# 作业二 关联规则挖掘

**姓名：李晓波**

**学号：2620170054**

### 一、实验环境及语言

**语言及环境依赖**

语言： python

依赖的包：numpy,pandas, chain, combinations

### 二、实现要求

1. 对数据集进行处理，转换成适合关联规则挖掘的形式；
2. 找出频繁项集；
3. 导出关联规则，计算其支持度和置信度
4. 对规则进行评价，可使用Lift，也可以使用教材中所提及的其它指标

### 三、实现方法

#### 采用数据集Building\_Permits

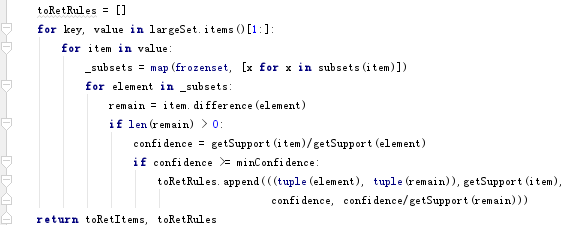
#### 对数据集进行处理的源程序

首先对数据集进行处理，因为数据较大，选出部分属性：：['Permit Type', 'Permit Type Definition', 'Plansets', 'TIDF Compliance', 'Existing Construction Type', 'Existing Construction Type Description', 'Proposed Construction Type', 'Proposed Construction Type Description']。

为方便查看，将属性简化为由各单词首字母串联而成，则变成：['PT', 'PTD', 'P', 'TC', 'ECT', 'ECTD', 'PCT', 'PCTD']等属性。为了关联挖掘，将源数据集中的各个属性映射拼接成为该元素的取值。处理程序preprocess.py，生成较处理的程序predata.csv

#### 找出频繁项集

实现Apriori算法，可得到1到K项的频繁项集,其中Aprior算法中计算支持度和置信度：



所有的频繁项集保存在frequent\_items.txt中

#### 导出关联规则，计算其支持度和置信度

运行结果一部分为

rule support confidence lift

('PCT\_5.0',) ==> ('PCTD\_wood frame (5)',) 0.575 1.000 1.739

('PCTD\_wood frame (5)',) ==> ('PCT\_5.0',) 0.575 1.000 1.739

('PTD\_otc alterations permit',) ==> ('PT\_8',) 0.899 1.000 1.112

('PT\_8',) ==> ('PTD\_otc alterations permit',) 0.899 1.000 1.112

('ECT\_5.0',) ==> ('ECTD\_wood frame (5)',) 0.570 1.000 1.755

('ECTD\_wood frame (5)',) ==> ('ECT\_5.0',) 0.570 1.000 1.755

('ECT\_5.0', 'PCTD\_wood frame (5)') ==> ('ECTD\_wood frame (5)',) 0.563 1.000 1.755

('ECTD\_wood frame (5)', 'PCTD\_wood frame (5)') ==> ('ECT\_5.0',) 0.563 1.000 1.755

('P\_2.0', 'ECT\_5.0') ==> ('ECTD\_wood frame (5)',) 0.282 1.000 1.755

('P\_2.0', 'ECTD\_wood frame (5)') ==> ('ECT\_5.0',) 0.282 1.000 1.755

所有关联规则保存在association\_rules\_by\_confidence.txt文件中。

#### List评价

所有文件保存在association\_rules\_by\_lift中