Instructions:
1. Please answer all questions in Part A.
2. You can optionally answer the bonus question in Part B. You can obtain additional marks for the bonus question if you answer it correctly.
3. The total marks in Part A are 100.
4. The total marks in Part B are 20.
5. The total marks you can obtain in this assignment are 100 only.
   If you answer the bonus question in Part B correctly, you can obtain additional marks. But, if the total marks you obtain from Part A and Part B are over 100, your marks will be truncated to 100 only.
6. You can use a coupon to waive any question you want in Part A only (not Part B) and obtain full marks for this question.
7. You can waive at most one question in Part A only (not Part B) in each assignment.
8. You can also answer the question you will waive. We will also mark it but will give full marks to this question.
9. The coupon is non-transferrable. That is, the coupon with a unique ID can be used only by the student who obtained it in class.
10. Please staple the coupon to the submitted assignment.
11. Please write down the question no. you want to waive on the coupon.

Part A (Compulsory Questions)

Q1 [20 Marks]

(a) Consider N = 10.
Suppose s = 0.4 and \( \varepsilon = 0.2 \).

<table>
<thead>
<tr>
<th>Item</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ideal Algorithm</td>
</tr>
<tr>
<td>I1</td>
<td>4</td>
</tr>
<tr>
<td>I2</td>
<td>3</td>
</tr>
<tr>
<td>I3</td>
<td>2</td>
</tr>
<tr>
<td>I4</td>
<td>1</td>
</tr>
</tbody>
</table>

Suppose each algorithm has the following output. An ideal algorithm is the algorithm which gives the correct frequent items and the correct infrequent items.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideal Algorithm</td>
<td>Frequent Items: I₁, Infrequent Items: I₂, I₃, I₄</td>
</tr>
<tr>
<td>Algorithm 1</td>
<td>Frequent Items: I₁, I₄, Infrequent Items: I₂, I₃</td>
</tr>
<tr>
<td>Algorithm 2</td>
<td>Frequent Items: I₁, I₂, Infrequent Items: I₃, I₄</td>
</tr>
<tr>
<td>Algorithm 3</td>
<td>Frequent Items: I₁, I₃, Infrequent Items: I₂, I₄</td>
</tr>
</tbody>
</table>

Which algorithm(s) (except Ideal Algorithm) has/have an \( \varepsilon \)-deficient synopsis? Why?
Which algorithm(s) (except Ideal Algorithm) does/do not have an \( \varepsilon \)-deficient synopsis? Why?
(b) In class, we learnt that the memory consumption of the lossy counting algorithm is $\lceil 1/\varepsilon \log (\varepsilon N) \rceil$ entries when there is only one item at one time instance. Suppose that there is only one transaction instead of one item at one time instance where a transaction is a set of items. Let $m$ be the maximum size of a transaction. Suppose that we want to find frequent itemsets instead of frequent items.

(i) What is the greatest possible memory consumption of this algorithm in this case?

(ii) Is the greatest possible memory consumption in (i) nicely bounded? A memory consumption is said to be nicely bounded if the consumption does not include any exponential factor on some variables which may be potentially large. One nicely-bounded example is “$N^2 \log N$” and one non-nicely bounded example is “$2^N$”.

Q2 [20 Marks]

(a) Consider Sticky Sampling Algorithm. (Hints: You can read the paper about Sticky Sampling Algorithm for reference. The proof there is too brief. Please use your own words and elaborate the proof in this question.)

Prove that Sticky Sampling Algorithm has an $\varepsilon$-deficient synopsis with probability at least 1-\( \delta \).

(b) What are the differences between Sticky Sampling Algorithm and Lossy Counting Algorithm?

(c) What are the differences between Lossy Counting Algorithm and Space-Saving Algorithm?

Q3 [20 Marks]

We want to find frequent items over a sliding window.

Suppose that we want to re-use the Space-Saving algorithm.

Assume that we use the batch-based approach for this purpose. Let $B$ be the batch size. The first $B$-th data points form the first batch. The next $B$-th data points form the second batch. We can also form other batches for the remaining data points.

(a) Suppose that we are interested in finding frequent items over the 4 recent batches after reading $100 \times B$ data points. Note that $100 \times B$ is a multiple of the batch size $B$.

(i) Please design an algorithm for this problem such that the error introduced (in terms of the frequency of an item in fraction) is minimized.

(ii) What is the greatest error in any estimated frequency (in fraction) for the above algorithm?

