National University of Computer and Emerging Sciences

School of Computing **Spring 2021** **Islamabad Campus**

Serial No:

DS-301 Big Data Analytics

Saturday, July 10^{th}, 2021

Course Instructor(s)

Hammad Majeed

Final Exam
Total Time: 3 Hours
Total Marks: 90

____

Signature of Invigilator

Student Name Roll No Section Signature

DO NOT OPEN THE QUESTION BOOK OR START UNTIL INSTRUCTED. *After asked to commence the exam, please verify that you have* ***11*** *different printed pages excluding the cover page. There are total of* ***6*** *questions.*

- Attempt on question paper. Attempt all of them. Read the question carefully, understand the question, and then attempt it.
- No additional sheet will be provided for rough work.
- Use permanent ink pens only. Any part done using soft pencil will not be marked and cannot be claimed for rechecking.
- Calculator sharing is strictly prohibited.
- Use permanent ink pens only. Any part done using soft pencil will not be marked and cannot be claimed for rechecking.

Question:	1	2	3	4	5	6	Total
Points:	6	28	8	20	16	12	90
Score:							

Question 1 .... **(6 Marks)**

(a) **(3 Marks)** Compute the Jaccard bag similarity of each pair of the following three bags: $A = \{1, 1, 1, 2\}$, $B = \{1, 1, 2, 2, 3\}$, and $C = \{1, 2, 3, 4\}$.

Solution:
Sim(A,B) = 3/9 = 1/3 – **2/3**
Sim(A,C)=2/8=1/4 – **2/4 = 1/2**
Sim(B.C)=3/9=1/3 – **3/4**

(b) **(3 Marks)** What are the first ten 3-shingles in the sentence *The most effective way to represent documents as sets, for the purpose of identifying lexically similar documents is to construct from the document the set of short strings that appear within it.*?

Solution: The set of the first 10 3-shingles is "The", "he ", "e m", " mo", "mos", "ost", "st ", "t e", " ef", "eff" . Or "The most effective", "most effective way", "effective way to", "way torepresent", "to represent documents", "represent documents as", "documents as sets", "as sets for", "sets for purpose", "for purpose of"

Question 2 .... **(28 Marks)**

Using table below, answer the following questions.

Element	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

(a) **(6 Marks)** Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.
You can redraw the above table and add three columns for the output of each hash function

Element	S1	S2	S3	S4	2x+1 mod 6	3x+2 mod 6	5x+2 mod 6
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

(b) **(2 Marks)** Which of these hash functions are true permutations?

(b) **h3**

(c) **(5 Marks)** How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

similarities	1-2	1-3	1-4	2-3	2-4	3-4
col/col	0	0	0.25	0	0.25	0.25
sig/sig	0.33	0.33	0.67	0.67	0.67	0.67

The estimated Jaccard similarities are not close to the true ones at all.

