DSCI6001P 数据科学基础

作业4

Problem 1

1. 考虑下表的购物篮事务：

事务 ID	购买项
1	{牛奶，啤酒，尿布}
2	{面包，黄油，牛奶}
3	{牛奶，尿布，饼干}
4	{面包，黄油，啤酒}
5	{啤酒，饼干，尿布}
6	{牛奶，尿布，面包，黄油}
7	{面包，黄油，尿布}
8	{啤酒，尿布}
9	{牛奶，尿布，面包，啤酒}
10	{啤酒，饼干}

问题：

(1) 从这些数据中，能够提取出的关联规则的最大数量是多少（包括零支持度的规则）？

(2) 能够提取的频繁项集的最大长度是多少（假定最小支持度>0）？

(3) 写出从该数据及中能够提取的 3-项集的最大数量的表达式。

(4) 找出具有最大支持度的项集（长度为 2 或更大）。

答：

问题（1）：

分析可知，该数据集有6个项，即牛奶、啤酒、尿布、面包、黄油、饼干。对这6个项进行排列组合。对于这个关联规则的求解首先提出如下引理。

引理：假设某个数据集包含n个项，那么从该数据集提取的可能规则的总数是$3^n - 2^{n+1} + 1$。

>下面对于该引理进行证明，可以将总情况视为先选出k个数，其中$k \geq 2$。然

后，再将k个数分成两个部分（两个部分的数目都不小于2）。这样可以得到总情况数如公式（1）所示。

$$\sum_{k=2}^{n}\binom{n}{k}\sum_{i=1}^{k-1}\binom{k}{i}=\sum_{k=2}^{n}\binom{n}{k}(2^k-2)=3^n-2^{n+1}+1 \tag{1}$$

故该数据集能够提取的关联规则的最大数为：$3^6-2^{6+1}+1=602$。

问题（2）：

由于项集{牛奶，尿布，面包，黄油}和{牛奶，尿布，面包，啤酒}长度最大且支持度均为$\frac{1}{10}>0$。因此能够提取的频繁项集的最大长度为4。

问题（3）：

相当于从6个项之中选择3个，因此能够提取出的3-项集数目为$\binom{6}{3}=20$

问题（4）：

{面包，黄油}的支持度最大为$\frac{4}{10}=0.4$。/font>

Problem 2

2. 数据库有 5 个事务。设 $min_sup=60\%,\ min_conf=80\%$。

TID	购买的商品
T100	|M, O, N, K, E, Y|
T200	|D, O, N, K, E, Y|
T300	|M, A, K, E|
T400	|M, U, C, K, Y|
T500	|C, O, O, K, I, E|

问题：

(1) 分别使用$Apriori$算法和$FP-growth$算法找出频繁项集。比较两种挖掘过程的有效性。

(2) 列举所有与下面的元规则匹配的强关联规则(给出支持度 s 和置信度 c)，其中，X 是代表顾客的变量，$itemi$是表示项的变量(如“A”，“B”等)：

$$\forall x\in transaction, buys(X,item_1)\wedge buys(X,item_2)\Rightarrow buys(X,item_3)[S,C] \tag{2}$$

答：

问题（a）：

Apriori算法：首先商品的一项集及其支持度计数。

项集	支持度计数
{M}	3

项集	支持度计数
{O}	3
{N}	2
{K}	5
{E}	4
{Y}	3
{D}	1
{A}	1
{U}	1
{C}	2
{I}	1

基于上表过滤到低于阈值$5 \times 0.6 = 3$的项集。因此剩下的1-项集有{M}、{O}、{K}、{E}、{Y}。根据频繁1项集生成候选集C_2并计数。

项集	支持度计数
{M, O}	1
{M, K}	3
{M, E}	2
{M, Y}	2
{O, K}	3
{O, E}	3
{O, Y}	2
{K, E}	3
{K, Y}	3
{E, Y}	2

基于上表过滤到低于阈值$5 \times 0.6 = 3$的项集。因此剩下的2-项集有{M, K}、{O, K}、{O, E}、{K, E}、{K, Y}。根据频繁2项集生成候选集C_3并计数。

