UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

INFORMATION RETRIEVAL AND RECOMMENDER SYSTEMS

Exercise 3: Implementation and indexing

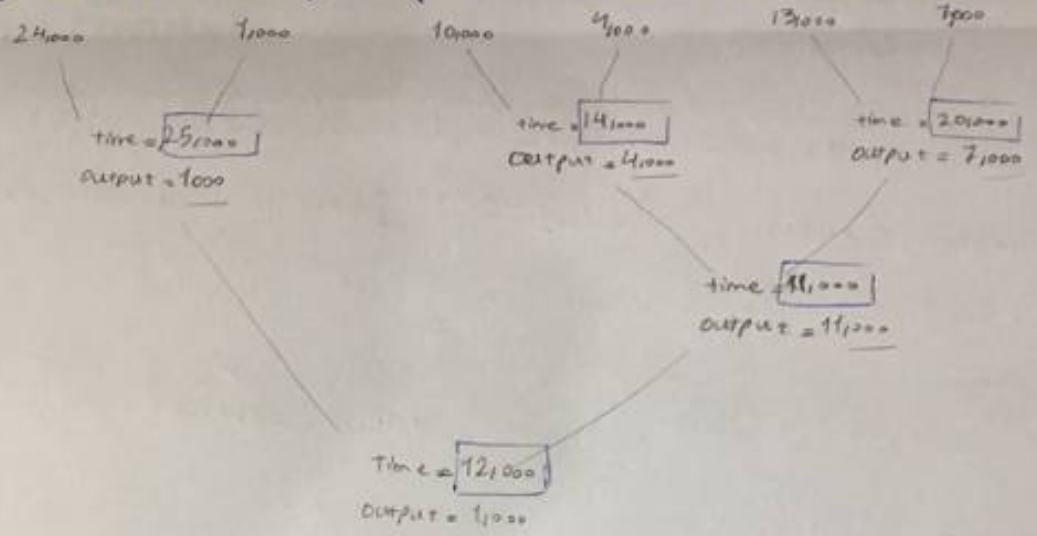
Ashwin Kumar Gururajan, Mohana Fathollahi

13th October, 2022

Exercise 4:

Main:

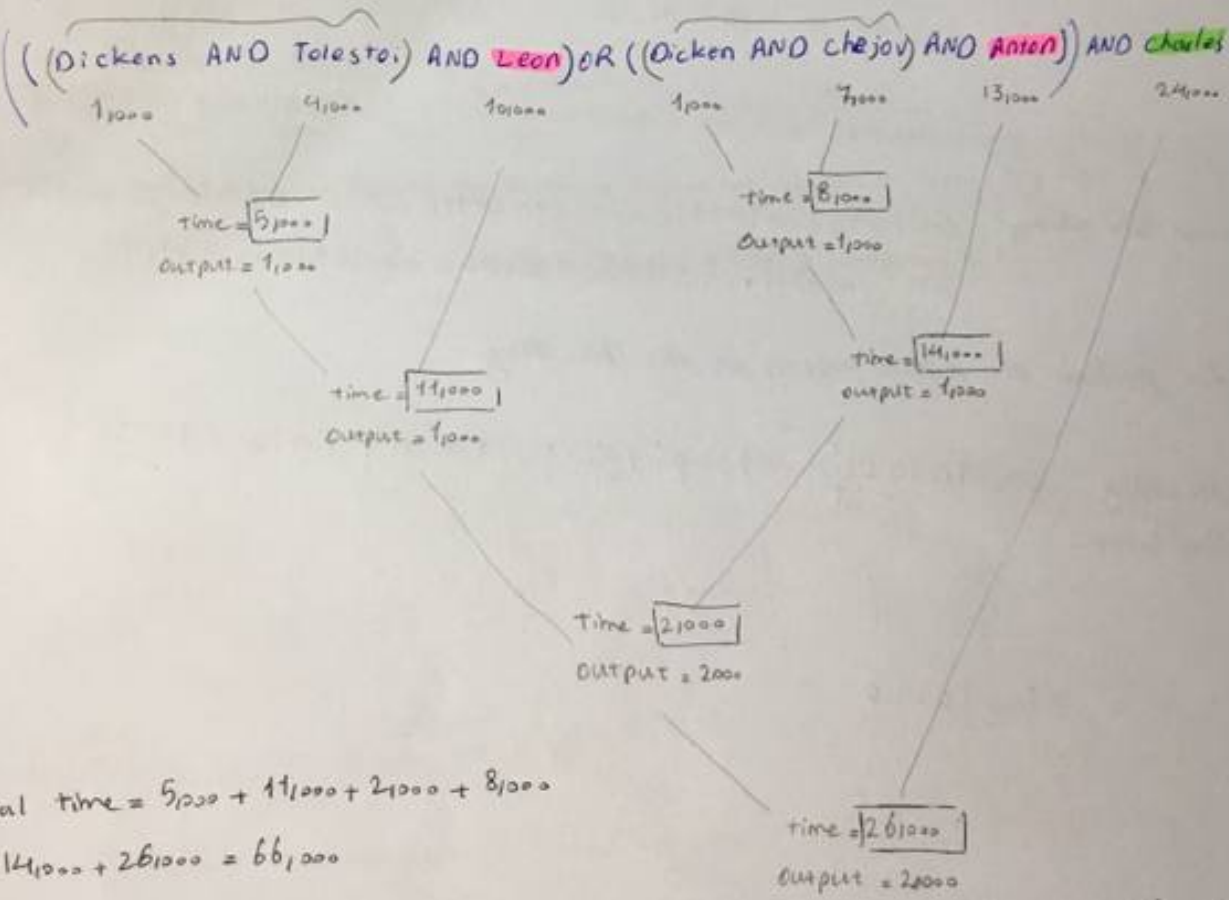
(Charles AND Dicken) AND ((Leon AND Tolestoi) OR (Anton AND chejov))



