ourdiu.com | Shared by MD. GOLAM MUSTAKIM NAEEM

Daffodil International University (DIU)

Department of Computing and Information System (CIS)

Course: Big Data Analytics and Applications (CIS312)

Class Test 4 (CT4) — Section A

Time: 30 Minutes | **Full Marks:** 15 *(3 Questions × 5 Marks each)*

Question 1: Bloom Filter (5 Marks)

(a) Why does a Bloom Filter guarantee zero **False Negatives**, whereas a **False Positive** can still occur? *(2 Marks)*

(b) Consider an empty Bloom Filter of size $m = 8$ bits (indices $0 \dots 7$) with $k = 3$ hash functions:

- **Insert** `"dhaka"` $\rightarrow \{1, 3, 6\}$
- **Insert** `"rangpur"` $\rightarrow \{2, 3, 7\}$
- **Query** `"sylhet"` $\rightarrow \{1, 4, 7\}$
- **Query** `"khulna"` $\rightarrow \{2, 6, 7\}$

Write down the final bit array $B[0 \dots 7]$ after both insertions, and state the query result (**Definitely Not** or **Probably Present / False Positive**) for `"sylhet"` and `"khulna"`. *(3 Marks)*

Question 2: HyperLogLog & Count-Min Sketch (5 Marks)

(a) **(HyperLogLog):** A HyperLogLog sketch uses $b = 2$ prefix bits ($m = 4$ buckets, initialized to $[0, 0, 0, 0]$). Update the registers for the following 3 events and show the final array $M[0 \dots 3]$ *(harmonic mean calculation is **not** required)*:

- `user_1: 01 001011`
- `user_2: 11 100100`
- `user_1: 01 001011` *(duplicate)*

(2.5 Marks)

(b) **(Count-Min Sketch):** A 3×4 Count-Min Sketch (rows $0 \dots 2$, columns $0 \dots 3$, initially all 0) processes the stream `["A", "B", "A"]` with row hash indices (h_0, h_1, h_2):

- `"A"` $\rightarrow (1, 2, 0)$
- `"B"` $\rightarrow (1, 0, 3)$

Show the updated 3×4 matrix and find the estimated count $\hat{f}(\texttt{"A"}) = \min(C[0, 1], C[1, 2], C[2, 0])$ *(2.5 Marks)*

Question 3: MinHashing (5 Marks)

(a) State the formula for **Jaccard Similarity** $J(A, B)$ and calculate the exact Jaccard Similarity for $A = \{\texttt{"big data"}, \texttt{"data spark"}, \texttt{"spark sql"}\}$ and $B = \{\texttt{"big data"}, \texttt{"data spark}$ *(2 Marks)*

(b) Given two documents D_1, D_2 and two permutations showing **row visit order** (1 = visited first):

Shingle	D_1	D_2	π_1 (Order)	π_2 (Order)
r_1	1	1	2	1
r_2	0	0	3	2
r_3	1	0	1	3

Construct the 2×2 Signature Matrix $[\text{Sig}(D_1), \text{Sig}(D_2)]$ and compute the estimated similarity $\hat{J}(D_1, D_2)$. *(3 Marks)*