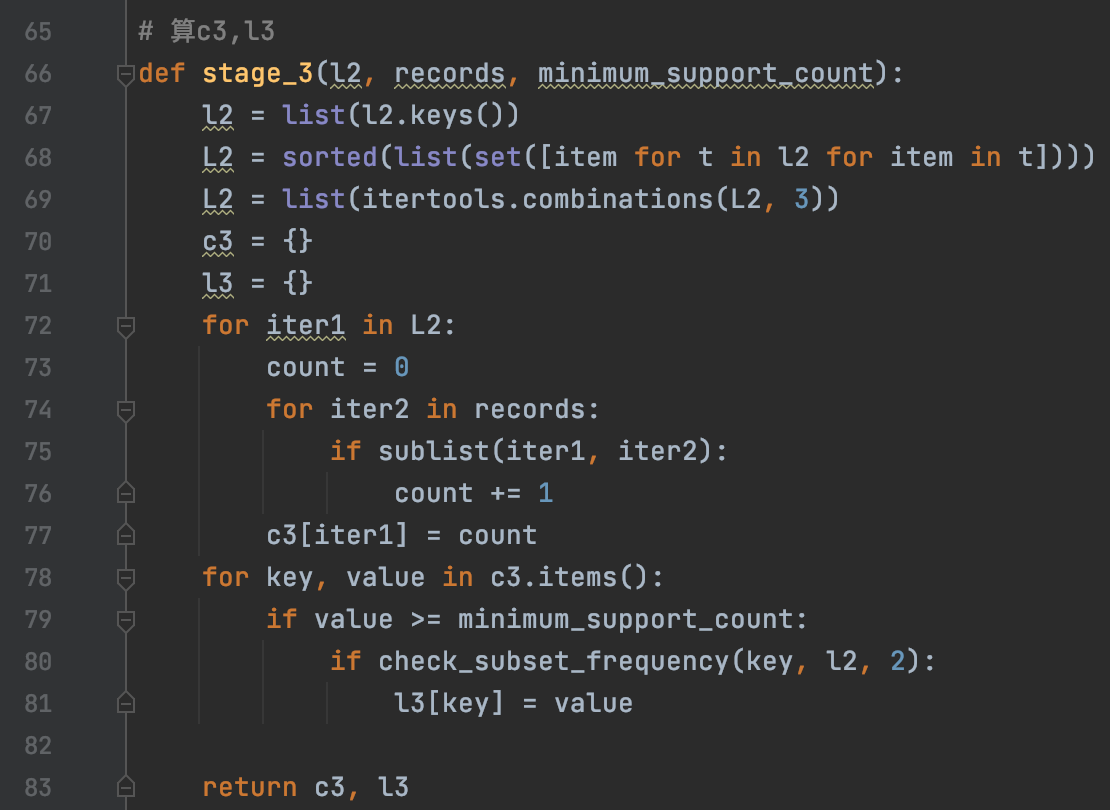


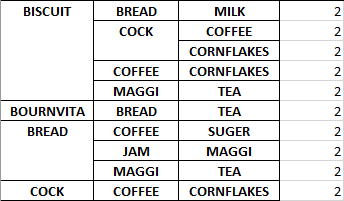


L(2)如下：



1. 再由L(2)产生C(3)，按照同样的手法得出L (3)：





1. 循环往复，直到再也得不到任何一个新的频繁项集时，算法终止。

