参考答案

备注：目前参考答案内容以英文为主，仅供参考。

第一题

（a）To compute the cosines of the angles between the vectors for each pair of the three computers in terms of α and β, you can use the formula for the cosine similarity (cosine distance):

Cosine Similarity (A,B) =$\frac{A \cdot B}{\|A\| * \|B\|}$

Let's calculate the cosine similarity for each pair of computers:
For computers A and B:
A=[3.06,500α,6β]
B=[2.68,320α,4β]

Then, Cosine Similarity (A,B) =$\frac{3.06*2.68+500\alpha*320\alpha+6\beta*4\beta}{\sqrt{(3.06)^2+(500\alpha)^2+(6\beta)^2}*\sqrt{(2.68)^2+(320\alpha)^2+(4\beta)^2}}$

For computer A and C:
A=[3.06,500α,6β]
C=[2.92,640α,6β]

Then, Cosine Similarity (A,C) =$\frac{3.06*2.92+500\alpha*640\alpha+6\beta*6\beta}{\sqrt{(3.06)^2+(500\alpha)^2+(6\beta)^2}*\sqrt{(2.92)^2+(640\alpha)^2+(6\beta)^2}}$

For computer B and C:
B=[2.68,320α,4β]
C=[2.92,640α,6β]

Then, Cosine Similarity (B,C) =$\frac{2.68*2.92+320\alpha*640\alpha+4\beta*6\beta}{\sqrt{(2.68)^2+(320\alpha)^2+(4\beta)^2}*\sqrt{(2.92)^2+(640\alpha)^2+(6\beta)^2}}$

（b）

When α = 0.01 and β = 0.5, the angles between the vectors are computed using the cosine similarity formula with α = 0.01 and β = 0.5

Cosine Similarity (A,B) =$\frac{3.06*2.68+500*0.01*320*0.01+6*0.5*4*0.5}{\sqrt{(3.06)^2+(500*0.01)^2+(6*0.5)^2}*\sqrt{(2.68)^2+(320*0.01)^2+(4*0.5)^2}}$= 0.9908815005407525

Cosine Similarity (A,C) =$\frac{3.06*2.92+500*0.01*640*0.01+6*0.5*6*0.5}{\sqrt{(3.06)^2+(500*0.01)^2+(6*0.5)^2}*\sqrt{(2.92)^2+(640*0.01)^2+(6*0.5)^2}}$= 0.9915547143332561

Cosine Similarity (B,C) =$\frac{2.68*2.92+320*0.01*640*0.01+4*0.5*6*0.5}{\sqrt{(2.68)^2+(320*0.01)^2+(4*0.5)^2}*\sqrt{(2.92)^2+(640*0.01)^2+(6*0.5)^2}}$= 0.9691779219936828

第二题

（a）布尔代数如下：

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

因此，

J(A, B)=1-4/8=0.5；

J(A,C)=1-4/8=0.5；

J(B,C)=1-4/8=0.5

（b）cos(A, B)=4/6;

cos(A, C)=4/6;

cos(B, C)=4/6;

（c）阈值处理后

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

cos(A, B)=1/12^(1/2);

cos(A, C)=0;

cos(B, C)=1/3;

（d）归一化后如下：

$$\begin{bmatrix} 1 & 2 & \square & 2 & -2 & \square & 0 & -1 \\ \square & 0 & 1 & 0 & -2 & -1 & -2 & \square \\ -1 & \square & -2 & 0 & \square & 1 & 2 & 0 \end{bmatrix}$$

因此，cos(A, B)=4/140^(1/2);

cos(A, C)=-1/140^(1/2);

cos(B, C)=-7/10;

第三题

（a）设MM^T=A，　M^TM=B，那么

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

$$\mathrm{B} = \begin{bmatrix} 1 & 3 & 5 & 0 & 1 \\ 2 & 4 & 4 & 2 & 3 \\ 3 & 5 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 4 & 3 \\ 0 & 2 & 4 \\ 1 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 36 & 37 & 38 \\ 37 & 49 & 61 \\ 38 & 61 & 84 \end{bmatrix}$$

(b)特征值保留 4 位小数:

$$|A - \lambda \cdot I| = \begin{bmatrix} 14-\lambda & 26 & 22 & 16 & 22 \\ 26 & 50-\lambda & 46 & 28 & 40 \\ 22 & 46 & 50-\lambda & 20 & 32 \\ 16 & 28 & 20 & 20-\lambda & 26 \\ 22 & 40 & 32 & 26 & 35-\lambda \end{bmatrix} = 0$$

求得A的特征值:

λ_1 =153.5670, λ_2 =15.4330, $\lambda_3 = \lambda_4 = \lambda_5$ =0

类似地,

求得B的特征值:

λ_1 =153.5670, λ_2 =15.4330, λ_3 =0

(c)

根据特征值求特征向量:

$(A - \lambda \cdot I) \cdot y = 0$ 将y单位化: $x = \frac{y}{|y|}$

λ_1对应的单位特征向量:

$$x_1 = \begin{pmatrix} 0.2977 \\ 0.5705 \\ 0.5207 \\ 0.3226 \\ 0.4590 \end{pmatrix}$$

λ_2对应的单位特征向量:

$$x_2 = \begin{pmatrix} 0.1591 \\ -0.0332 \\ -0.7359 \\ 0.5104 \\ 0.4143 \end{pmatrix}$$

λ_3对应的单位特征向量:

$$x_3 = \begin{pmatrix} 0.1870 \\ 0.5705 \\ -0.2340 \\ 0.2327 \\ -0.7284 \end{pmatrix}$$

λ_4对应的单位特征向量:

$$x_4 = \begin{pmatrix} 0.9186 \\ -0.3401 \\ 0.0399 \\ -0.1713 \\ -0.0980 \end{pmatrix}$$

λ_5对应的单位特征向量:

$$x_5 = \begin{pmatrix} -0.0853 \\ -0.4820 \\ 0.3619 \\ 0.7429 \\ -0.2784 \end{pmatrix}$$

因此 A 矩阵的特征向量，也就是 SVD 分解中的 U 如下:

$$U = \begin{bmatrix} 0.2977 & 0.1591 & 0.1870 & 0.9186 & -0.0853 \\ 0.5705 & -0.0332 & 0.5705 & -0.3401 & -0.4820 \\ 0.5207 & -0.7359 & -0.2340 & 0.0399 & 0.3619 \\ 0.3226 & 0.5104 & 0.2327 & -0.1713 & 0.7429 \\ 0.4590 & 0.4143 & -0.7284 & -0.0980 & -0.2784 \end{bmatrix}$$

类似地，先求解 B 矩阵中特征值对应的单位特征向量:

λ_1对应的单位特征向量:

$$x_1 = \begin{pmatrix} 0.4093 \\ 0.5635 \\ 0.7176 \end{pmatrix}$$

λ_2对应的单位特征向量:

$$x_2 = \begin{pmatrix} -0.8160 \\ -0.1259 \\ 0.5642 \end{pmatrix}$$

λ_3对应的单位特征向量:

$$x_3 = \begin{pmatrix} -0.4082 \\ 0.8165 \\ -0.4082 \end{pmatrix}$$

因此，B 矩阵的特征向量，也就是 SVD 分解中的 V 如下:

$$V = \begin{bmatrix} 0.4093 & -0.8160 & -0.4082 \\ 0.5635 & -0.1259 & 0.8165 \\ 0.7176 & 0.5642 & -0.4082 \end{bmatrix}$$

那么，也可以求解V^T结果如下:

$$V^T = \begin{bmatrix} 0.4093 & 0.5635 & 0.7176 \\ -0.8160 & -0.1259 & 0.5642 \\ -0.4082 & 0.8165 & -0.4082 \end{bmatrix}$$

(d) 那么，$\sum$=

$$\begin{bmatrix} \sqrt{153.5670} & 0 & 0 \\ 0 & \sqrt{15.4330} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 12.3922 & 0 & 0 \\ 0 & 3.9285 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

因此，M=UΣV$^{\mathrm{T}}$=

$$= \begin{bmatrix} 0.2977 & 0.1591 & 0.1870 & 0.9186 & -0.0853 \\ 0.5705 & -0.0332 & 0.5705 & -0.3401 & -0.4820 \\ 0.5207 & -0.7359 & -0.2340 & 0.0399 & 0.3619 \\ 0.3226 & 0.5104 & 0.2327 & -0.1713 & 0.7429 \\ 0.4590 & 0.4143 & -0.7284 & -0.0980 & -0.2784 \end{bmatrix} \cdot \begin{bmatrix} 12.3922 & 0 & 0 \\ 0 & 3.9285 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.4093 & -0.8160 & -0.4082 \\ 0.5635 & -0.1259 & 0.8165 \\ 0.7176 & 0.5642 & -0.4082 \end{bmatrix}^T$$

$$= \begin{bmatrix} 0.2977 & 0.1591 & 0.1870 & 0.9186 & -0.0853 \\ 0.5705 & -0.0332 & 0.5705 & -0.3401 & -0.4820 \\ 0.5207 & -0.7359 & -0.2340 & 0.0399 & 0.3619 \\ 0.3226 & 0.5104 & 0.2327 & -0.1713 & 0.7429 \\ 0.4590 & 0.4143 & -0.7284 & -0.0980 & -0.2784 \end{bmatrix} \cdot \begin{bmatrix} 12.3922 & 0 & \cancel{0} \\ 0 & 3.9285 & \cancel{0} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.4093 & -0.8160 & \cancel{-0.4082} \\ 0.5635 & -0.1259 & \cancel{0.8165} \\ 0.7176 & 0.5642 & \cancel{-0.4082} \end{bmatrix}^T$$