项集	支持度计数
{M, O, K}	1
{M, E, K}	2
{M, Y, K}	2

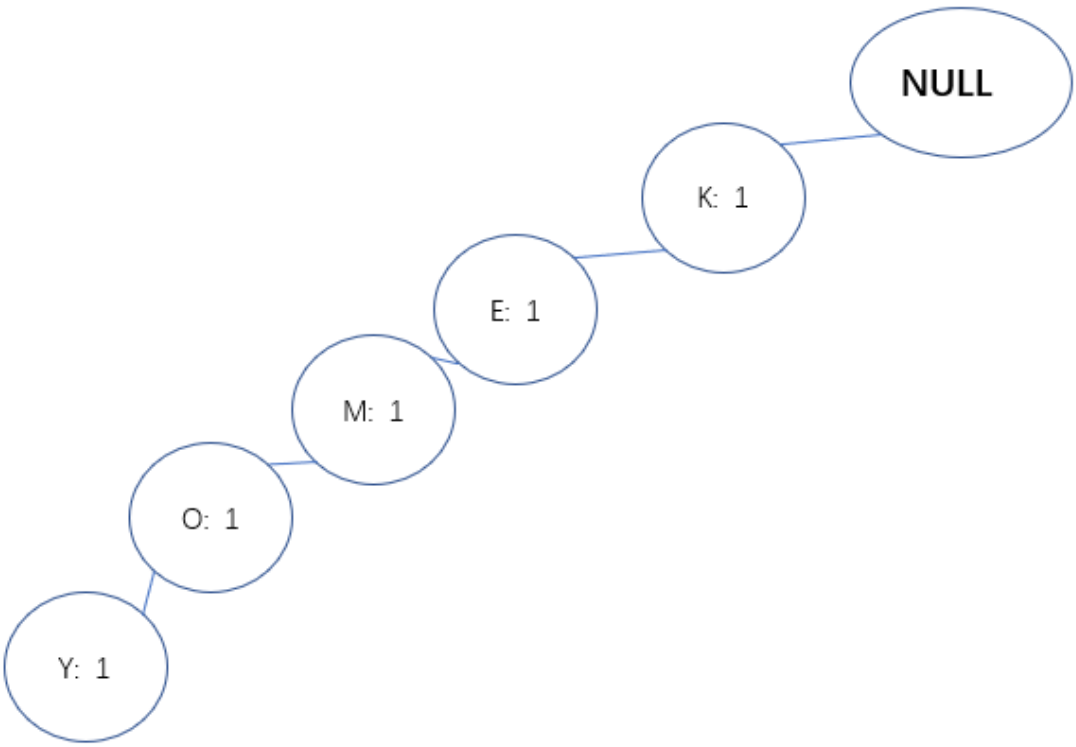
项集	支持度计数
{O, K, E}	3
{O, K, Y}	2
{K, E, Y}	2

基于上表过滤到低于阈值$5 \times 0.6 = 3$的项集。因此剩下的3-项集有{O, K, E}。由于L_3中只有一个项集，不再生成候选集，$Apriori$算法结束。