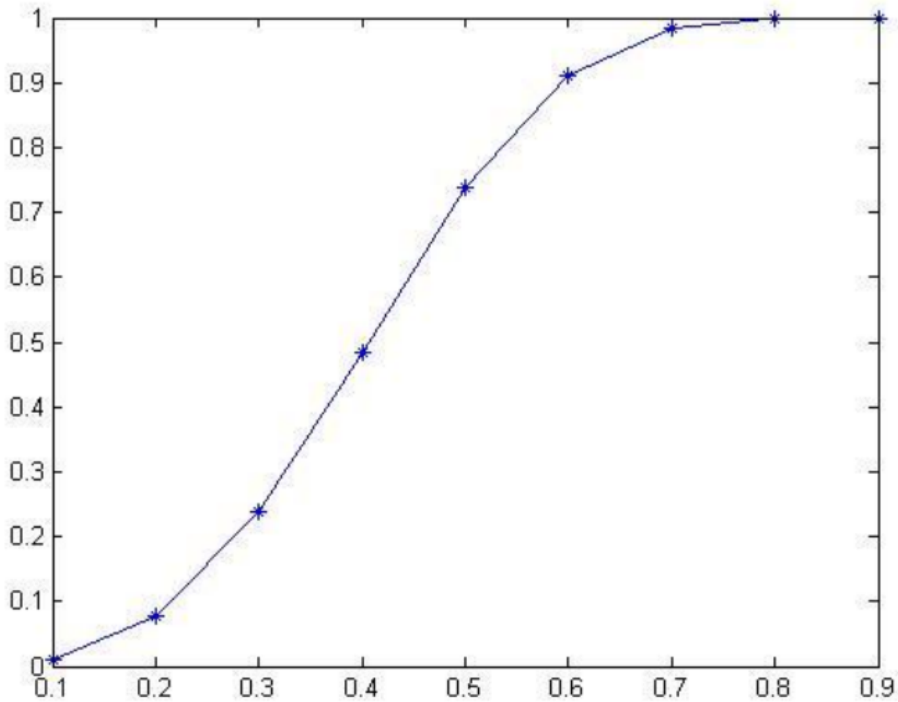
(d) Evaluate the S-curve $1-(1-s^r)^b$ for $s = 0.1, 0.2, \ldots, 0.9$, for the following values of r and b:

i. **(3 Marks)** r = 3 and b = 10.

Values of the S-curve for b=10 and r=3

s	$1-(1-s^r)^b$
0.1	0.0100
0.2	0.0772
0.3	0.2394
0.4	0.4839
0.5	0.7369
0.6	0.9123
0.7	0.9850
0.8	0.9992
0.9	1.0000

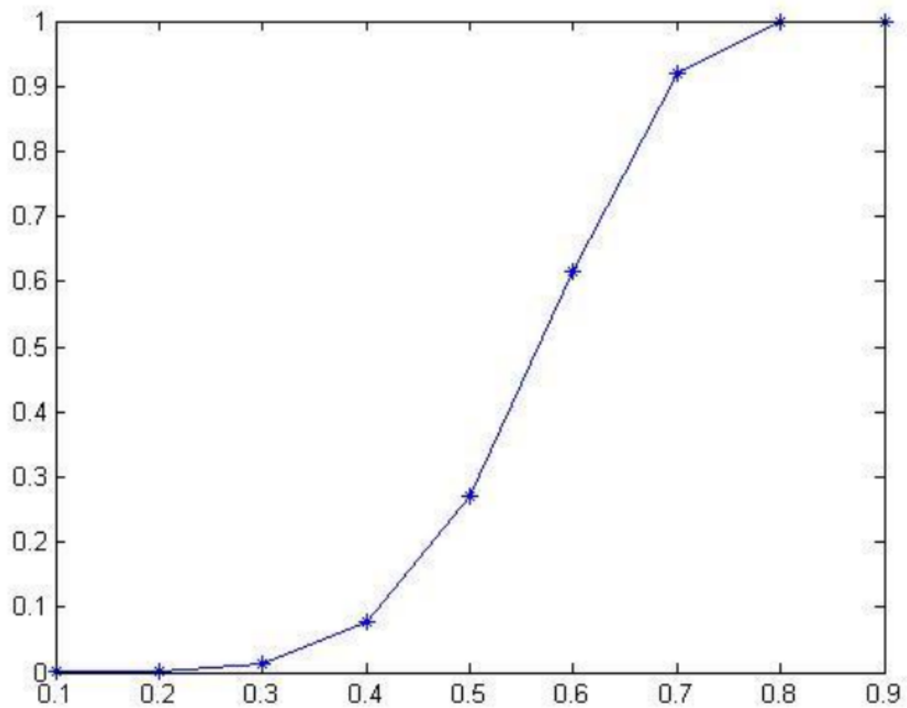
The figure is as follows:



ii. **(3 Marks)** r = 6 and b = 20.

