

**课 程 实 验 报 告**

**课程名称： 大数据分析**

**专业班级： CS1806**

**学 号： U201814670**

**姓 名： 李田田**

**指导教师： 崔金华**

**报告日期： 2020.12.17**

**计算机科学与技术学院**

**目录**

[实验三 关系挖掘实验 1](#_Toc58398474)

**[3.1 实验内容](#_Toc58398475)** [1](#_Toc58398475)

**[3.2 实验过程](#_Toc58398476)** [1](#_Toc58398476)

[3.2.1 编程思路 1](#_Toc58398477)

[3.2.2 遇到的问题及解决方式 2](#_Toc58398478)

[3.2.3 实验测试与结果分析 2](#_Toc58398479)

**[3.3 实验总结](#_Toc58398480)** [3](#_Toc58398480)

# 实验三 关系挖掘实验

## **3.1 实验内容**

1. 实验内容

编程实现Apriori算法，要求使用给定的数据文件进行实验，获得频繁项集以及关联规则。

2. 实验要求

以Groceries.csv作为输入文件

输出1~3阶频繁项集与关联规则，各个频繁项的支持度，各个规则的置信度，各阶频繁项集的数量以及关联规则的总数

固定参数以方便检查，频繁项集的最小支持度为0.005，关联规则的最小置信度为0.5

## **3.2 实验过程**

### 3.2.1 编程思路

Apriori算法：

每一个项都是集合C1的成员。扫描所有的事件，获得每个项出现的次数，生成C1。然后对每个项进行计数。然后根据最小支持度从C1中删除不满足的项，从而获得频繁1项集L1。

对L1的自身连接生成的集合执行剪枝策略产生候选2项集的集合C2，然后，扫描所有事务，对C2中每个项进行计数。同样的，根据最小支持度从C2中删除不满足的项，从而获得频繁2项集L2。

对L2的自身连接生成的集合执行剪枝策略产生候选3项集的集合C3，然后，扫描所有事务，对C3每个项进行计数。同样的，根据最小支持度从C3中删除不满足的项，从而获得频繁3项集L3。

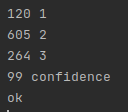
### 3.2.2 遇到的问题及解决方式

一开始用元组tuple实现该算法过程，后来发现元组实现在查找的过程中很慢，便将tuple换成set类型，以便于查找和使用是否为子集的函数 issubset()。

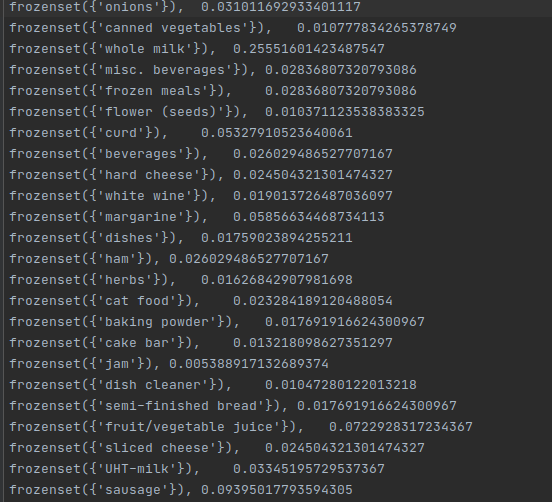
刚开始用元组实现的时候非常麻烦，需要经常在tuple和set、list类型进行转换，而且对于一个元素tuple类型会出现逗号，输出不是很标准。

### 3.2.3 实验测试与结果分析

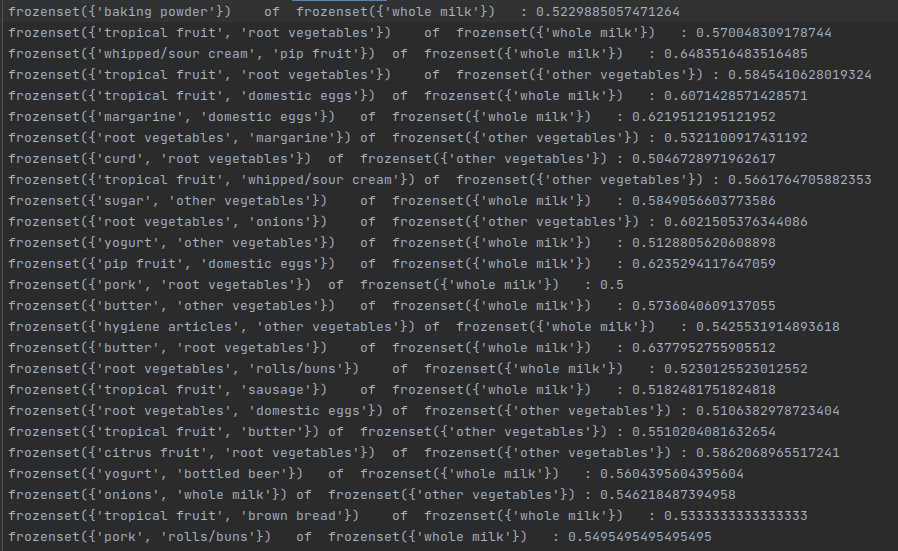
各阶频繁项集的数量以及关联规则的总数：



频繁项的支持度：



规则的置信度：



## **3.3 实验总结**

通过此次实验，最大的体会便是对于python语言的数据结构的体会，如list：有序集合、可重复、随时增删；tuple：有序列表、初始化之后无法修改；dict：查找速度快；set：无序集合，不可重复、便于查找，可直接进行集合操作。

此外，对于apriori算法，啤酒和婴儿用品的例子和源头更验证了大数据源于生活。很多大数据的算法，与其说是寻找解决办法，倒不如说是寻找生活中各种各样的规律，再运用于生活。