Total time = 25,000 + 14,000 + 20,000 + 11000 + 12,000 = 82,000

Suggestion:

((Dickens AND Tolestoi) AND Leon) OR ((Dicken AND chejov) AND Anton) AND Charles



Total time = 5,000 + 11,000 + 21000 + 8,000 + 141,000 + 261,000 = 66,000

To have smaller size of output, it is better to consider terms that have lower frequency inside of boolean operations and put terms that have higher frequency outside of boolean operations. In this case Leon / Anton / Charles have highest frequency.

Exercise 5 : -1-

[10, 1, 15, 3, 22, 2, 23.4, 34,1 , 44,1, 50, 2, 58. 8, 90, 1, 101, 1, 112, 2]

1. Gap Compression [(id 1, f1), . . . , (idk, fk)] —compressed→ [(id1, f1), . . . , (id k − id k−1, fk)]

[A 10, 1, B 5,3, C 7,2, D 1,4, E 11,1, F 15,1, G 6,2, H 8,8, I 32,1, J 11,1, K 11,2]

2. Elias' Gamma

[A 0001010, 1 , B 00101,3, C 00111,2 , D 1,4 , E 0001011, 1 , F 0001010,1 , G 00110, 2

H 0001000, 8 , I 000001 00000, 7 , J 0001011, 1 , K 0001011, 2]

3. Compress frequency , using unary self-delimiting

[A 0001010, 0 , B 00101, 110 , C 00111, 10 , D 1, 1110 , E 0001011, 0 , F 0001010, 0,

G 00110, 10 , H 0001000, 11111110, I 000001 00000, 0 , J 0001011, 1 , K 0001011, 10]

Final bit string: 000101000010111000111101111000010110000101000011010

0001000011111110000001000000000010111000101110

2- perform the inverse process on the bit string

1. Decoding Elias' Gamma

0000101011 0 | 00100 | 1040 | 00100 0 | 10 | 00 1 10 110 |

N=4 → 2^4 + 5 → [21, 10]

N=2 → 2^2 + 0 → [4, 0] [1,0] [1,0] [4,0]

2^2 2^2 2

[6, 110]

0100 | 0110

N=1 → 2^1 + 0 → [2,0]

2^1 1 → [3,0]

2.