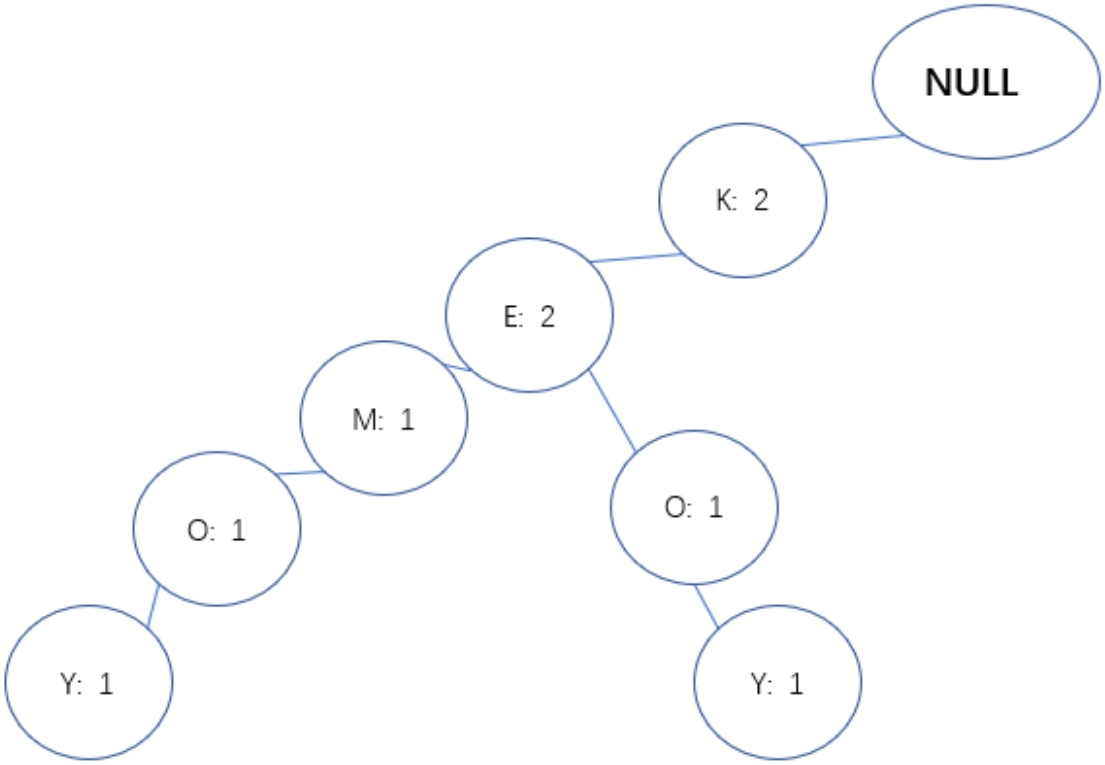
FP-growth算法：同样生1-项集并按照降序排列选择出频繁1-项集。

项集	支持度计数
{K}	5
{E}	4
{M}	3
{O}	3
{Y}	3

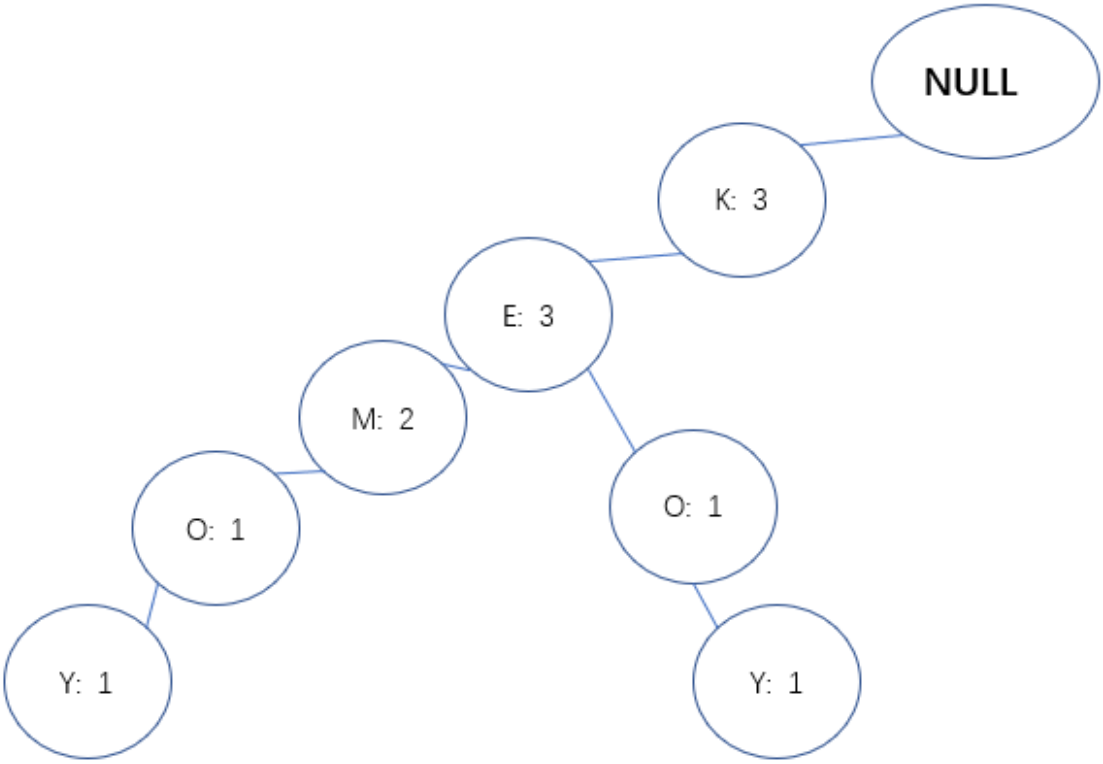
根据L_1提取出$T100$中的特征项为：{M}，{O}，{K}，{E}，{Y}。按照降序排列得到{K, E, M, O, Y}，然后开始创建$FP-tree$。



