Data Mining Assignment - 1

Name - Ritu Kumari
Roll No. - 2016078
Branch - CSE, 3rd Year

1. To obtain a uniform sample of size s from a population set of size N (s << N) and N is known beforehand, we can use the following algorithm. Let the N tweets be represented as integers.

   Algorithm

   a. Create an array sample[0...s-1] and population[0...N-1].
   b. Create an array unique[0….s-1] to keep only unique integers in the sample and initialise all its elements to some -ve value.
   c. Populate the population array with the N integers.
   d. Generate s random integers between 0 and N-1. (Random integers are indices)
   e. Store the s generated integers into
   f. Retrieve elements from the population array that are positioned at those s generated integers.

   Pseudocode:

```
Sampling (P[0....N-1], N, s)    // P - population array, N - No of tweets, s - sample size
        S[0....s-1]                 // create a sample array of size s
        U[0....s-1]          // create an array U to store unique indices
         for i = 0 to s-1
              j = random(0, N)  // generates random integers in range [0, N-1]
              if j is in U    // requires a for loop from i = 0 to s-1 to check if j is already in U
                    i--
              else
                    S[i] = P[j]
```

   Time Complexity: $O(s^2)$

2(d).

Item	N = 100	N = 500	N = 1000	N = 10000
A_1 = 1	47	247	431	4600
A_2 = 2	32	179	387	3770
A_3 = 3	15	87	232	1933
A_4 = 4	20	76	137	1223
A_5 = 5	10	54	118	1072
A_6 = 6	18	87	176	1732
A_7 = 7	7	50	116	894
A_8 = 8	5	42	119	1002
A_9 = 9	55	238	464	4566
A_{10} = 10	14	96	209	2023
A_{11} = 11	8	68	120	1412
A_{12} = 12	29	147	229	2569
A_{13} = 13	25	106	209	2223
A_{14} = 14	22	146	302	3048
A_{15} = 15	21	64	147	1613
A_{16} = 16	49	244	475	4613
A_{17} = 17	47	232	467	4726
A_{18} = 18	17	73	108	1158
A_{19} = 19	12	63	111	1238
A_{20} = 20	47	201	443	4585
Sum of frequencies of elements	500	2500	5000	50000

2(e). Observations are:

A. There is exactly the same five elements that are included in every sample and have almost the same frequency count i.e A_1, A_9, A_{16}, A_{17} and A_{20}.
B. The positive and negative peaks are obtained at (generally) the same points.
C. The sum of the frequencies of the elements is 5N.
D. A case of reservoir sampling.

The sampling process is biased since only those five items are included every time in the final sample as observed in point A. Also the probability of every item to be included into the sample is not same.

2(f). **Proof -** We will prove this by induction. We need to show that after n+1 element arrives the sample maintains the property that the probability that a stream point is included in the sample after (n+1)th item arrives is s/(n + 1).

Base Case - (n = s) After n = s stream points have arrived,
Probability of any stream point being included in the sample of size s = s/s = 1.

Inductive Hypothesis - After n stream points have arrived,
Probability of any stream point being included in the sample of size s = s/n
--------- (1)

Inductive Step - After n+1 element arrives,
probability that the elements in the sample are kept = (Probability that (n+1)th element is discarded) + (Probability that (n+1)th element is included) X (Probability that the element in the sample is not picked for replacement by (n+1)th element)
= [1 - s / (n + 1)] + [s / (n + 1)] X [(n - 1) / n]
= n / (n + 1) **--------- (2)**

So, from **(1)** and **(2)**,
Probability that a stream point is included in the sample after (n+1)th item arrives = s/n X n/ (n + 1) = s / (n + 1)

=> Hence proved.