Decoding gap compression

Result of step 1: [21, 10, 4, 0, 1, 0, 1, 0, 4, 0, 1, 0, 6, 110, 2, 0, 3, 0]

[21, 40, 25, 0, 26, 0, 27, 0, 31, 0, 32, 0, 38, 110, 40, 0, 43, 0]

3. Decoding unary self-delimiting

[21, 2, 25, 1, 26, 1, 27, 1, 31, 1, 32, 1, 38, 3, 40, 1, 43, 1]

Exercise 6:

((A and B) and C) and D) and E

10,000 20,000 40,000 80,000 120,000

$10{,}000 \times \log_2 20{,}000 + 10{,}000 \log_2 40{,}000 + 10{,}000 \log_2 80{,}000 + 10{,}000 \log_2 120{,}000$

$= 10{,}000\,(14{,}29 + 15{,}29 + 16{,}29 + 16{,}87) = 627{,}400$

A and (B and (C and (D and E)))

20,000 40,000 80,000 120,000

$80{,}000 \log_2 120{,}000 + 40{,}000 \log_2 80{,}000 + 20{,}000 \log_2 40{,}000 + 10{,}000 \log_2 20{,}000$

$= 80{,}000 * 15{,}87 + 40{,}000 * 16{,}29 + 20{,}000 * 15{,}29 + 10{,}000 * 14{,}29$

$= 2{,}449{,}900$

10,000 40,000 120,000

((A and B) or (C and D)) or (E and F)

A first or

$10{,}000 \log_2 20{,}000 + 40{,}000 \log_2 80{,}000 + 150{,}000 \log_2 120{,}000 + (10{,}000 + 40{,}000) +$

second or

$(50{,}000 + 120{,}000) = 10{,}000 \times 14{,}29 + 40{,}000 * 16{,}29 + 150{,}000 \times 16{,}87 + 50{,}000 + 170{,}000$

$= 3{,}545{,}000$

First or second or

(A and B) or ((C and D) or (E and F))

10,000 40,000 120,000

second or

$10{,}000 \log_2 20{,}000 + 40{,}000 \log_2 80{,}000 + 120{,}000 \log_2 150{,}000 + (40{,}000 + 120{,}000) + (160{,}000 +$

$10{,}000) = 3{,}320{,}000$

(A and E) or (B and E)

$10{,}000 \log_2 \frac{120{,}000}{10{,}000} + 20{,}000 \log_2 \frac{120{,}000}{20{,}000} + (10{,}000 + 20{,}000) = 540{,}000$

(A and B) or E

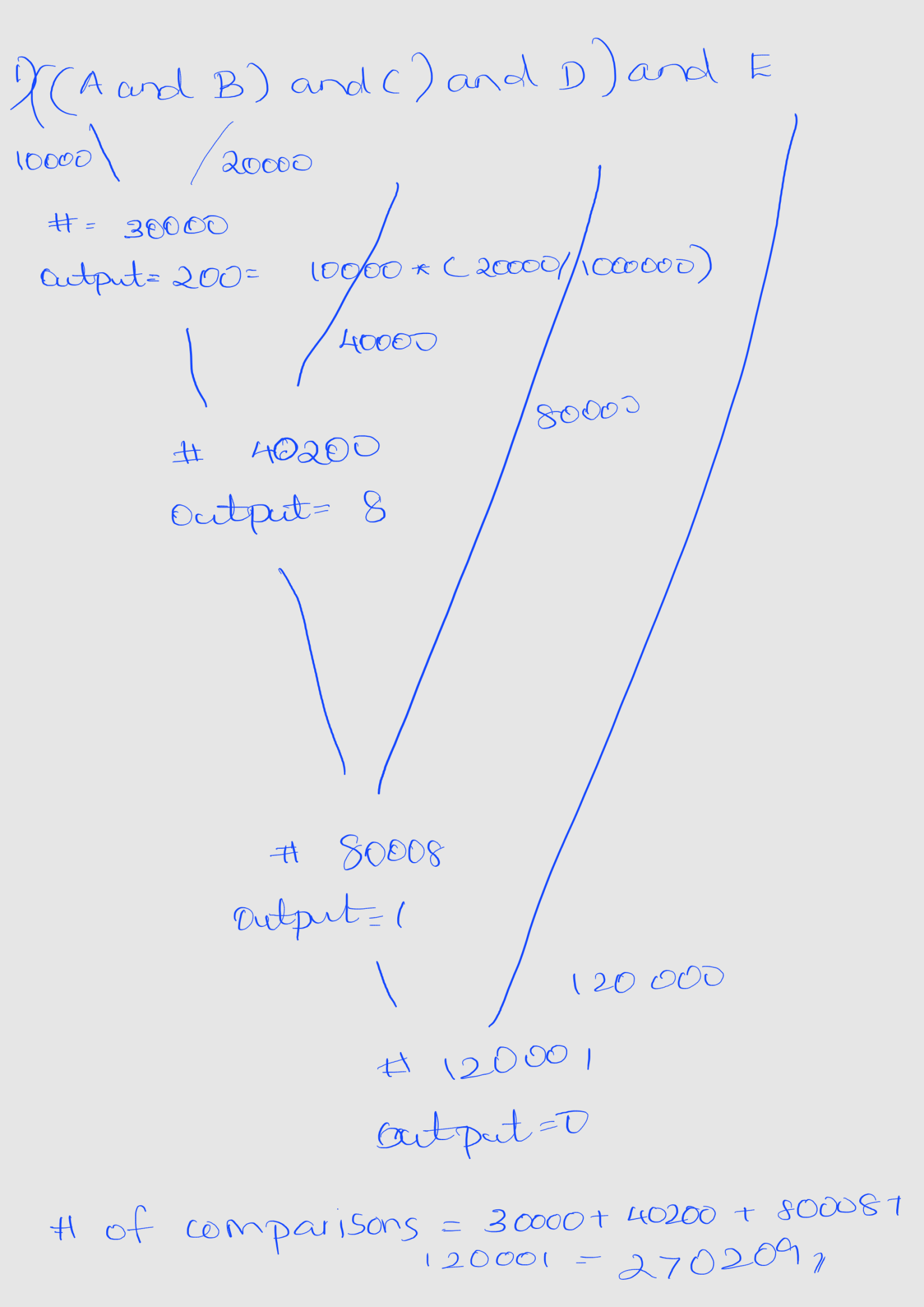
$10{,}000 \log_2 \frac{20{,}000}{10{,}000} + (10{,}000 + 120{,}000) = 280{,}000$