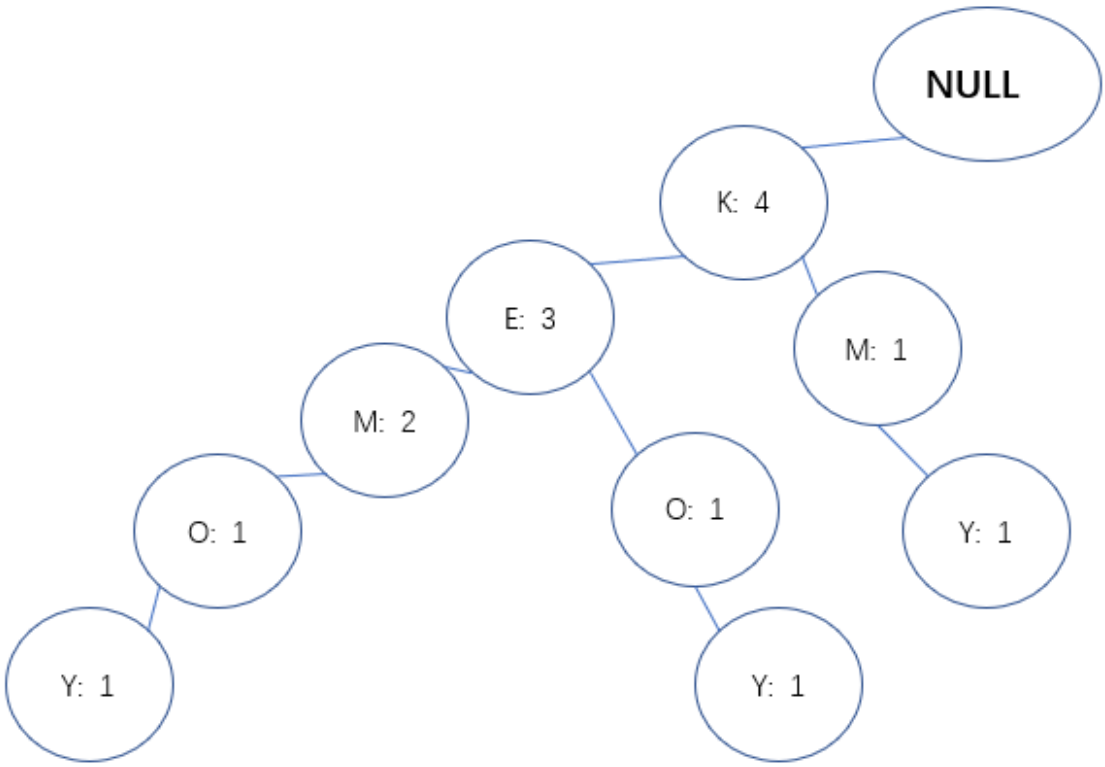
对于$T200$的数据获取降序排列的频繁项集为：{K, E, O, Y}。它与$T100$共享前缀{K, E}。因此更新共享结点支持度。对于{O, Y}重新创建树的节点。



