Assignment 4:

Question 1:

Only N of that million items are frequent, so we're only interested in keeping an array of length N to keep the counts of the frequent items for our second pass. This will take up 4N bytes and will replace the 4 million mentioned above.

The remainder of the memory required will be taken up with our count of the frequent pairs, which we can do using a triangular array or a hash table of triples.

The triangular array will consist of a one dimensional array of length (N^2) which at 4 bytes per slot, takes up $2N^2$ bytes. So for the triangular array method, our total number of bytes S, will be calculated as follows

$$S = 2N^2 + 4N$$ //

Question 2:

	B	C	P	M	J
1	x		x		
2		x		x	
3	x	x			x
4			x	x	
5	x	x		x	
6				x	x
7			x		x
8	x	x		x	x

B → Beer
C → Coke
P → Pepsi
M → Milk
J → Juice

S(B,C) = 3
S(B,P) = 1
S(B,J) = 2
S(B,M) 2

S(C,P) : 0
S(C,M) : 3
S(C,J) : 2

S(P,M) : 1
S(P,J) = 0

S(M,J) : 2

Suppose Threshold (s) = 2
so select item which have threshold equal or greater than 2.

[B,C], [B,M], [B,J]

[C,M], [C,J], [M,J]

Question 8

Expected memory for pass 2:

Expected no. of frequent buckets = 1,000,000

Each frequent bucket hashes to $\left[1 + \frac{P}{\text{buckets}}\right]$ pairs which will simplify to $\frac{P}{\text{buckets}}$ pairs.

Probability of bucket to be frequent & map to frequent bucket = $\frac{1,000,000}{\text{buckets}}$

Total expected memory consumption for pass 2

12 bytes/pair = $\frac{P \times 12,000,000}{\text{buckets}}$

In pass 1

Let us use $\frac{1}{4}$ mb space to count item & remainder of S as a hash table to help eliminate non frequent pairs

(1) integers (4 bytes) = $\frac{S \text{ mb} \cdot 8}{4} \approx S/4$ buckets hash table

Make table is compressed to bit map before pass 2

(2) But no of bucket is $S/4$. Two buckets will have less space in pass 2.

$\therefore S/4$ → upper bound for buckets

$S/4$ buckets are compressed to $S/32$ bytes using bitmap

$S * \frac{31}{32}$ bytes for counting pairs, as per the assumption, 2nd pass

$\left[P * \frac{12,000,000}{buckets} \right]$ bytes for counting so it buckets $= S/4$

we need $P * 12,000,000 / (S/4)$

$= \frac{48,000,000 * P}{S}$ bytes for counting since we have $S * \frac{31}{32}$ bytes free

$$S * \frac{31}{32} = \frac{48,000,000 * P}{S}$$

$$S^2 * 31 = 48,000,000 \times 32 \times P$$

$$P = \frac{S^2 \times 31}{48,000,000 \times 32}$$

Our bound is $P < \frac{S^2}{49,548,387}$

Question 4

Set of item : {A, B, C, D, E, F, G, H}

Maximal frequent itemset

{A, B} {A, C} {A, D}, {B, C}, {E}, {F}

→ For singleton sets

Singletons sets which are not in frequent sets are in negative border.

{G} {H} → -ve border

→ For pairs, item will be in -ve border when all of its subsets are in frequent subset but not set.

{A E} {A F} {B D} {B E} {B F} {C D}

{C E} {C F} {D E} {D F} {E F}

→ For triplen → {A, B, C}

sets in -ve border : {G} {H}

{A, E}, {A, F}, {B, D}, {B, E}, {B, F}, {E, D},

{C, F}, {C, E}, {D, E}, {D, F} {E, P}, {A, B, C}