BIRZEIT UNIVERSITY
Electrical and Computer Engineering Department

ENCSENCS5342: “Information Retrieval, Web Search and NLP”

Assignment #1: Instructor: Dr. Adnan H. Yahya,

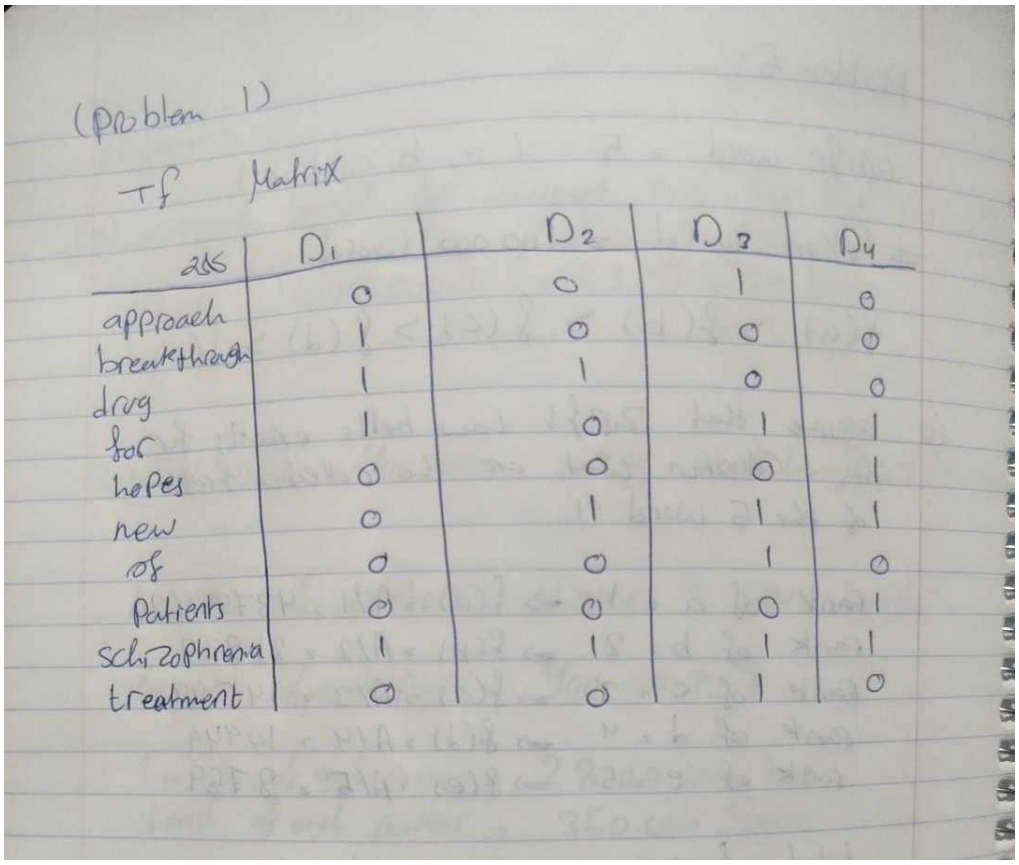
Leyan Burait 1211439

Due date: May 25, 2025, strongly encouraged to submit before the exam (5/5/2025).

Problem1. Create the term frequency matrix for the four document titles below:

```
Doc_1 = "breakthrough drug for Schizophrenia"
Doc_2 = "new schizophrenia drug"
Doc_3 = "new approach for treatment of schizophrenia"
Doc_4 = "new hopes for Schizophrenia patients"
```

Retain every word (no stopword removal). The indexed units should be words converted to lower case, no stemming. The entry in each cell should be the number of occurrences of the term in the associated document.