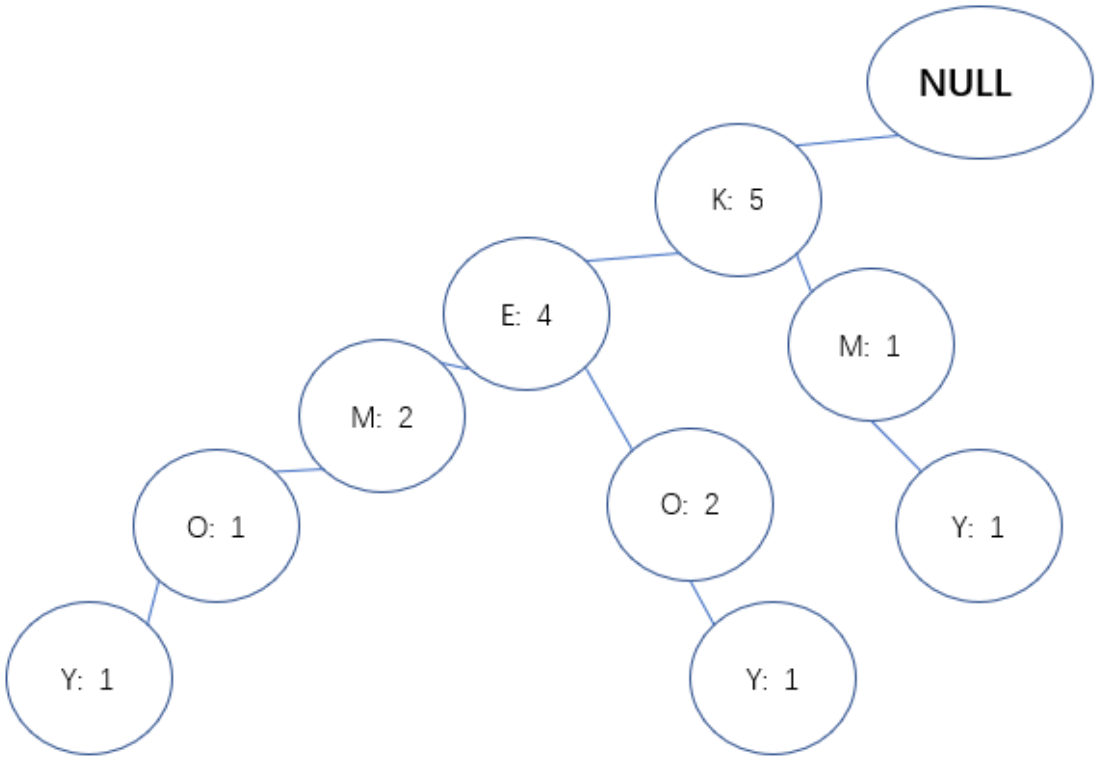
对于$T300$的数据获取降序排列的频繁项集为：{K, E, M}。构建树如下。



对于$T400$的数据获取降序排列的频繁项集为：{K, M, Y}。构建树如下。



对于$T500$的数据获取降序排列的频繁项集为：{K, E, O}。构建树如下。



通过构建完成的树获得每个条件项的CPB。

条件项	条件模式库（PCB）
K	{}
E	{K: 4}

条件项	条件模式库（PCB）
M	{KE: 2, K: 1}
O	{KEM: 1, KE: 2}
Y	{KEMO: 1, KEO: 1, KM: 1}

然后对于每个条件项构建条件频繁模式树。最后产生的频繁项集。

条件项	产生的频繁项集
K	{}
E	{K, E}
M	{K, M}
O	{O, E}, {K, E, O}, {O, K}
Y	{K, Y}

分析两个算法的过程很容易发现，$Apriori$ 算法是基于逐层搜索迭代方式。算法会产生大量的候选集。空间和时间开销较大，没有考虑各个属性之间的重要性。而$FP-growth$算法通过$FP-Tree$存储了用于挖掘的相关信息，利于分治的思想进行递归挖掘，需要遍历数据集2次。 $FP-Growth$算法不产生无用的候选项集。但是需要利用树这种数据结构进行存储，对于大量子节点的树会造成其效率下降。