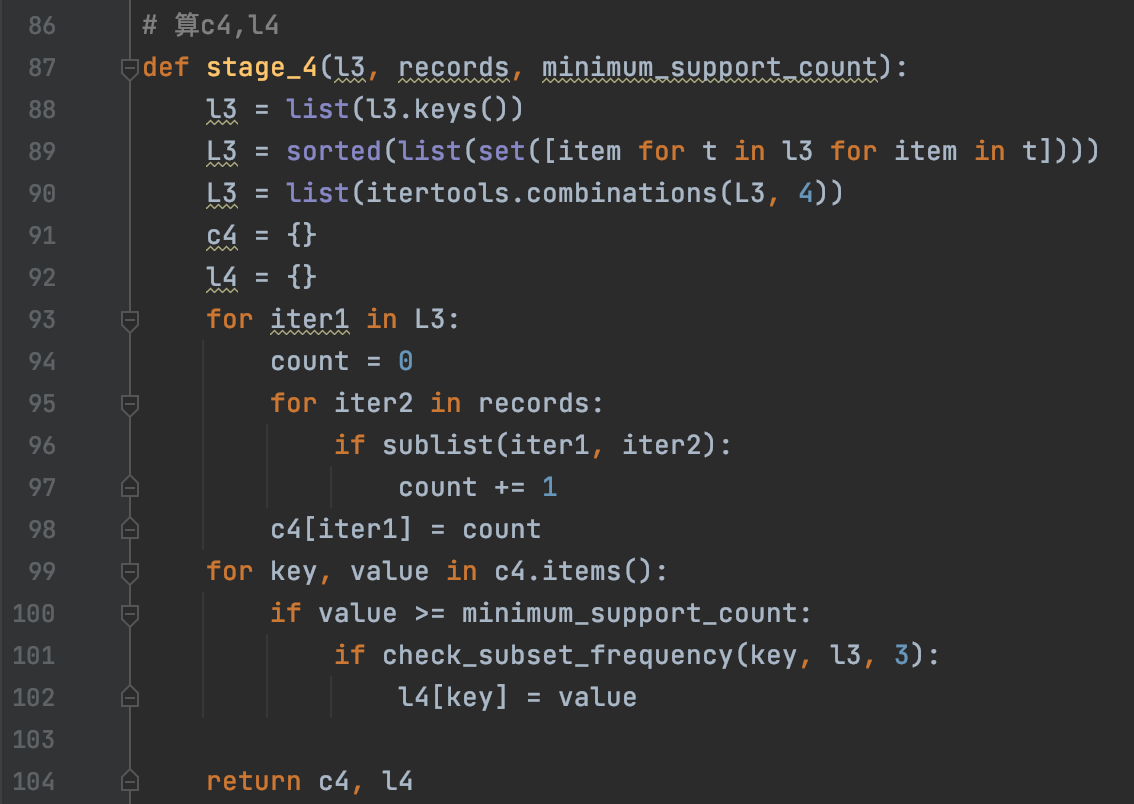
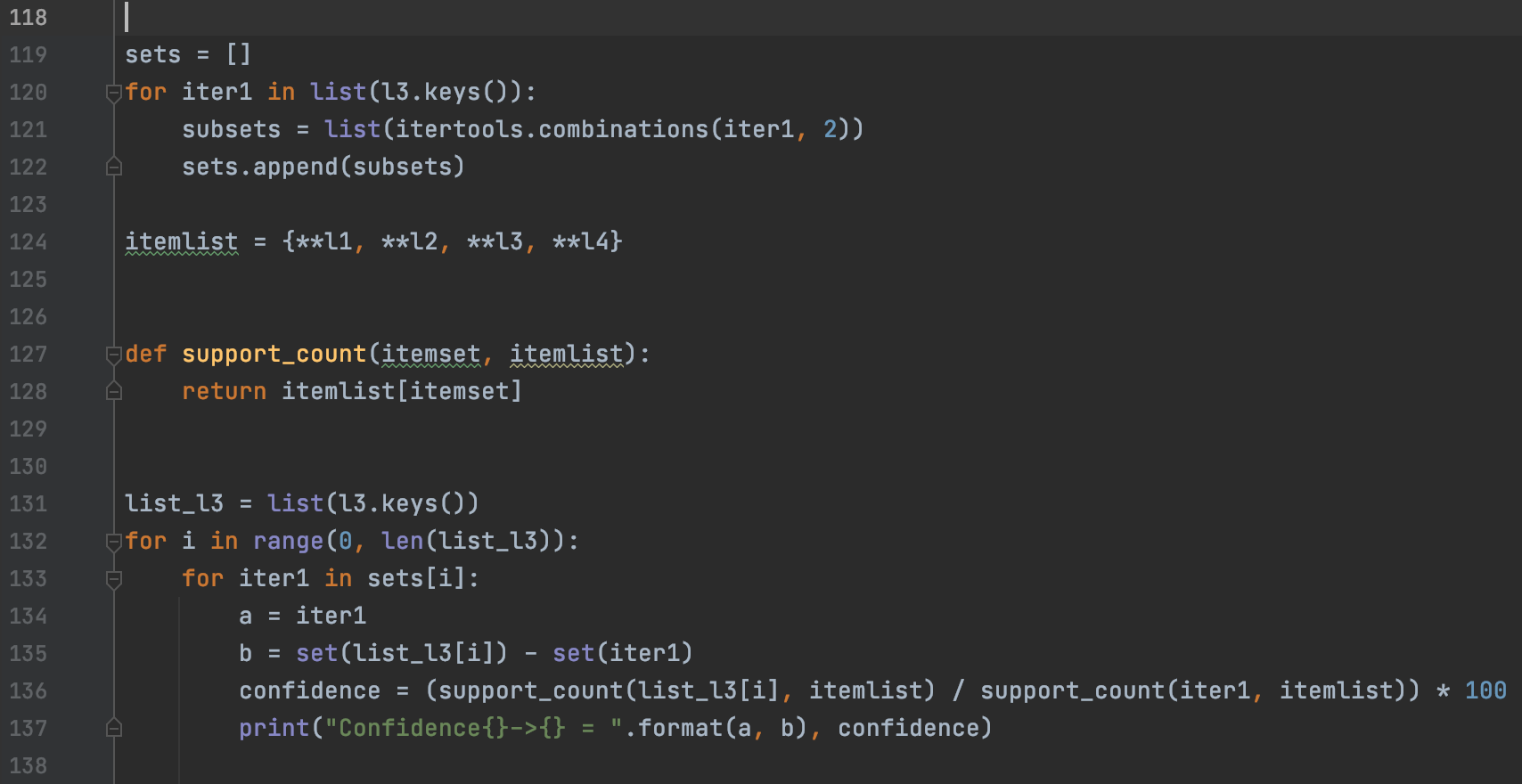
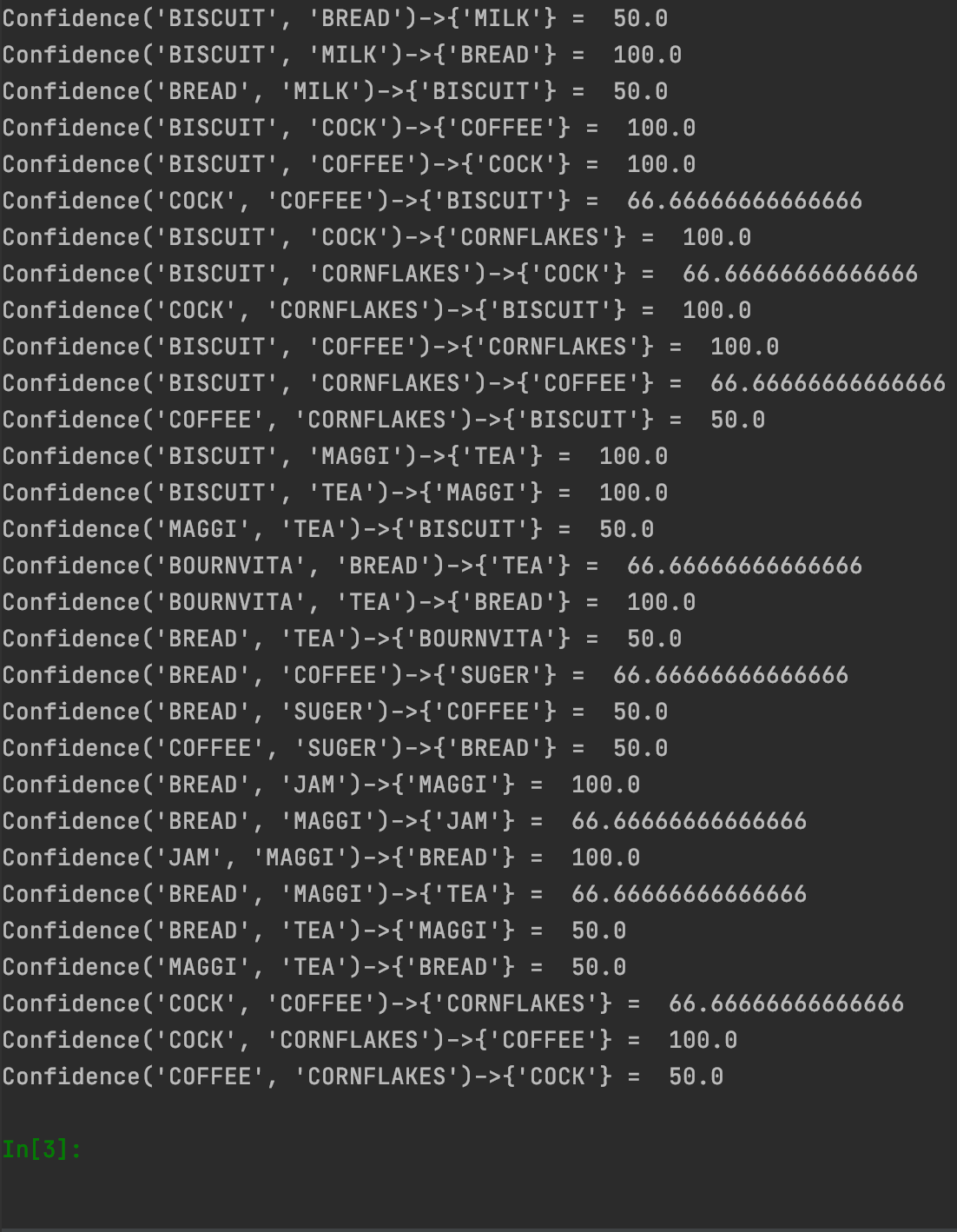


Image for post

1. 现在有了频繁项集，可以利用它们找关联规则了。方法是算出每条规则的置信度，选出不小于我们预设的最小置信度的规则们。如对L3中的商品们，有：





实验思考：

1. 集合的表示以及相关操作的实现方法。

利用字典，见实验步骤1。

1. 项集的数据结构描述。

用列表，见实验步骤1。

实验体会：

在上学期的机器学习课程中已经利用一些工具直接使用过该算法，现在我知道了算法背后的原理，可以帮助我在应用时更好地理解结果和参数的意义。

感谢吴老师的悉心指导。