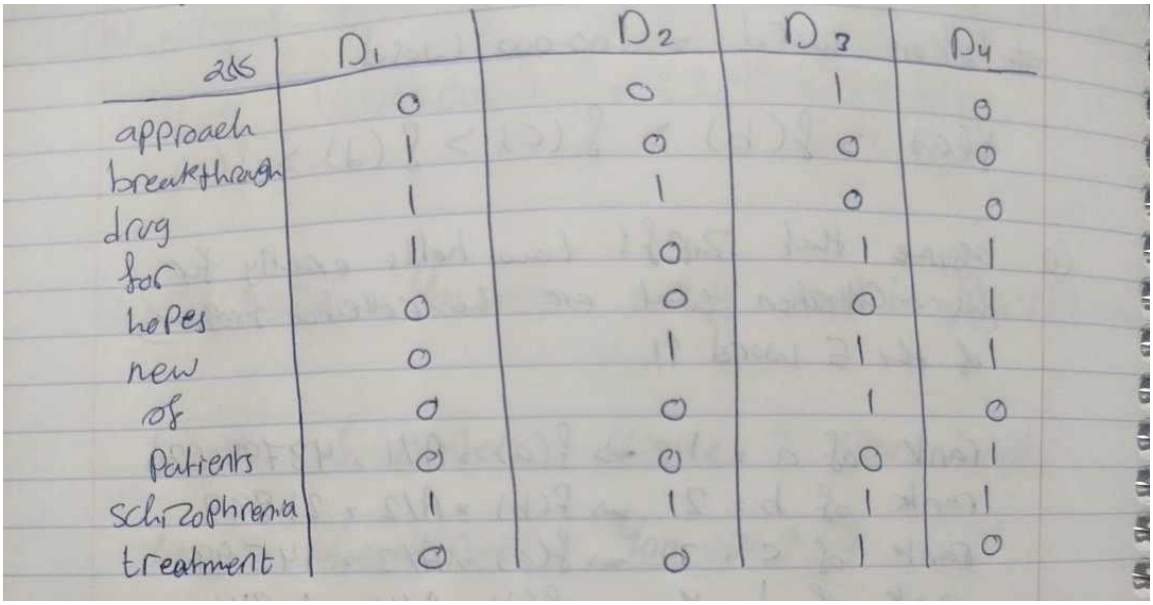


(Problem 1)

Tf Matrix

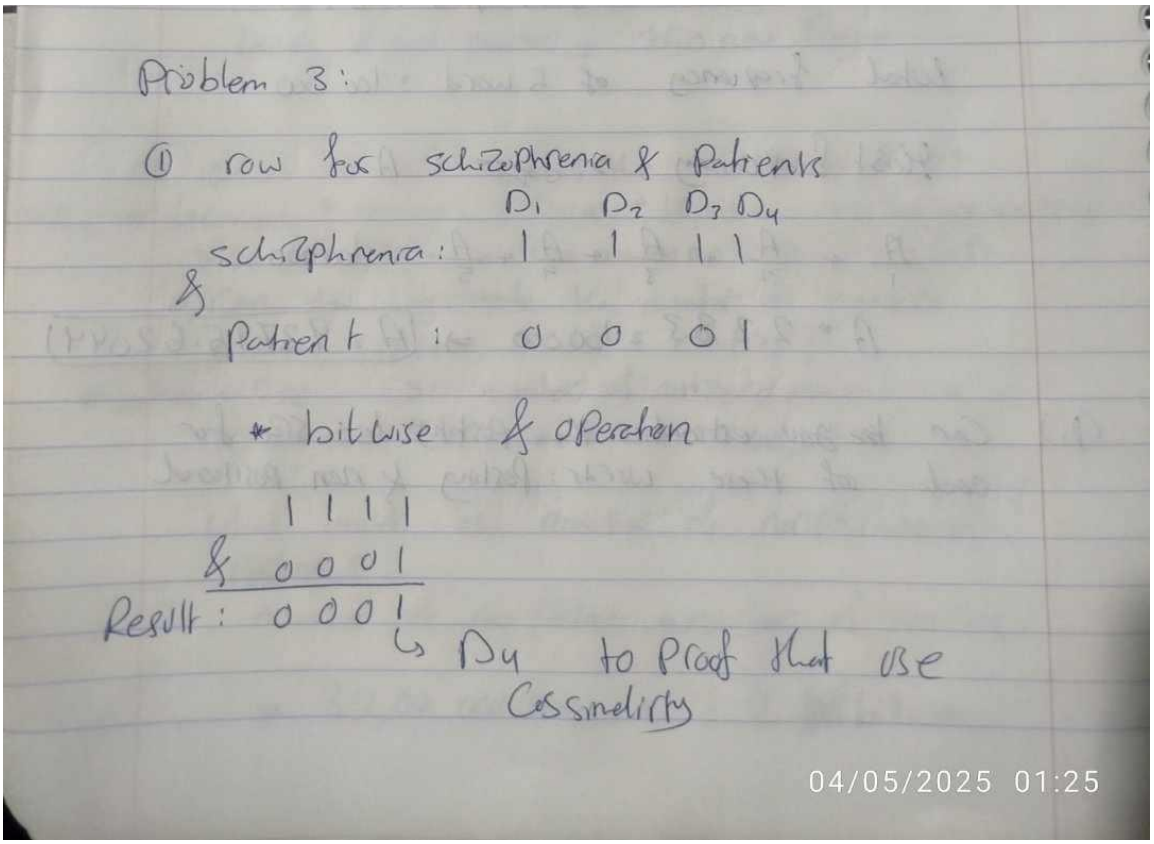
Docs	D_1	D_2	D_3	D_4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
hopes	0	0	0	1
new	0	1	1	1
of	0	0	1	0
Patients	0	0	0	1
Schizophrenia	1	1	1	1
treatment	0	0	1	0