2. Assuming mutual independence between occurences

For this part we used formula for independence probablity.

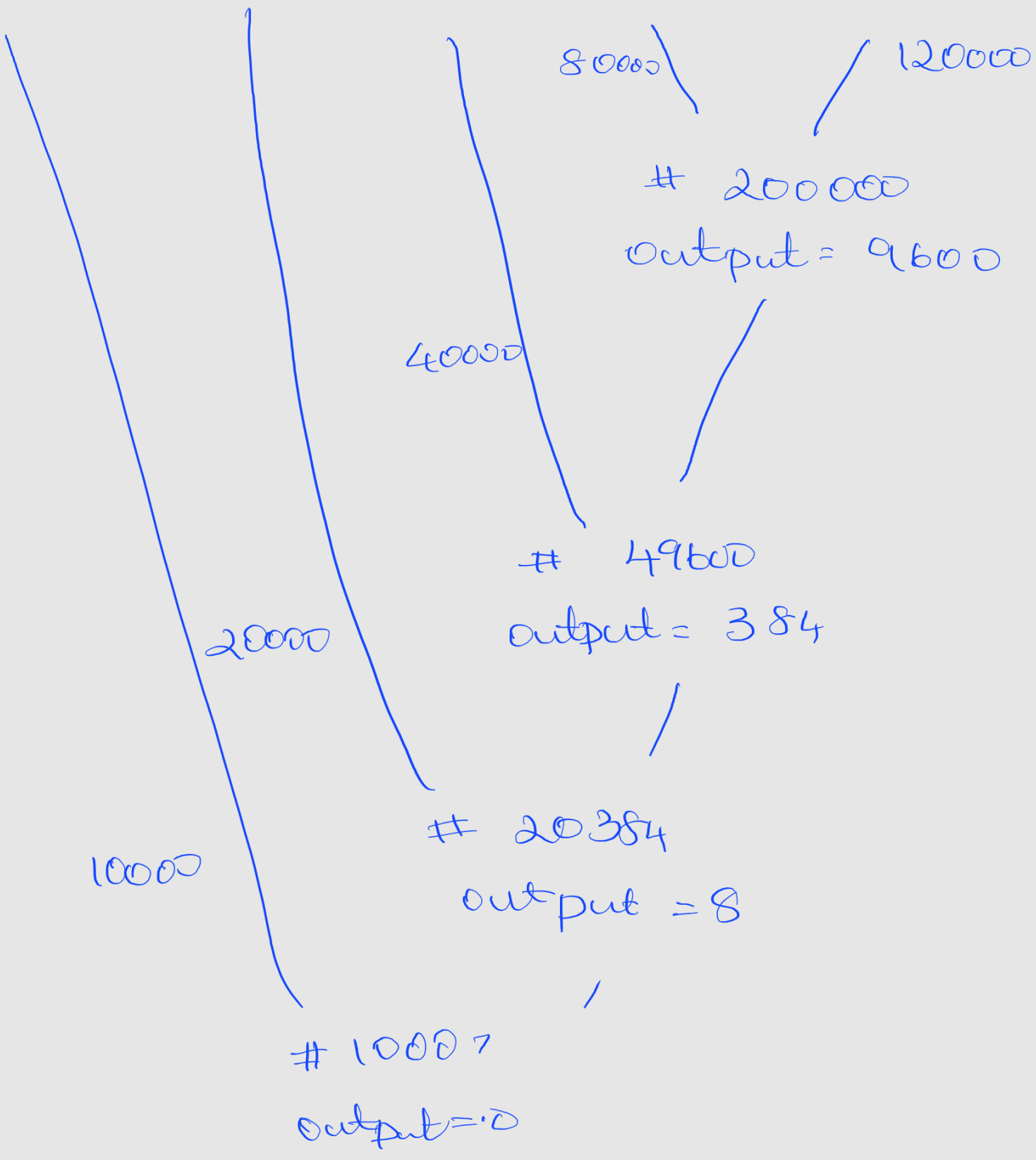
$P(A \cap B) = P(A) \cdot P(B)$