问题（b）：

对于该问题需要找出满足：

$C(item_1 \bigwedge item_2 \Rightarrow item_3)=\frac{S(item_1,item_2,item_3)}{S(item_1,item_2)}>0.8$。发现：

$$\frac{S(K,E,O)}{S(K,E)}=\frac{0.6}{0.6}=1>0.8 \tag{3}$$

$$\frac{S(K,E,O)}{S(K,O)}=\frac{0.6}{0.6}=1>0.8 \tag{4}$$

$$\frac{S(K,E,O)}{S(E,O)}=\frac{0.6}{0.6}=1>0.8 \tag{5}$$

因此，强关联规则有3条。$K \bigwedge E \Rightarrow O$，$K \bigwedge O \Rightarrow E$，$O \bigwedge E \Rightarrow K$。

Problem 3

3. 设计一种方法，对无限的数据流进行有效的朴素贝叶斯分类（即只能扫描数据流一次）。如果想发现这种分类模式的演变（例如，将当前的分类模式与较早的模式进行比较，如与一周以前的模式相比），你有何修改建议？
答：

由于只能扫描数据流一次，我们可以在内存（或者储存在硬盘上）之中使用一个属性值计数表进行更新。因为朴素贝叶斯分类的计算公式是基于先验知识来进行预测。只需要通过这个属性值计数表不断更新先验知识即可完成无线数据流的朴素贝叶斯分类。

```
Input Stream Data:
    for steaming input x:
        add it to the Bayes
        train and get the model
        update model
```

如果像发现这种分类模式的演变，算法只需要设置固定的周期就好。例如，我们将上周的模式先存储下来保存其计算过程。对于这周的计算结果，用这周的数据进行替换然后再得到新的计算结果和计算过程。对于下周依次类推，这样就可以分析出分类模式的演变规律。但是这个方法缺点是需要保存之前的分类器，会造成额外的存储开销。

```
Input Stream Data:
    for steaming input x:
        add it to the Bayes
        train and get the model
        update model
        if time % 7 == 0: #以7天为周期
            update weeklymodel as model
```

Problem 4

4. 假设一个布隆过滤器的容量为 8×10^9位，集合中有 1×10^9个元素。如果使用 3 个哈希函数，试计算误判率。如果使用 4 个哈希函数呢？
答：

根据误判率公式：

$$f = (1 - e^{-\frac{kn}{m}})^k \tag{6}$$

如果使用 3 个哈希函数：

$$f_3 = (1 - e^{-\frac{3\times1\times10^9}{8\times10^9}})^3 \approx 0.25 \tag{7}$$

如果使用 4 个哈希函数：

$$f_3 = (1 - e^{-\frac{3\times1\times10^9}{8\times10^9}})^4 \approx 0.16 \tag{8}$$

Problem 5

5. 假定全集 A 有 n 个元素，随机从中抽取出两个子集 A_1和 A_2，且每个子集都有 m 个元素，求 A_1和A_2两个集合的期望相似度。
答：

不妨假设选出两个子集中只有t个元素相同。这个事件发生的概率为：

$$P(\text{只有}t\text{个元素相同}) = \frac{\binom{n}{t}\binom{n-t}{m-t}\binom{n-m}{m-t}}{\binom{n}{m}\binom{n}{m}} \tag{9}$$

故可得相似度期望：

$$E = \sum_{t=0}^{m} \frac{t}{2m-t} \frac{\binom{n}{t}\binom{n-t}{m-t}\binom{n-m}{m-t}}{\binom{n}{m}\binom{n}{m}} \tag{10}$$

Problem 6

6. 给定输入流$< b, a, c, a, d, e, a, f, a, d >$，计数器个数 $k = 3$。请逐步写出 $Misra - Gries$ 算法执行的结果。
答：

输入	操作	结果
b	插入	F = {(b, 1)}
a	插入	F = {(b, 1), (a, 1)}
c	插入 删除	F = {(b, 1), (a, 1), (c, 1)} F = {}
a	插入	F = {(a, 1)}
d	插入	F = {(a, 1), (d, 1)}
e	插入 删除	F = {(a, 1), (d, 1), (e, 1)} F = {}

输入	操作	结果
a	插入	F = {(a, 1)}
f	插入	F = {(a, 1), (f, 1)}
a	更新	F = {(a, 2), (f, 1)}
d	插入 删除	F = {(a, 2), (f, 1), (d, 1)} F = {(a, 1)}