Problem2. Create a new incidence matrix from the matrix you prepared in problem 1 (1s and 0's only).



Docs	D_1	D_2	D_3	D_4
approach	0	0	1	0
breakthrough	1	0	0	0
drug	1	1	0	0
for	1	0	1	1
hopes	0	0	0	1
new	0	1	1	1
of	0	0	1	0
Patients	0	0	0	1
Schizophrenia	1	1	1	1
treatment	0	0	1	0

Problem3. Extract the rows for the terms "schizophrenia" and " patients" from the matrix and write them down, one above the other. Each should have four binary digits (0 or 1) in them. Then draw a horizontal line under them and compute the AND of the two rows. This should produce a single row in which a 1 appears for each document that has both "schizophrenia" AND " patients". What are these documents? Show your work.



Q : schizophrenia & patients

Q : [0, 0, 0, 0, 0, 0, 0, 1, 1, 0]

D_4 : [0, 0, 0, 1, 1, 1, 0, 1, 1, 0]

$$\text{Cos similarity} = \frac{Q \cdot D_4}{|Q||D_4|}$$

$Q \cdot D_4 = 1*1 + 1*1 + 0 = 2$

$|Q| = \sqrt{(1)^2 + (1)^2} = \sqrt{2} = 1.4$

$|D_4| = \sqrt{(1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2} = \sqrt{5} = 2.236$

$$\text{Cos similarity}(Q, D_4) = \frac{2}{3.13} = \boxed{0.6388}$$

Q with D_3

D_3 = [1, 0, 0, 1, 0, 1, 1, 0, 1, 1]

$$\text{cos} = \frac{Q \cdot D_3}{|Q||D_3|} = \frac{1}{1.4 * 2.449} = \frac{1}{3.4292} = \boxed{0.2916}$$

$|D_3| = \sqrt{6 * (1)} = 2.449$

Q with D_2

D_2 = [0, 0, 1, 0, 0 1, 0, 0, 1, 0]

