第1题

(a) 基于α,β

三台计算机的向量分别为：

- A：[3.06, 500α, 6β]
- B：[2.68, 320α, 4β]
- C：[2.92, 640α, 6β]

则sim(A,B)：

$$sim(A,B)=\frac{3.06\times2.68+500\alpha\times320\alpha+6\beta\times4\beta}{\sqrt{3.06^2+(500\alpha)^2+(6\beta)^2}\cdot\sqrt{2.68^2+(320\alpha)^2+(4\beta)^2}}$$

其余 sim(A,C)、sim(B,C) 同理。

(b) α=0.01,β=0.5

三台计算机的向量分别为：

- A：[3.06, 5,3]
- B：[2.68, 3.2, 2]
- C：[2.92, 6.4 3]

计算余弦相似度：

$$sim(A,B)=\frac{3.06\times2.68+5\times3.2+3\times2}{\sqrt{3.06^2+5^2+3^2}\cdot\sqrt{2.68^2+3.2^2+2^2}}\approx\frac{28.9408}{\sqrt{43.3636}\times\sqrt{22.5424}}\approx0.926$$

$$sim(A,C)\approx0.961\quad sim(B,C)\approx0.910$$

第2题

效用矩阵：

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

(a) Jaccard 距离

转为布尔值（有评分为1，无为0）：

- A: [1 1 0 1 1 0 1 1]
- B: [0 1 1 1 1 1 1 0]
- C: [1 0 1 1 0 1 1 1]
- A ∩ B = 4, A ∪ B = 8 ⇒ 距离 = 1 - 4/8 = 0.5
- A ∩ C = 4, A ∪ C = 8 ⇒ 距离 = 1 - 4/8 = 0.5

- B ∩ C = 4, B ∪ C = 8 ⇒ 距离 = 1 - 4/8 = 0.5

(b) 布尔余弦距离

用布尔向量计算余弦相似度，再取 1-相似度为距离：

- A·B = 4, |A|=√6, |B|=√6 ⇒ sim ≈ 0.667 ⇒ dist ≈ 0.333
- A·C = 4, |A|=√6, |C|=√6 ⇒ sim ≈ 0.667 ⇒ dist ≈ 0.333
- B·C = 4, |B|=√6, |C|=√6 ⇒ sim ≈ 0.667 ⇒ dist ≈ 0.333

(c) 评分3-5为1，1-2为0，其余为0

- A: [1 1 0 1 0 0 1 0]
- B: [0 1 1 1 0 0 0 0]
- C: [0 0 0 1 0 1 1 1]

得出余弦距离：

- A·B=2, |A|=√4, |B|=√3 ⇒ sim ≈ 0.577 ⇒ dist ≈0.423
- A·C=2, |A|=√4, |C|=√4 ⇒ sim = 0.5⇒ dist = 0.5
- B·C=1, |B|=√3, |C|=√4 ⇒ sim ≈ 0.289 ⇒ dist ≈ 0.711

(d) 评分减去各用户均值

用户均值：

- A: (4+5+5+1+3+2)/6 = 3.33
- B: (3+4+3+1+2+1)/6 = 2.33
- C: (2+1+3+4+5+3)/6 = 3.0

归一化：

	a	b	c	d	e	f	g	h
A	0.67	1.67		1.67	-2.33		-0.33	-1.33
B		0.67	1.67	0.67	-1.33	-0.33	-1.33	
C	-1.0		-2.0	0.0		1.0	2.0	0.0

计算余弦距离（找出共同评分项）：

用户对	归一化余弦相似度	归一化余弦距离
A-B	0.823	**0.177**
A-C	-0.263	**1.263**
B-C	-0.933	**1.933**

第3题

原始矩阵：

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix}$$

(a)

$$MM^T = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3 & 5 & 0 & 1 \\ 2 & 4 & 4 & 2 & 3 \\ 3 & 5 & 3 & 4 & 5 \end{bmatrix} = \begin{bmatrix} 14 & 26 & 22 & 16 & 22 \\ 26 & 50 & 46 & 28 & 40 \\ 22 & 46 & 50 & 20 & 32 \\ 16 & 28 & 20 & 20 & 26 \\ 22 & 40 & 32 & 26 & 35 \end{bmatrix}$$

$$M^TM = \begin{bmatrix} 1 & 3 & 5 & 0 & 1 \\ 2 & 4 & 4 & 2 & 3 \\ 3 & 5 & 3 & 4 & 5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 36 & 37 & 38 \\ 37 & 49 & 61 \\ 38 & 61 & 84 \end{bmatrix}$$