因此，频繁元素为a。

Problem 7

7. 给定数据流$< 4,1,3,5,1,3,2,6,7,0,9 >$，若哈希函数形如$h(x) = (ax + b)\ mod\ 8$，其中 a 和 b 是任意给定的常数。假设给定如下哈希函数：

(1) $h(x) = (3x + 2)\ mod\ 8$；

(2) $h(x) = (7x + 5)\ mod\ 8$；

(3) $h(x) = (5x + 3)\ mod\ 8$。

请利用 $Count - Min\ sketch$ 算法估计频繁项。
答：

$Hash$值

数据	4	1	3	5	1	3	2	6	7	0	9
h_1	6	5	3	1	5	3	0	4	7	2	5
h_2	1	4	2	0	4	2	3	7	6	5	4
h_3	7	0	2	4	0	2	5	1	6	3	0

Algorithm 5.5: CM sketch 算法

输入: 数据流，查询元素 a
输出: 元素 a 出现的频数

1 **初始化**：$C[1\cdots d][1\cdots w]\leftarrow 0, w=\frac{2}{\epsilon}, d=\lceil log(1/\delta)\rceil$; 选择$d$个独立的哈希函数$h_1, h_2, \cdots, h_d : [n]\rightarrow[w]$
2 **处理**：(j, c)，其中$c=1$;
3 **for** i = *1 to* d **do**
4 $\quad C[i][h_i(j)]\leftarrow C[i][h_i(j)]+c$;
5 **return** $\widehat{f}_a = min_{1\le i\le d} C[i][h_i(a)]$;

按照算法构建出C表。

深度d\宽度w	0	1	2	3	4	5	6	7
h_1	1	1	1	2	1	3	1	1
h_2	1	1	2	1	3	1	1	1
h_3	3	1	2	1	1	1	1	1

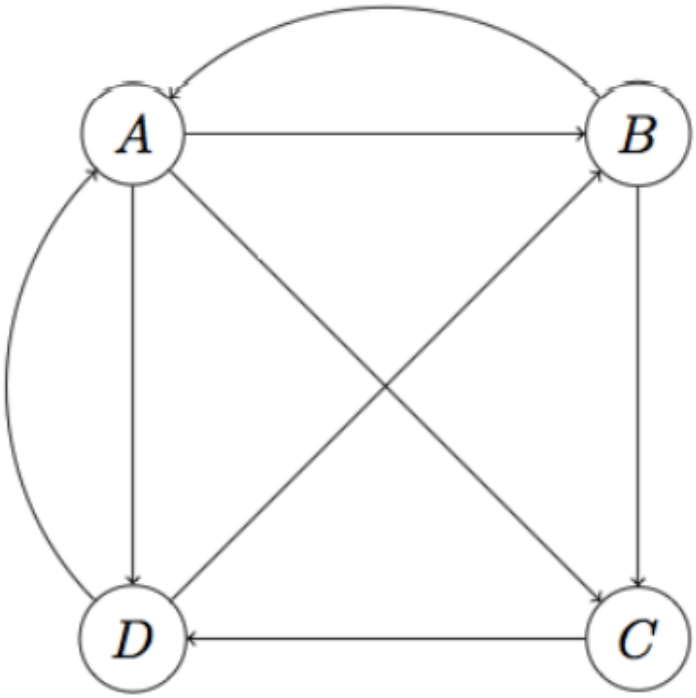
从C表可以看出频繁项对应的$hash$应该为$(5,4,0)$。因为对于h_1、h_2与h_3而言，$(5,4,0)$分别出现次数最多。而$(5,4,0)$对应的数据是1和9，因此频繁项为1、9。

Problem 8

8. 在不考虑 $dampingfactor$ 的情况下，对于连通图 $PageRank$ 的定义如下：

> 设 u 是有向图 G 中的一个顶点，$N(u)$ 表示图 G 中指向顶点 u 的顶点集合。则图顶点 u 的 PageRank 值可以计算为
>
> $$PR(u)=\sum_{v\in N(u)}\frac{PR(v)}{N(v)}, \tag{7.30}$$
>
> 其中 $PR(u)$ 表示顶点 u 的 PageRank 值。