And to get expected length of the list, we multiply the above probablity by the total number of words.



# of comparisons = 30000 + 40200 + 80008 + 120001 = 270209

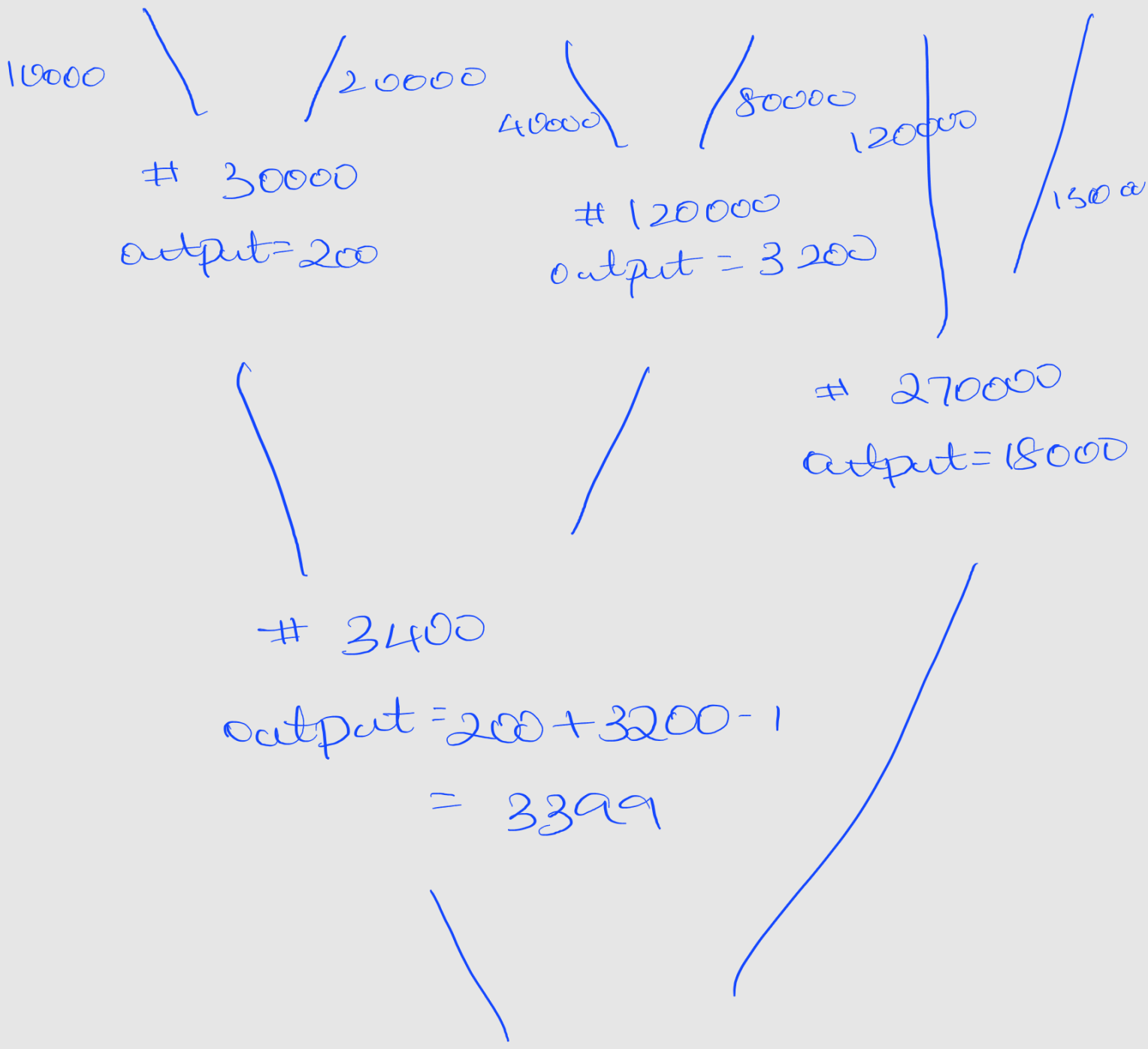
2) A and (B and (C and (D and E)))