s	$1 - (1 - s^r)^b$
0.1	0.0000
0.2	0.0013
0.3	0.0145
0.4	0.0788
0.5	0.2702
0.6	0.6154
0.7	0.9182
0.8	0.9977
0.9	1.0000

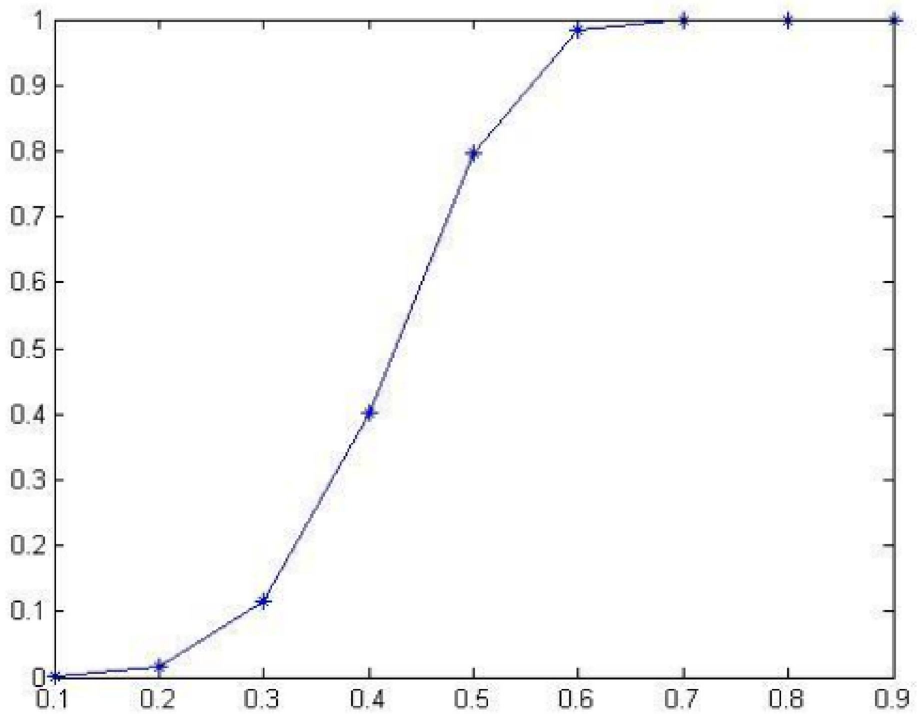
The figure is as follows:



iii. **(3 Marks)** r = 5 and b = 50.

Values of the S-curve for b=50 and r=5

s	$1-(1-s^r)^b$
0.1	0.0005
0.2	0.0159
0.3	0.1145
0.4	0.4023
0.5	0.7956
0.6	0.9825
0.7	0.9999
0.8	1.0000
0.9	1.0000



(e) **(4 Marks)** For each of the (r, b) pairs in the previous exercise, compute the threshold, that is, the value of s for which the value of $1-(1-s^r)^b$ is exactly 1/2.

Solution: Use $S = (1-0.5^{1/b})^{1/r}$ to calculate S values.
For b=10 and r = 3, S=0.4060881
For b=20 and r=6, S=0.5693534
For b=50 and r = 5, S=0.4243945

(f) **(2 Marks)** How does this value compare with the estimate of $(1/b)^{1/r}$

Solution: All the values are quite close to estimates $(1/b)^{1/r}$
For b=10 and r = 3, S=0.4060881 $\approx$ 0.4641589
For b=20 and r=6, S=0.5693534 $\approx$ 0.6069622
For b=50 and r = 5, S=0.4243945 $\approx$ 0.4573051

Question 3 . . . . **(8 Marks)**

(a) **(3 Marks)** Suppose we use Bloom filter to filter 1 billion members of set S using 8 billion bits. calculate the false-positive rate if we use three hash functions?