$|D_2| = \sqrt{3} = 1.73$

04/05/2025 01:26

Cos Similarity $(Q, D_2) = \frac{Q \cdot D_2}{|Q|\,|D_2|}$

$Q \cdot D_2 = 1$
$|Q| = 1.4$
$|D_2| = 1.73$

$= \frac{1}{1.4 * 1.73} = \frac{1}{2.424}$

$Cos(Q, D_2) = \boxed{0.412}$

Cos similarity $(Q, D_1) = \frac{Q \cdot D_1}{|Q|\,|D_1|}$

$D_1 = [0, 1, 1, 1, 0, 0, 0, 0, 1, 0]$

$|D_1| = \sqrt{4} = 2$
$|Q| = 1.4$

$\Rightarrow Cos(Q, D_1) = \frac{1}{2 * 1.4}$

$Q \cdot D_1 = 1$

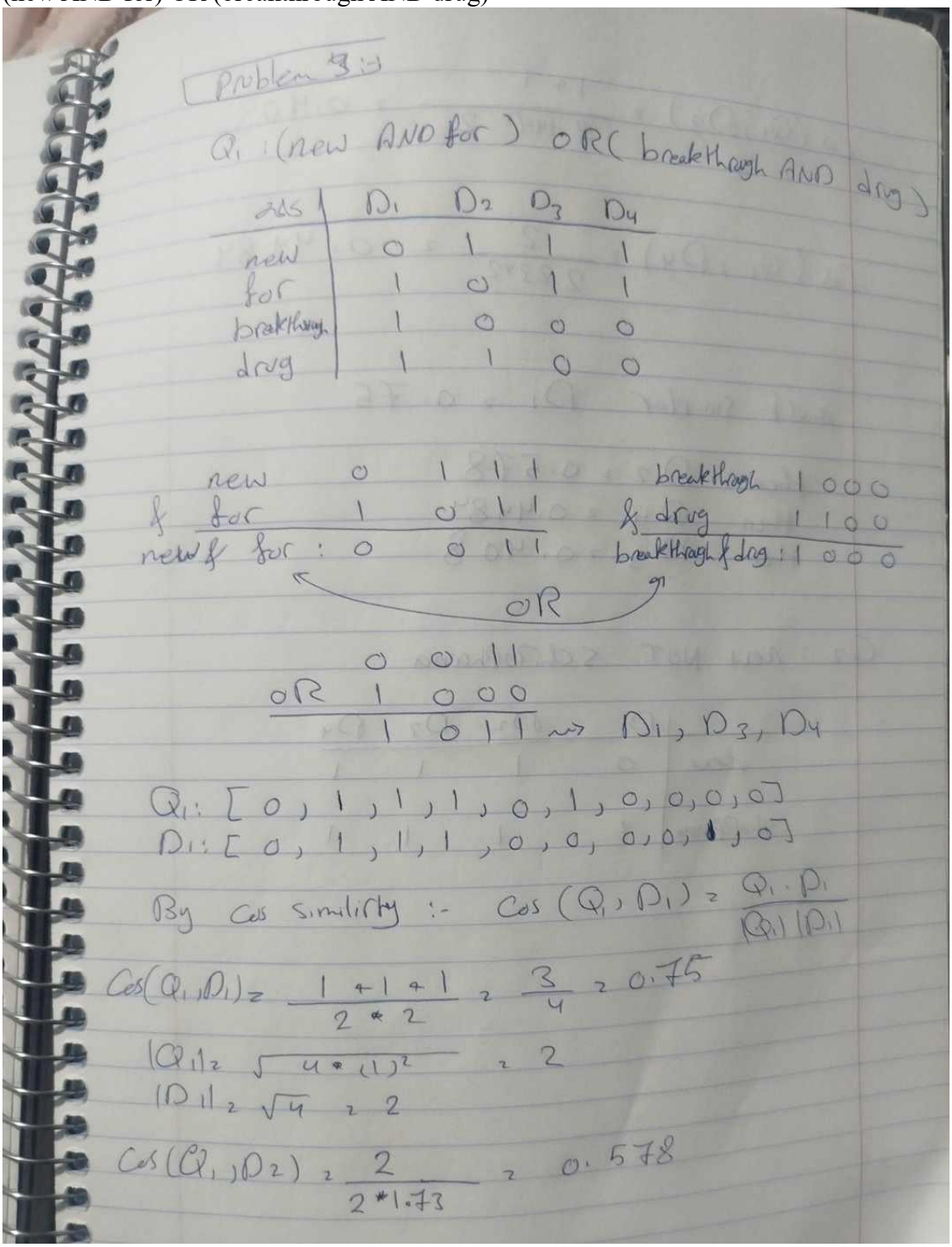
$\boxed{Cos(Q, D_1) = 0.357}$

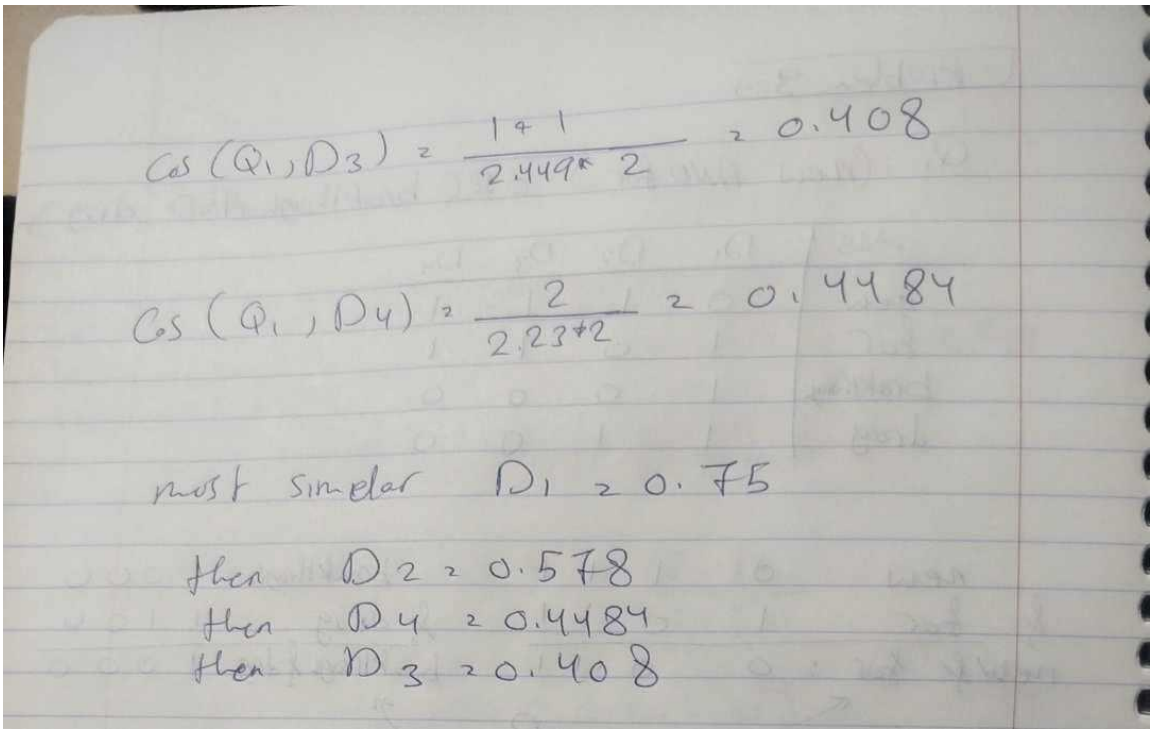
Most similarity $\boxed{Q \text{ with } D_4 = 0.6388}$

Then Q with $D_2 = 0.412$
then Q with $D_1 = 0.357$
then Less similarity Q with $D_3 = 0.2916$

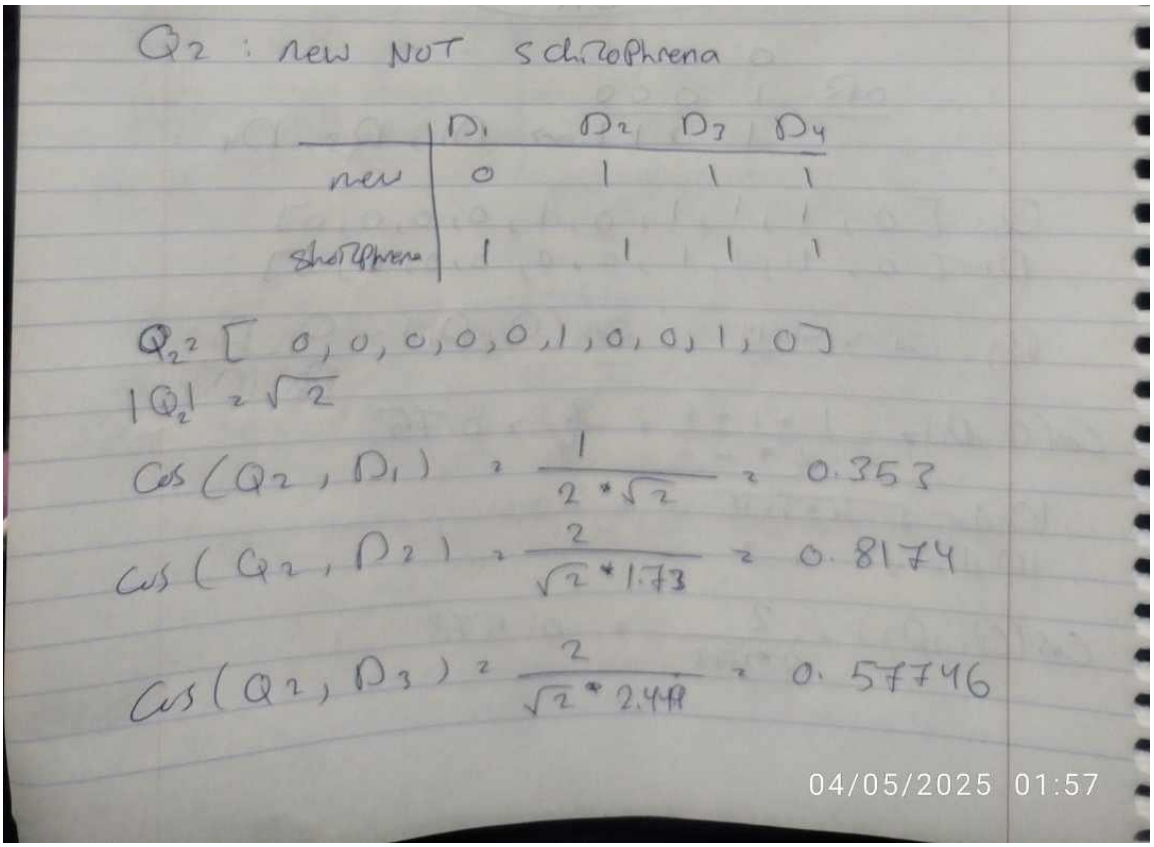
4. Perform the same computation for the following queries:

- (new AND for) OR (breakthrough AND drug)





- (new NOT schizophrenia)



$$\cos(Q_2, D_4) = \frac{2}{\sqrt{2} * 2.236} = 0.632$$

Q_2 Similty most with $D_2 = 0.8174$

then Q_2 with 0.632 D_4

then Q_2 with $D_3 = 0.57746$

then Q_2 with $D_1 = 0.353$

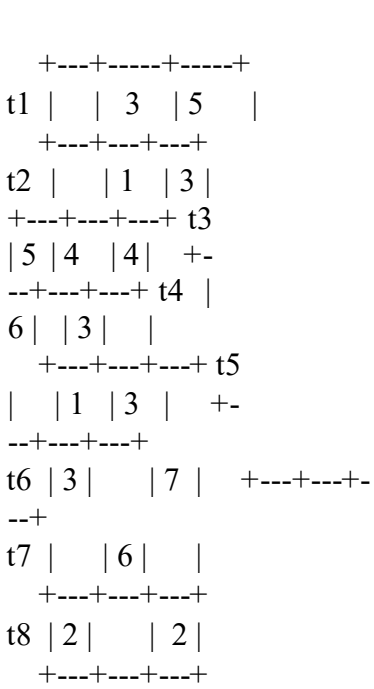
Matrix with Posting List

DS	Posting List (DocIDs)
approach	3
breakthrough	1
drug	1, 2
for	1, 3, 4
hopes	4
new	2, 3, 4
of	3
Patients	4
schizophrenia	1, 2, 3, 4
treatment	3

04/05/2025 01:59

5. Draw the matrix with posting lists.

Problem4. For the following questions, use the term-by-document matrix provided below to perform vector space retrieval (blanks indicate zeros):



1. Build the w matrix in which each element is computed as TF * IDF, where TF is what is specified in each element above and the IDF for term i is computed as:

2. Term-Document Matrix with TF-IDF weights:

Term	D1	D2	D3
T1	0	1.755	2.925
T2	0	0.585	1.755
T3	0	0	0
T4	3.51	1.755	0
T5	0	0.585	1.755
T6	1.755	0	4.095

Term	D1	D2	D3
T7	0	9.51	0
T8	1.17	0	1.17

- IDF Calculations (for N=3N=3 documents):
    - Terms t1,t2,t7$t1$,$t2$,$t7$: df=1df=1 → IDF=log$f_0$2(3/1)≈1.585IDF=log2(3/1)≈1.585
    - Terms t4,t5,t6,t8$t4$,$t5$,$t6$,$t8$: df=2df=2 → IDF=log$f_0$2(3/2)≈0.585IDF=log2(3/2)≈0.585
    - Term t3$t3$: df=3df=3 → IDF=log$f_0$2(3/3)=0IDF=log2(3/3)=0

Similarity Scores Calculation:

- For each document, sum the TF-IDF weights of the query terms t2$t2$ and t7$t7$:
    - d1$d1$: t2=0+t7=0=0$t2$=0+$t7$=0=0
    - d2$d2$: t2=0.585+t7=9.510≈10.095$t2$=0.585+$t7$=9.510≈10.095
    - d3$d3$: t2=1.755+t7=0≈1.755$t2$=1.755+$t7$=0≈1.755

log (total_number_of_documents / number_of_documents_containing_term_i). **Use base 2 logarithms**.