(b) Suppose that we are interested in finding frequent items over the most recent $4B$ data points after reading $100.5 \times B$ data points. Note that $100.5 \times B$ is not a multiple of the batch size. Besides, the most recent $4B$ data points discussed above come from 4 complete batches and 1 incomplete batch. Specifically, if the 4 complete batches are $B_1$, $B_2$, $B_3$ and $B_4$ where $B_1$ is the oldest batch, then conceptually, some older data points in $B_1$ should not be used for finding frequent items but some newer data points in $B_1$ should be used. Besides, conceptually, all data points in $B_2$, $B_3$ and $B_4$ should be used. All data points in the incomplete batch should be used.

(i) Please design an algorithm for this problem such that the error introduced (in terms of the frequency of an item in fraction) is minimized.

(ii) What is the greatest error in any estimated frequency (in fraction) for the above algorithm?
Q4 [20 Marks]

Assume that there are four web sites A, B, C and D. Suppose that A and B point to each other, and C and D point to each other. Furthermore, A points to C. What is the stochastic matrix created by the page rank method for the four sites?
What equation will be solved to decide on the ranking, taking into the consideration of the possibility of the spider trap problem?
What will be the resulting ranking?

Q5 [20 Marks]

Consider three dimensions (D₁, D₂, D₃). Let \( D = \{D₁, D₂, D₃\} \).
There are the following five data points:

\[
a: (4, 2, 1), b: (6, 5, 5), c: (1, 2, 2), d: (7, 2, 1), e: (1, 2, 3)
\]

Suppose that a smaller value is more preferable.

(a) Which points are in the skyline?

(b) If we project all data points to a subspace \( D' \) (\( D' \subset D \)), then we can determine the subspace skyline in \( D' \) (an object whose projection is not dominated by the projections of other objects is in the subspace skyline). We call \( D \) the full space.

(i) For the above data set, determine the skyline for each of the 2-dimensional subspaces, which are \( \{D₁, D₂\}, \{D₂, D₃\} \) and \( \{D₁, D₃\} \). That is, which points are in the subspace skyline in each subspace?

(ii) Is it true that a skyline object in the full space \( D \) is always a skyline object in a subspace \( D' \subset D \)? Why?

(iii) Is it true that a skyline object in a subspace \( D' \) (\( D' \subset D \)) is always a skyline object in the full space \( D \)? Why?

(iv) Let \( S' \) be the subspace skyline for a subspace \( D' \) and \( S \) be the skyline for the full space \( D \). Prove that there exists a point \( p \) in \( S' \) such that \( p \) is in \( S \).
Part B (Bonus Question)

Note: The following bonus question is an OPTIONAL question. You can decide whether you will answer it or not.

Q6 [20 Additional Marks]

In the skyline analysis we learnt in class, we assume that each attribute follows a particular ordering. For example, in attribute Price, a smaller value is more preferable. However, some categorical attributes (e.g., brand name) do not have this pre-defined ordering. Some customers prefer one brand name a to another brand name b while some prefer b to a. If a customer prefers a to b, we denote it by a pair (a, b). A customer can specify a set P of preferences which include multiple pairs. Denote the skyline for a set P of preferences to be SKY(P). Consider three hotels with two attributes, Price and Brand Name: h₁ (100, Royal), h₂ (200, Marriott) and h₃ (250, Marriott). If P = {}, then SKY(P) = {h₁, h₂}. If P = {(Royal, Marriott)}, then SKY(P) = {h₁}. If P = {(Marriott, Royal)}, then SKY(P) = {h₁, h₂}. Suppose that C is a set containing all values (e.g., Royal and Marriott) in the categorical attribute (e.g., Brand Name). Let Pa = {(a, y) | y ∈ C\{a}}, Pb = {(b, y) | y ∈ C\{b}}, and Pa<b = {(a, y) | y ∈ C\{a}}∪{(b, y) | y ∈ C\{a, b}}

Let X = SKY(Pa<b) – SKY(Pa) ∩ SKY(Pb). Given a set S of points and a in C, we denote K(S, a) to be the set of points in S which categorical attribute value is a.

(a) Is “X ⊆ SKY(Pₐ)” true? If yes, prove it formally. If no, please give a counter example.
(b) Is “X ⊆ SKY(Pₜ)” true? If yes, prove it formally. If no, please give a counter example.
(c) Is “X ⊆ K(SKY(Pₐ), a)” true? If yes, prove it formally. If no, please give a counter example.
(d) Is “X ⊆ K(SKY(Pₜ), b)” true? If yes, prove it formally. If no, please give a counter example.