Solution: rate of FP $=(1-exp(-km/n))^k$
$k = 3$
$m = 1 * 10^9$
$n = 8 * 10^9$
$(1-exp(-3*1/8))^3 = 0.03057935$

(b) **(3 Marks)** What if we use four hash functions in the above situation?

Solution: rate of FP $=(1-exp(-km/n))^k$
$k = 4$
$m = 1 * 10^9$
$n = 8 * 10^9$
$(1-exp(-4*1/8))^4 = 0.02396865$

(c) **(2 Marks)** What will be count of false-negatives in both the cases ?

(c) **0**

Question 4 . . . . **(20 Marks)**

Suppose our stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Our hash functions will all be of the form $h(x) = ax + b$ mod 32 for some a and b. You should treat the result as a 5-bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is:

(a) **(3 Marks)** $h(x) = 2x + 1$ mod 32. ____2^0____

(b) **(3 Marks)** $h(x) = 3x + 7$ mod 32. ____2^4____

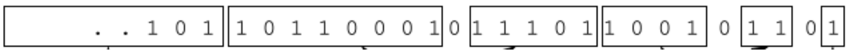
(c) **(3 Marks)** $h(x) = 4x$ mod 32. ____2^4____

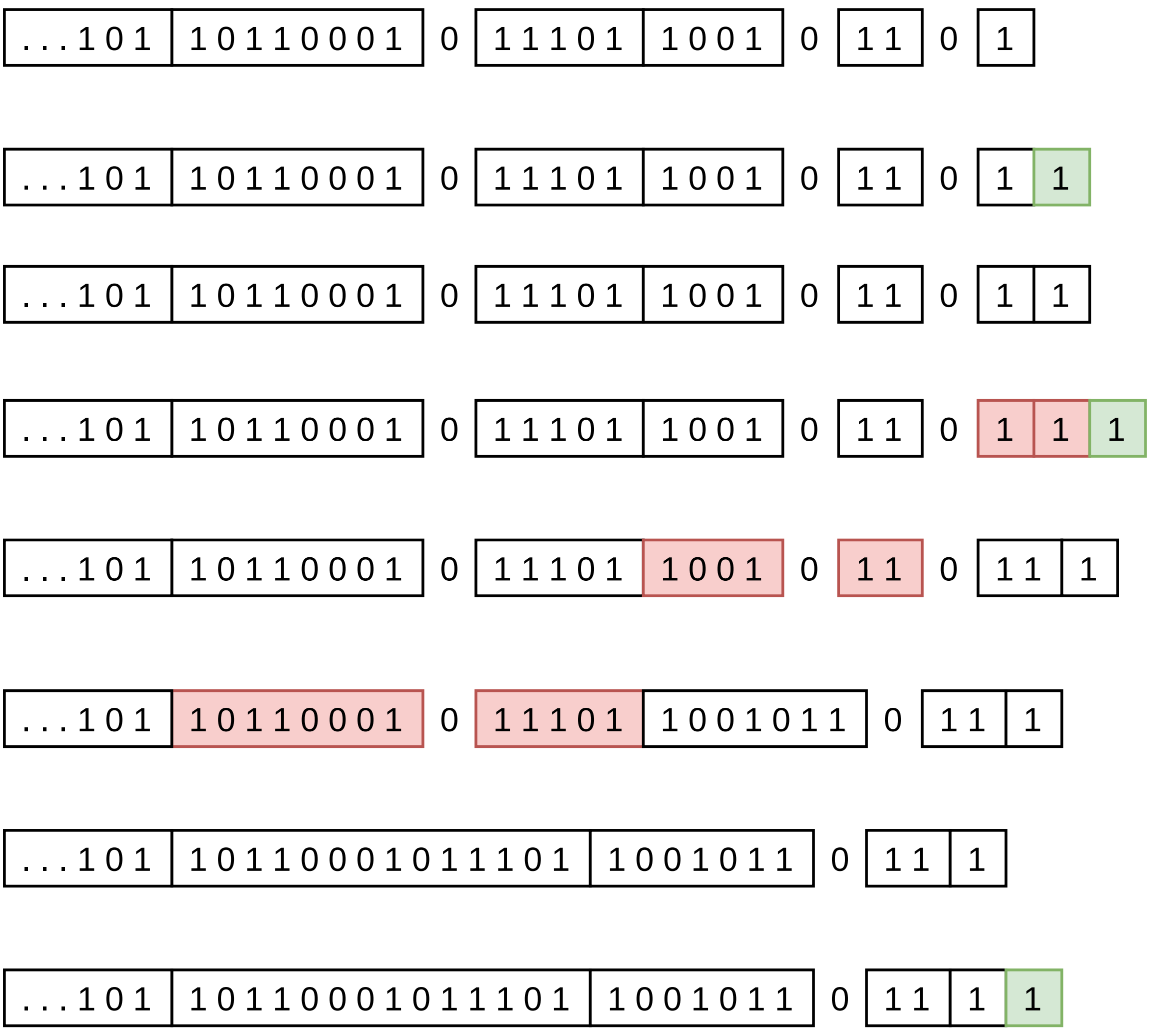
(d) Suppose the window .. 1 0 1 1 0 1 1 0 0 0 1 0 1 1 1 0 1 1 0 0 1 0 1 1 0. Estimate the number of 1's for the last k positions, for:

i. **(3 Marks)** k = 5, ____**3**____, what is the correct value ? ____**3**____

ii. **(3 Marks)** k=15, ____**8**____, what is the correct value ? ____**9**____

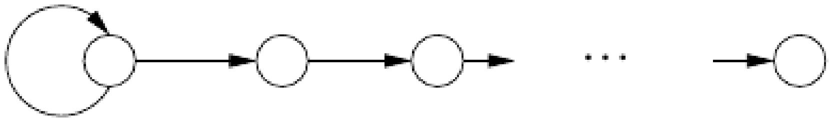
(e) **(5 Marks)** Describe what happens to the buckets if three more 1's enter the window represented by the figure below. You may assume none of the 1's shown leave the window.





Question 5 . . . . **(16 Marks)**

(a) **(6 Marks)** Suppose we recursively eliminate dead ends from the graph given below, solve the remaining graph, and estimate the PageRank for the dead-end pages. Suppose the graph is a chain of dead ends, headed by a node with a self-loop, as shown in the figure below. What would be the PageRank assigned to each of the nodes?



→ Recursively delete the dead-ends starting from right.

Transition Matrix M = [1], Starting PageRank will be [1]
and
$M^i v$ will result [1]

→ Now we will calculate the page Rank of each deleted node in the reverse order.
Node connected with the above node will have PageRank of $\frac{1}{2} \times 1 = \frac{1}{2}$.
All the right successors of this node will have PageRanks $\frac{1}{2}$.

(b) For the Web graph given in Figure 1, assuming only B is a trusted page:

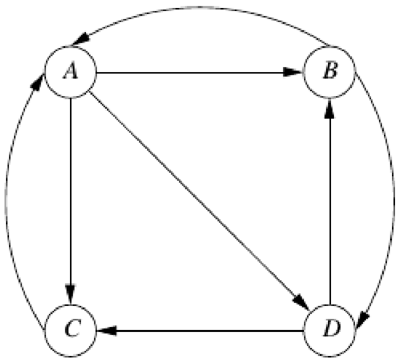


Figure 1: An imaginary model of web

i. **(3 Marks)** Compute the TrustRank of each page.

$\beta = 0.8$

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \quad V = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \quad S = \{B\} \; |S| = 1$$

$$e_s = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$V = \beta M V + (1-\beta) \frac{e_s}{|S|}$$

$$V = 0.8 \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.2 \\ 0 \\ 0 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.3 \\ 0.366 \\ 0.166 \\ 0.166 \end{bmatrix} \begin{matrix} - A \\ - B \\ - C \\ - D \end{matrix}$$

Trust Rank

ii. **(3 Marks)** Compute the spam mass of each page.

