**实验过程报告**

3220190975-宋迎新

# 数据集

[Trending YouTube Video Statistics](https://www.kaggle.com/datasnaek/youtube-new)

选择下列标签进行预处理：category\_id、comments\_disabled、ratings\_disabled、video\_error\_or\_removed， 发现有部分数据存在缺失值的情况。去掉缺失值的行后，由原来的359876条数据找出频繁模式。

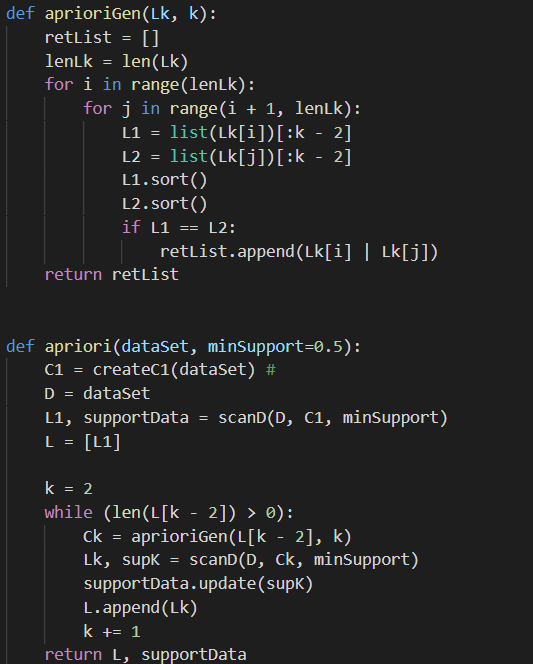
# 算法介绍

Apriori算法是一种挖掘关联规则的频繁项集算法，其核心思想是通过候选集生成和情节的向下封闭检测两个阶段来挖掘频繁项集。而且算法已经被广泛的应用到商业、网络安全等各个领域。 算法简介 Apriori算法1是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