no. of comparisons = 200000 + 49600 + 20384 + 10007

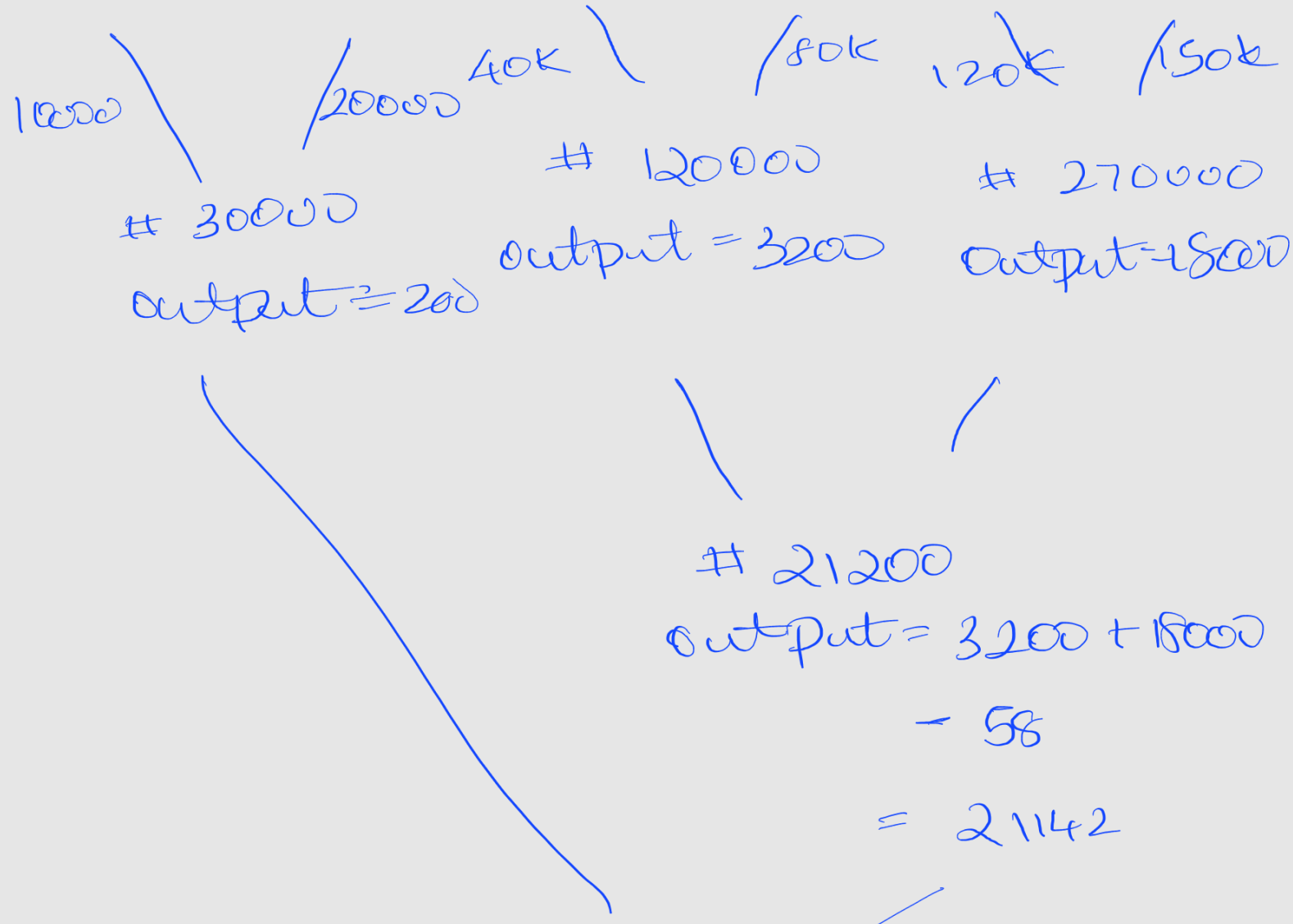
= 279991

3) ((A and B) or (C and D)) or (E and F)



= 30000 + 120000 + 270000 + 3400 + 273399 = 696799 -(U but it'll be small)

4. (A and B) or ((C and D) or (E and F))



= 30000 + 120000 + 270000

