数据挖掘互评作业2：频繁模式与关联规则挖掘

姓名：曹健 学号：3120190978

1. 数据
2. 数据集选择：**Oakland Crime Statistics 2011 to 2016**
3. 数据集描述：

包含records-for-2011.csv到records-for-2016.csv共六个文件

属性列表：

Agency，Create Time，Location，Area Id，Beat，Priority，Incident Type Id，Incident Type，Description，Event Number，Closed Time

1. 问题

* 对数据集进行处理，转换成适合进行关联规则挖掘的形式；
* 找出频繁模式；
* 导出关联规则，计算其支持度和置信度;
* 对规则进行评价，可使用Lift、卡方和其它教材中提及的指标, 至少2种；
* 对挖掘结果进行分析；
* 可视化展示。

三、关联规则挖掘

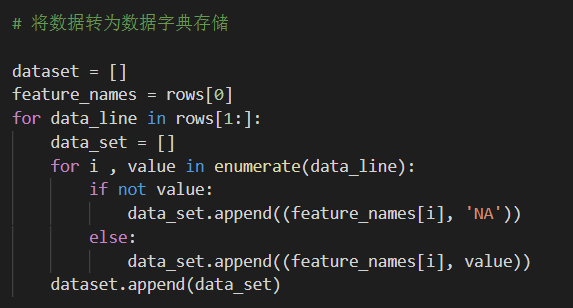
1. 数据集处理

将数据集处理成适合关联规则挖掘的格式，对Agency、Location、Area Id、Beat、Priority、Incident Type Id、Incident Type Description、Event Number这些属性进行关联规则挖掘。

将不同属性值转化为可以产生频繁项集的形式：



将每一个属性名与属性值的组合表示为一个元组的形式：（属性名，属性值）作为一个单项，并使用python中的frozenset类型表示项集。



2. 产生频繁项集

使用Apriori算法在经过预处理的数据集上构建频繁项集。关联规则的强度主要可以使用两个指标来衡量：支持度和置信度。

首先为这两个超参数设定阙值，设定最小支持度为0.01（将支持度设得偏小是因为数据集中有几个属性的取值较为分散，如果支持度太高则会忽略掉这些属性），将最小置信度设为0.5。

Apriori算法使用频繁项集的先验知识，使用一种称作逐层搜索的迭代方法，k项集用于探索(k+1)项集（如果事件A中包含k个元素，那么称这个事件A为k项集，并且事件A满足最小支持度阈值的事件称为频繁k项集）。首先，通过扫描事务（交易）记录，找出所有的频繁1项集，该集合记做L1，然后利用L1找频繁2项集的集合L2，L2找L3，如此下去，直到不能再找到任何频繁k项集。

Apriori算法流程如下：

* 扫描一次数据库D；计算出各个1项集的支持度，得到频繁1项集的集合。
* 从2项集开始循环，进行由频繁k-1项集生成频繁频繁k项集。
  + 连接步：将2个只有一个项不同的属于的频集做一个（k-2）JOIN运算得到。
  + 剪枝步：因为是超集，所以可能有些元素不是频繁的。舍弃掉子集不是频繁项集即不在频繁k-1项集中的项集
  + 扫描数据库，计算2.3步中过滤后的k项集的支持度，舍弃掉支持度小于阈值的项集，生成频繁k项集。
* 当生成的频繁k项集中只有一个项集时循环结束

**Apriori 算法实现代码细节见associasion.py文件**

3. 导出关联规则

使用上面基于Apriori算法产生的频繁项集，进行强关联规则的挖掘。过程如下：

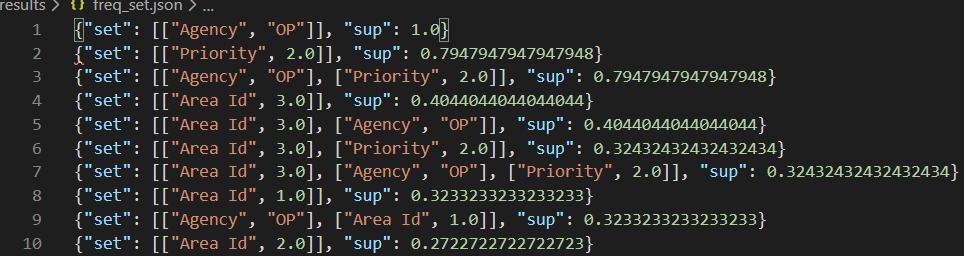
* 根据选定的频繁项集，找到它所有的非空子集。然后建立规则列表。
* 强关联规则需要满足最小支持度和最小置信度，对每一条规则计算指标。
* 根据以上指标，找到所有可能的关联规则

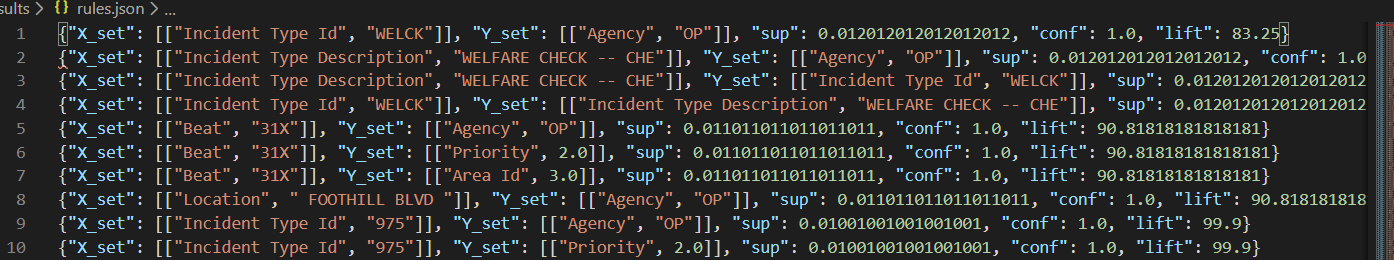
**实现代码细节见associasion.py文件**

4. 结果分析

**产生的频繁项集结果见 results/freq\_set.json 文件**

**产生的关联规则结果见 results/rules.json 文件**

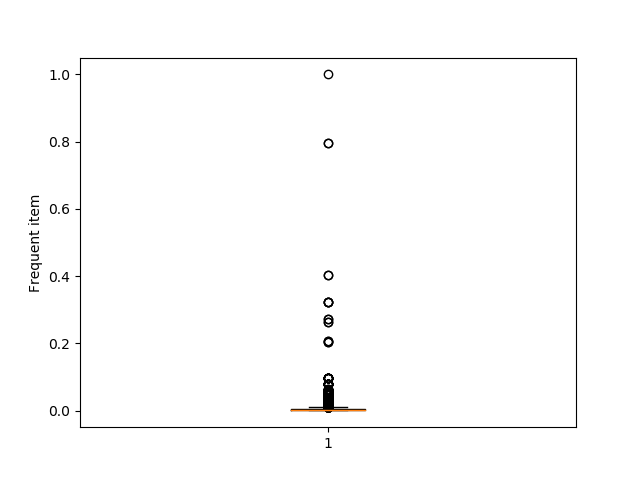




上图中是对应的前十条结果。

5. 结果可视化

对频繁项集结果做可视化：



对关联规则结果做可视化：