(b)

解如下特征多项式：

$$\det(M^TM - \lambda I) = 0$$

$$\lambda 1 \approx 153.57,\ \lambda 2 \approx 15.43,\ \lambda 3 \approx 0$$

(c)

这些特征向量已归一化，对应于上面三个特征值（列向量）：

$$V = \begin{bmatrix} -0.409 & -0.816 & 0.408 \\ -0.563 & -0.126 & -0.816 \\ -0.718 & 0.564 & 0.408 \end{bmatrix}$$

(d)

SVD 形式为：

$$M = U\Sigma V^T$$

- **奇异值 Σ（对角线）为特征值平方根：**

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} = \begin{bmatrix} 12.392 & 0 & 0 \\ 0 & 3.928 & 0 \\ 0 & 0 & \approx 0 \end{bmatrix}$$

- **矩阵 V**（即V^T 的转置）：

$$V = \begin{bmatrix} -0.409 & -0.563 & -0.718 \\ -0.816 & -0.126 & 0.564 \\ 0.408 & -0.816 & 0.408 \end{bmatrix}$$

- **矩阵 U（截取前三列）**：

$$U = \begin{bmatrix} -0.298 & 0.159 & 0.941 \\ -0.571 & -0.033 & -0.153 \\ -0.521 & -0.736 & -0.052 \\ -0.323 & 0.510 & -0.196 \\ -0.459 & 0.414 & -0.224 \end{bmatrix}$$

第4题

布隆过滤器参数：m = 80亿 bit，n = 10亿

$$f = (1 - e^{-kn/m})^k$$

令k=3：

$$f \approx (1 - e^{-3\times10^9/8\times10^9})^3 \approx (1 - e^{-0.375})^3 \approx (1 - 0.687)^3 \approx 0.313^3 \approx 0.031$$

令k=4 同理，结果 ≈ 0.017

第5题

(a) h(x)=2x+1 mod 32

流：3,1,4,1,5,9,2,6,5

哈希结果：

- 3 → 7 → 111 → 尾长0
- 1 → 3 → 011 → 尾长0
- ...尾长分别为：[0,0,0,0,0,0,0,0,0]

最大尾长=0 ⇒ 估计数目：2^0/φ ≈ 1/0.773 ≈ 1.29

(b) h(x)=4x mod 32

- 3 → 12 → 1100 → 尾长2
- 1 → 4 → 0100 → 尾长2
- ...尾长分别为：[2,2,4,2,2,2,2,3,2]

最大尾长=4 ⇒ 估计数目：2^4/φ ≈ 16/0.773 ≈ 20.69

第6题

给定的数据流为：

```
x = [3, 1, 4, 1, 3, 4, 2, 1, 2]
```

先计算样本的均值：

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu = \frac{3+1+4+1+3+4+2+1+2}{9} = \frac{21}{9} = 2.333\ldots$$

二阶中心矩

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

计算每一项：

x_i	x_i - μ	(x_i - μ)²
3	0.6667	0.4444
1	-1.3333	1.7778
4	1.6667	2.7778
1	-1.3333	1.7778
3	0.6667	0.4444
4	1.6667	2.7778
2	-0.3333	0.1111
1	-1.3333	1.7778
2	-0.3333	0.1111

求和：

$$\frac{0.4444 + 1.7778 + 2.7778 + 1.7778 + 0.4444 + 2.7778 + 0.1111 + 1.7778 + 0.1111}{9} = \frac{12.0}{9} = 1.3333$$

所以：

二阶矩（奇异数）=1.3333

三阶中心矩

$$\mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3$$

计算每一项：

x_i	x_i - μ	(x_i - μ)³
3	0.6667	0.296
1	-1.3333	-2.370
4	1.6667	4.630
1	-1.3333	-2.370
3	0.6667	0.296
4	1.6667	4.630
2	-0.3333	-0.037
1	-1.3333	-2.370

x_i	x_i - μ	(x_i - μ)³
2	-0.3333	-0.037

求和：

$$\frac{0.296 - 2.370 + 4.630 - 2.370 + 0.296 + 4.630 - 0.037 - 2.370 - 0.037}{9} = \frac{2.668}{9} \approx 0.296$$

所以：

三阶矩=0.296

第7题

k 值	真实值	估计值	误差
5	3	3	0
15	9	10	1