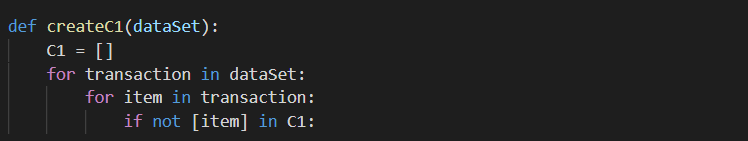
实验过程中频繁项阈值为15%，置信度的阈值为60%

相应代码为：

Apriori主函数：



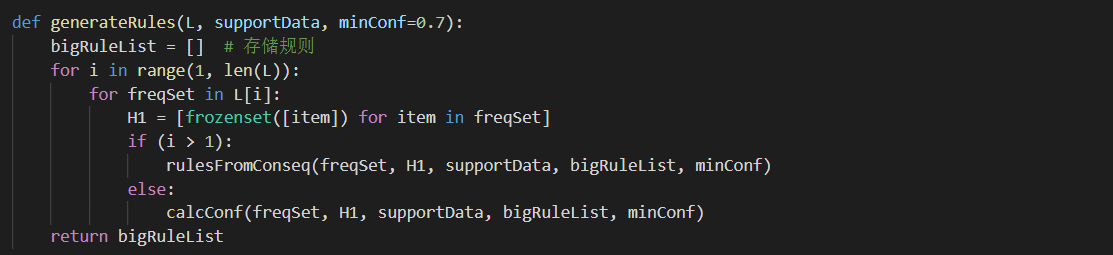
候选集生成：



过滤低支持度函数：



合并筛选函数：



保存结果为json格式，

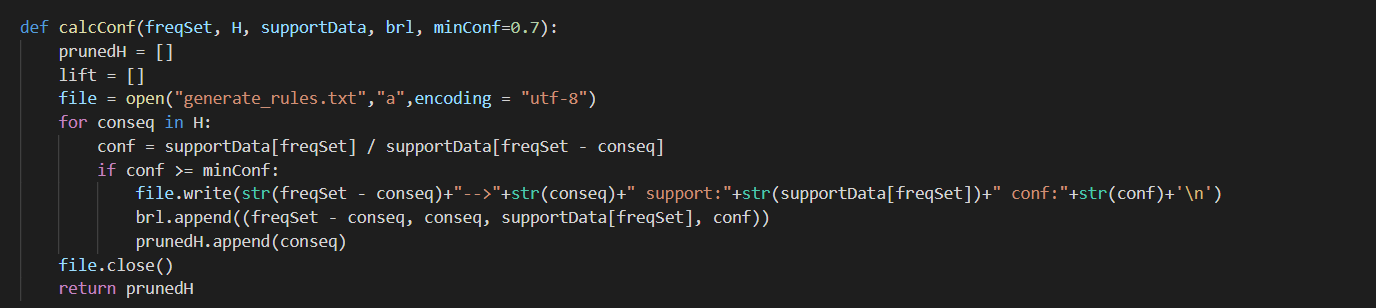
# 导出关联规则

评价指标为Lift和Jaccard。

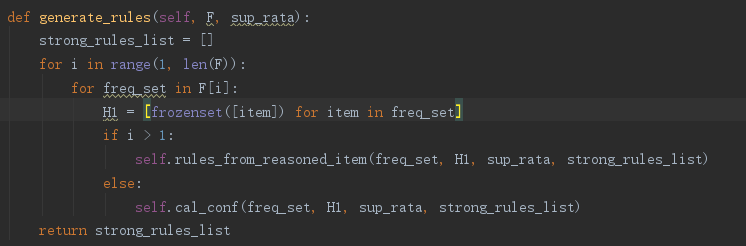
Lift:

Jaccard:

以上计算对应的代码为：

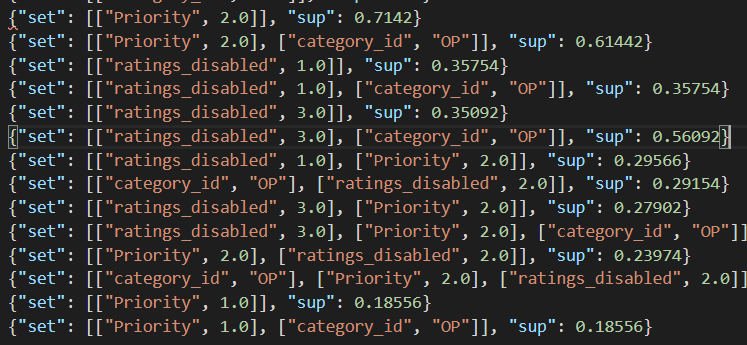


计算并保留达到置信度阈值的函数为：

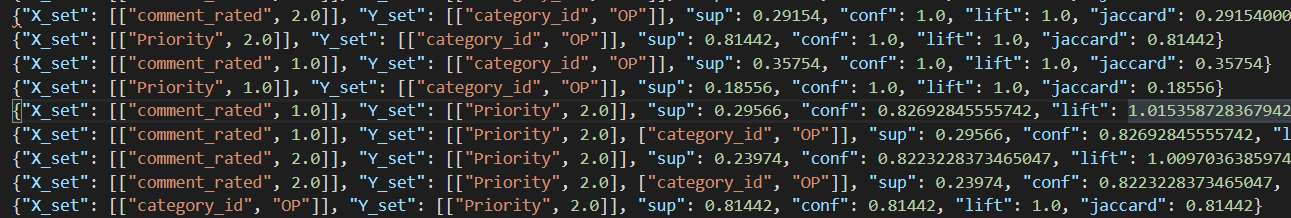


# 分析挖掘结果

频繁项集frequency.json频繁项集，排序由大到小



关联规则rules.json文件中，按置信度由大到小排列



由频繁项集.json得知，而且category\_id和ratings\_disabled的关联度较高。由rules.json得知，["category\_id", 0.8]与["ratings\_disabled", 1.5]的置信度较高，由此可以初步判断出category\_id和ratings\_disabled有较强的相关性。

# 可视化

（1）对频繁项集使用盒图可视化和 对规则的指出度与置信度使用散点图可视化