+ 21200 + 21342

= 462542 //

5. (A and E) or (B and E)

10K \ / 120K 20K \ / 120K

130K # 40K

output = 1200 output = 2400

\ /

= 3600

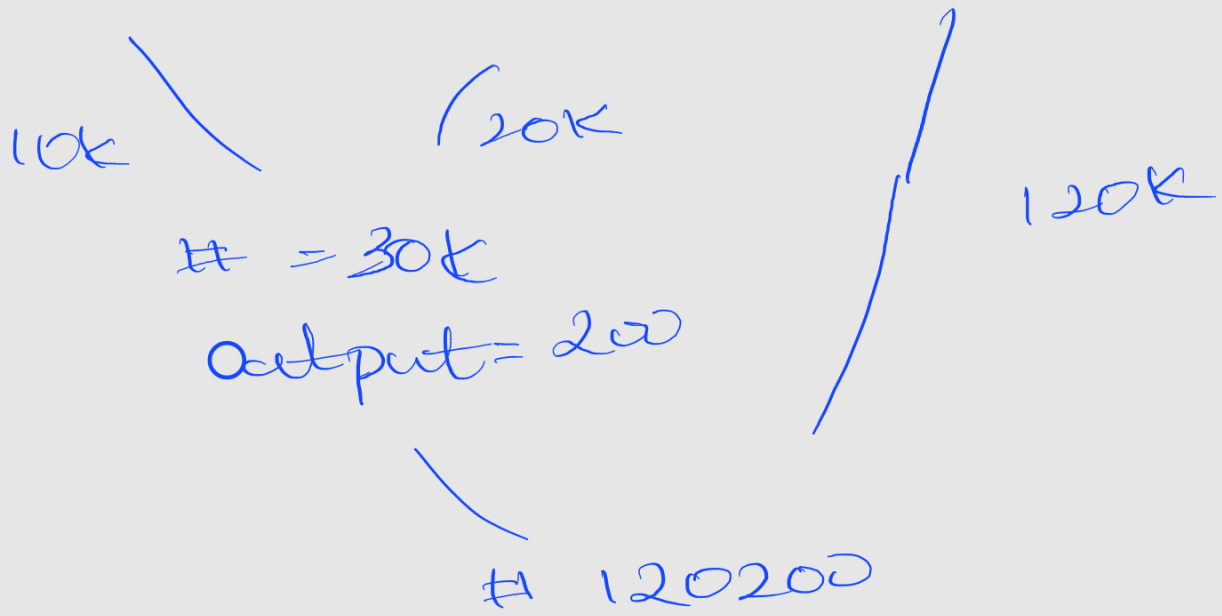
output = 3600 - 3

= 3797

= 130000 + 140000 + 3600

= 273600 //

6) (A and B) or E



₦ = 30000 + 120200

= 150200//