Page Rank.

$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \quad V = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

$$V = MV$$

$$= \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0.375 \\ 0.208 \\ 0.208 \\ 0.208 \end{bmatrix}$$

$$\text{Spam mass} = \frac{P_r - T_r}{P_r} \quad T_r \rightarrow \text{(trust rank)}$$

$$\begin{bmatrix} 0.375 \\ 0.208 \\ 0.208 \\ 0.208 \end{bmatrix} - \begin{bmatrix} 0.3 \\ 0.366 \\ 0.166 \\ 0.166 \end{bmatrix} \Bigg/ \begin{bmatrix} 0.375 \\ 0.208 \\ 0.208 \\ 0.208 \end{bmatrix}$$

$$= \begin{bmatrix} 0.20 \\ -0.76 \\ 0.20 \\ 0.20 \end{bmatrix} \begin{matrix} \rightarrow A \\ - B \\ - C \\ - D \end{matrix}$$

(c) **(4 Marks)** Compute the hubbiness and authority of each of the nodes in our original Web graph of Figure 1.

$$h = \lambda L a$$
$$a = \mu L^T h$$

$$h = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad L = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad L^T = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

$$L^T h = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}$$

$$a = \mu L^T h = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{here } \mu = \frac{1}{2}$$

$$La = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

$$h = \lambda L a = \begin{bmatrix} 1 \\ 2/3 \\ 1/3 \\ 2/3 \end{bmatrix} \quad \text{here } \lambda = 1/3$$

First iteration complete

hubbiness $\rightarrow h = \begin{bmatrix} 1 \\ 2/3 \\ 1/3 \\ 2/3 \end{bmatrix}$

authority $\rightarrow a = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

Question 6 .... **(12 Marks)**

(a) Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if b divides i with no remainder. Answer the following questions:

Note: This question is slightly different from the question in the quiz taken in the class.

i. **(3 Marks)** If the support threshold is 5, which items are frequent?

> **Solution:** All the items numbered from 1-100 with divisors count (number of basket membership) greater than and equals to 5 will be frequent items. For example divisors of 36 are 1, 2, 3, 4, 6, 9, 12 ,18, 36. Count is 9 therefore it is frequent item. Whereas 5 has divisors 1 and 5. The count is $2 < 5$ therefore its not a frequent item.

ii. **(3 Marks)** If the support threshold is 5, which pairs of items are frequent?

> **Solution:** Clearly, the of numbers picked from part (i) will make pairs that are frequent. Therefore, a pair (a, b) will be considered frequent if it satisfies following two conditions:
> a and b are frequent items
> $|divisors(a) \cap divisors(b)| \geq 5$

(b) For the data of previous part, what is the confidence of the following association rules?

i. **(3 Marks)** $\{24, 60\} \rightarrow 8$

> **Solution:** 24 is member of baskets numbered 1,2,3,4,6,8,12,24 (frequent item)
> 60 is member of baskets numbered 1,2,3,4,5,6,10,12,15,20,30,60 (frequent item)
> $|divisor(24) \cap divisor(60)| = |1, 2, 3, 4, 6, 12| = 6 > 5$ (frequent pair)
> 8 is member of baskets numbered 1,2,4,8
> $|divisor(24) \cap divisor(60) \cap divisor(8)| = |1, 2, 4| = 3$
> confidence $= 3/6 = 1/2$

ii. **(3 Marks)** $\{2, 3, 4\} \rightarrow 5$

> **Solution:** 2 is member of baskets numbered 1,2
> 3 is member of baskets numbered 1,3
> $|divisor(2) \cap divisor(3)| = |1| = 1$
> 5 is member of baskets numbered 1,5
> $|divisor(2) \cap divisor(3) \cap divisor(5)| = |1| = 1$
> confidence $= 1/1 = 1$