**关联规则挖掘 美国国会投票记录**

姓名:吕思勤 学号:19030130448

1. **实验内容**
2. 数据来源

http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

1. 使用Apriori算法，支持度设为30%，置信度为90%，挖掘高置信度的规则

**2、分析及设计**

1. 数据预处理:

依据数据的格式和内容，对数据进行整合、提取等操作在编程处理时尤其还要注意格式的转换

1. 频繁项集的产生:

1) 使用循环遍历当前数据集确定每一个单独的项对应的支持度值 根据支持度剪枝确定频繁1-项集

2) 采用Fk-1\*Fk-1的项集产生策略，对于频繁k-1项集生成当前频繁k项集

基于支持度和置信度阈值进行剪枝确定最终的频繁k项集

3) 重复上述第二个步骤，直到不再有新的频繁项集产生

1. 关联规则挖掘:
2. 规则产生性质

对于从同一个频繁项集产生的规则，置信度评估在规则箭头右边的项数目上具有反单调性，因此从频繁k+1项集中生成规则时，优先考虑在频繁k项集中且是频繁 k+1项集的子集元素

1. 规则合并方式

在规则左边取交集;规则左右两边的并集不发生改变

1. 程序实现

对于频繁项集的列表，循环遍历每个k+1项集，在频繁k项集中找是其子集的元素，计算二者的置信度，进行剪枝

1. 结果输出:

将结果输出并对其进行分析

**3、详细实现**

1.加载原始数据

一張含有 文字 的圖片

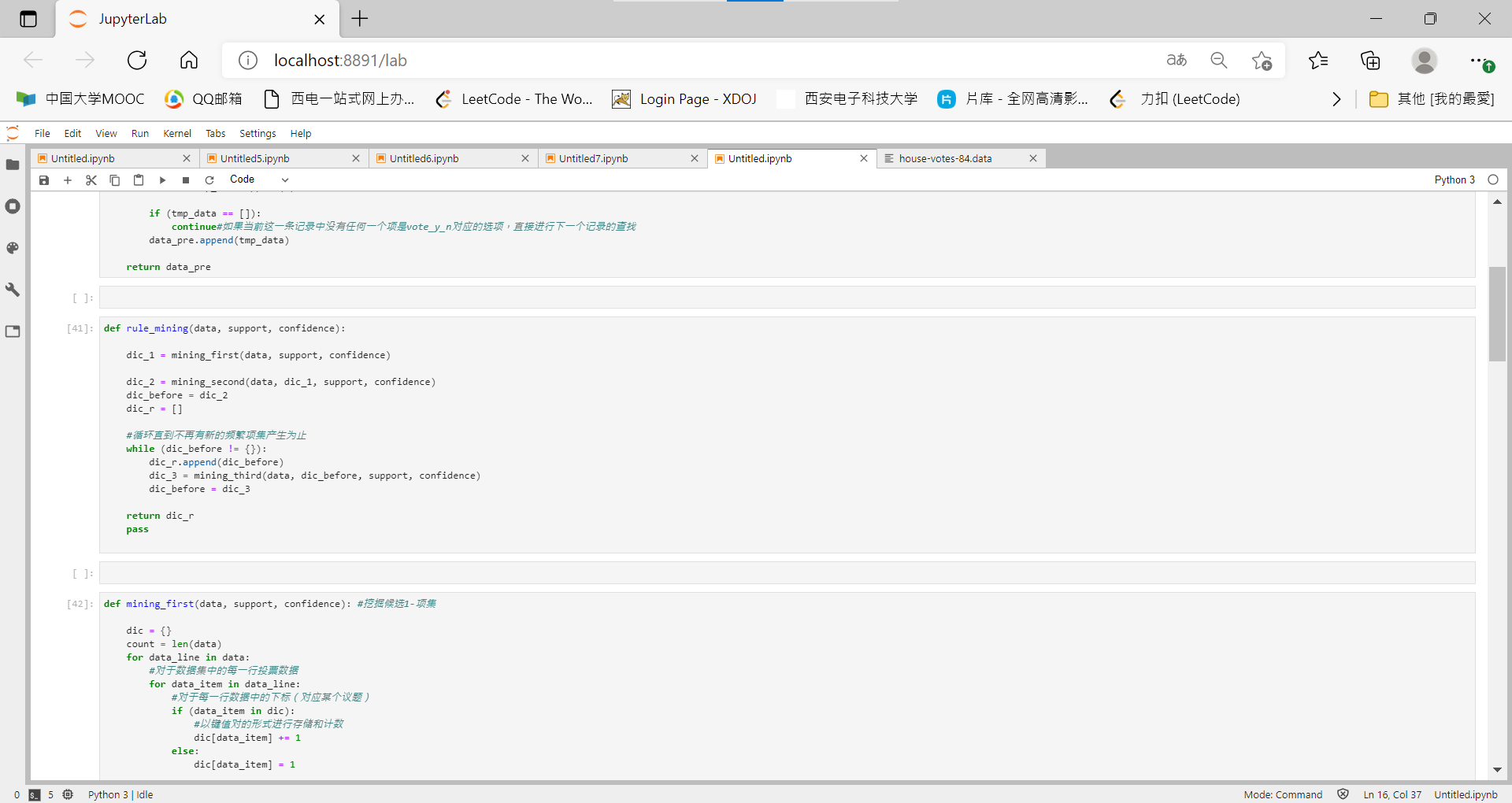
自動產生的描述

2.数据预处理 以按 y 或 n 寻找关联规则

一張含有 文字 的圖片

自動產生的描述

1. 挖掘关联规则



#第一次挖掘

一張含有 文字 的圖片

自動產生的描述

#第二次挖掘

一張含有 文字 的圖片

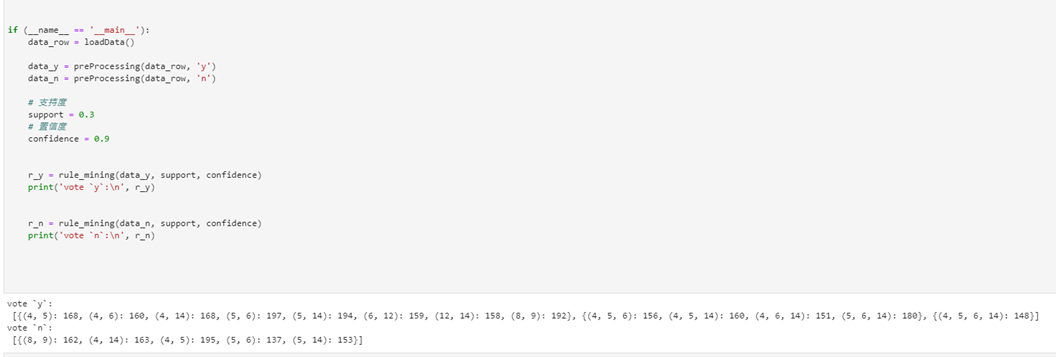
自動產生的描述

#第三次挖掘

一張含有 文字 的圖片

自動產生的描述

4.主函数



**4、实验结果**

一張含有 文字 的圖片

自動產生的描述

根据结果可以挖掘出高置信度的规则，将2-项集、3-项集和4-项集等显示输出，包括关联的项和频数。比如vote `y`中第一项规则表示，给决议4投了赞成票的人，往往也会给决议5投赞成票，其他同理。

**5、心得体会**

通过本次实验，我学习了python中函数的定义方式与调用技巧以及使用Apriori算法，挖掘高置信度的规则的编程实现方法，让我更加深入的了解了Apriori关联分析算法中的剪枝规则和反单调性性质，Apriori关联分析算法实现起来看似复杂，但是它的原理却是非常经典且简单粗暴。

除此之外我深刻的体会到关联规则分析的强大功能，在大规模数据集中使用简单的算法就能够寻找出有趣的关联规则。完成这次的实验后，我了解到数据挖掘知识的应用领域，也感受到这门学科有趣的一面。