$$= \begin{bmatrix} 0.2977 & 0.1591 & \cancel{0.1870} & \cancel{0.9186} & \cancel{-0.0853} \\ 0.5705 & -0.0332 & \cancel{0.5705} & \cancel{-0.3401} & \cancel{-0.4820} \\ 0.5207 & -0.7359 & \cancel{-0.2340} & \cancel{0.0399} & \cancel{0.3619} \\ 0.3226 & 0.5104 & \cancel{0.2327} & \cancel{-0.1713} & \cancel{0.7429} \\ 0.4590 & 0.4143 & \cancel{-0.7284} & \cancel{-0.0980} & \cancel{-0.2784} \end{bmatrix} \cdot \begin{bmatrix} 12.3922 & 0 & \cancel{0} \\ 0 & 3.9285 & \cancel{0} \\ \cancel{0} & \cancel{0} & 0 \\ \cancel{0} & \cancel{0} & 0 \\ \cancel{0} & \cancel{0} & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.4093 & -0.8160 & \cancel{-0.4082} \\ 0.5635 & -0.1259 & \cancel{0.8165} \\ 0.7176 & 0.5642 & \cancel{-0.4082} \end{bmatrix}^T$$

$$= \begin{bmatrix} 0.2977 & 0.1591 \\ 0.5705 & -0.0332 \\ 0.5207 & -0.7359 \\ 0.3226 & 0.5104 \\ 0.4590 & 0.4143 \end{bmatrix} \begin{bmatrix} 12.3922 & 0 \\ 0 & 3.9285 \end{bmatrix} \begin{bmatrix} 0.4093 & -0.8160 \\ 0.5635 & -0.1259 \\ 0.7176 & 0.5642 \end{bmatrix}$$

第四题

假阳率（假正例率）=$(1-e^{-km/n})^k$,其中 k represents the value of the hash function, n size of the bit array (the length of the Bloom Filter), m is the total number of members in the set s.

那么 k=3 时，$(1-e^{-km/n})^k = (1-e^{-3/8})^3 = 0.030579 \approx 3.1\%$
k=4 时，$(1-e^{-km/n})^k = (1-e^{-4/8})^4 = 0.0323969 \approx 3.2\%$

第五题

（a）**Step 1:** Hash value will be calculated by putting x value i.e., element into the given formula h(x)=2x+1 mod 32. For instance when x=3, then h(2)=7

Step 2: After finding the hash value convert that hash value into a binary digit. For instance, the calculated hash value for the element x=3 is 7, so now convert 7 into a binary digit by simply dividing the 2 by 7 and noting down the remainder, which will be 111. Also, place 0’s to make it a 5-bit binary integer, i.e., 00111

Step 3: Calculate the Tail Length =R Whenever the user apply a hash function h to a stream element a, the bit string h(a) will end in some number of 0s, possibly none. This is called Tail Length R for a and h.

Step 4: At last, calculate the number of distinct elements 2^R by simply placing the value of Tail Length R. By using the above Steps along with Flajolet-Martin Algorithm following calculation is done:

Element	Hashed value	Convert to Binary	Tail Length=R	Number of distinct elements 2^R
3	7	00111	0	1
1	3	00011	0	1
4	9	01001	0	1
1	3	00011	0	1
5	11	01011	0	1
9	19	10011	0	1
2	5	00101	0	1
6	13	01101	0	1
5	11	01011	0	1

From the above hash function, the maximum tail length R=0, so the number of distinct elements is estimated to be $2^R = 1$

（b）By using the above Steps along with Flajolet-Martin Algorithm following calculation is done:

Element	Hashed value	Convert to Binary	Tail Length=R	Number of distinct elements 2^R
3	12	01100	2	4
1	4	00100	2	4
4	16	10000	4	16
1	4	00100	2	4
5	20	10100	2	4
9	4	00100	2	4
2	8	01000	3	8
6	24	11000	3	8
5	20	10100	2	4

From the above hash function, the maximum tail length R=4, so the number of distinct elements is estimated to be $2^R = 16$.

第六题

这是一个长度为 9 的流，其中元素 3 出现 2 次，元素 4 出现 2 次，元素 1 出现 3 次，元素 2 出现 2 次。
因此奇异数=3*2^2+3^2=21
而三阶矩=3*2^3+3^3=51

第七题

k=5 时，预估 1 的数量=2*桶大小为 1+桶大小为 2 的一半=3。
真实 1 的数量=3，因此预估值和真实值差了 0；
k=15 时，预估 1 的数量=2*桶大小为 1+桶大小为 2+桶大小为 4+桶大小为 4 的一半=10。
真实 1 的数量=9。因此预估值和真实值差了 1。