若存在 A、B、C、D 四个网站，其链接结构如下图所示，计算其$PageRank$值（有向图，注意箭头）。



答：

基于*python*编程实现。

```
import numpy as np
#构建有向图矩阵
G = np.array([[0, 1 / 3, 1 / 3, 1 / 3], [1 / 2, 0, 1 / 2, 0], [0, 0, 0, 1], [1 / 2, 1/ 2, 0, 0]])
G = G.T
#初始话pr
pr0 = np.array([0, 0, 0, 0])
pr1 = np.array([0.25, 0.25, 0.25, 0.25])
#迭代至收敛
while(np.sum((pr1 - pr0) ** 2) > 1e-6):
        pr0 = pr1.copy()
        pr1 = G.dot(pr0)
        print(pr1)
print(pr1)
```

```
#迭代收敛过程
[0.25       0.20833333 0.20833333 0.33333333]
[0.27083333 0.25       0.1875     0.29166667]
[0.27083333 0.23611111 0.21527778 0.27777778]
[0.25694444 0.22916667 0.20833333 0.30555556]
[0.26736111 0.23842593 0.20023148 0.29398148]
[0.2662037  0.23611111 0.20833333 0.28935185]
[0.26273148 0.23341049 0.20679012 0.2970679 ]
[0.2652392  0.23611111 0.20428241 0.29436728]
[0.2652392  0.23559671 0.20646862 0.29269547]
[0.26414609 0.2347608  0.20621142 0.29488169]
[0.26482124 0.23548954 0.2054291  0.29426012]
[0.26487483 0.23540381 0.20601852 0.29370285]
```

因此A、B、C、D的$PageRank$ 值分别为：
0.26487483、0.23540381、0.20601852、0.29370285。

Problem 9

9. 给定一个转移矩阵 P 和状态向量π:

$$P=\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{4} & \frac{3}{4}\end{pmatrix} \text{和}\pi=\begin{pmatrix}1\\0\end{pmatrix}^T$$

a. 计算 πP, πP^2和πP^3 。

b. 证明 πP^n 的结果接近一个常数向量。

c. 给定任一个转移矩阵 P， 满足矩阵中每个元素为概率值，且每一行的元素和为 1。证明 P 和$\frac{1}{n}((n-1)I+P)$ 有相同的平稳分布，其中 I 表示单位矩阵。

答：

问题a.

$$\pi P=\begin{pmatrix}1\\0\end{pmatrix}^T\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{1}{4} & \frac{3}{4}\end{pmatrix}=\begin{pmatrix}\frac{1}{2} & \frac{1}{2}\end{pmatrix} \tag{11}$$

$$\pi P^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{3}{8} & \frac{5}{8} \end{pmatrix} \tag{12}$$

$$\pi P^3 = \begin{pmatrix} \frac{3}{8} & \frac{5}{8} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{9}{32} & \frac{21}{32} \end{pmatrix} \tag{13}$$

问题b.

很容易求解得到矩阵P的特征值$\lambda = \frac{1}{4}$、1。对应的特征向量为：$\begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}$、$\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$。对矩阵$P$进行特征分解很容易得到：

$$P = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^{-1} \tag{14}$$

显然有：

$$\begin{aligned} \lim_{n \to +\infty} P^n &= \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{4}^n & 0 \\ 0 & 1^n \end{pmatrix} \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^{-1} \end{aligned} \tag{15}$$

因此，πP^n 的结果接近一个常数向量。

问题c.

平稳分布即存在π使得$\pi = \pi P_i$。若对于不同i，π相同则说明具有相同的平稳分布。 因此：

$$\begin{aligned} & \pi = \pi \frac{1}{n}((n-1)I + P) \\ \iff & \pi = \frac{(n-1) \times \pi}{n} + \frac{\pi \times P}{n} \\ \iff & \frac{\pi}{n} = \frac{\pi \times P}{n} \\ \iff & \pi = \pi P \end{aligned} \tag{16}$$

因此，P 和$\frac{1}{n}((n-1)I + P)$ 有相同的平稳分布。