Compute the rank order of the documents that would be found using the vector space method for the UNWEIGHTED query (terms are of equal importance): t2 t7

Ranking:

1. d2$d2$ (Score ≈10.095≈10.095)
2. d3$d3$ (Score ≈1.755≈1.755)
3. d1$d1$ (Score =0=0)

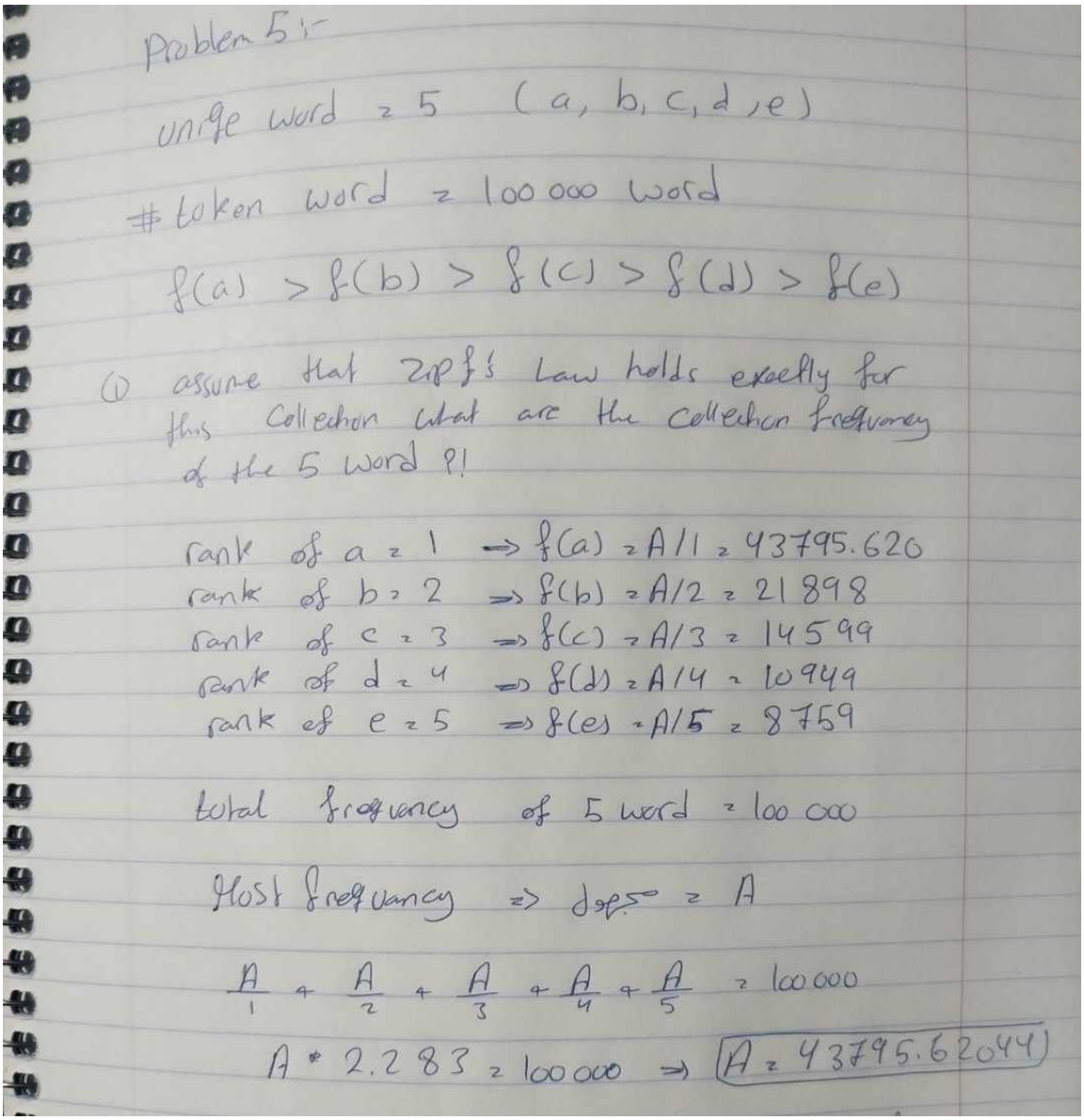
Final Ranked List:
d2>d3>d1*d2>d3>d*1

In your answer, give the similarity score that is computed for each document and give the ranked list that the system would return.

Problem 5.

Given a collection that contains 5 different words a, b, c, d,e only.
The frequency order is f(a) >f(b) >f(c)>f (d) >f (e). The total number of tokens in the collection is 100000.

2-a. Assume that Zipf's law holds exactly for this collection. What are the collection frequencies of the 5 words?



2-b. Can you estimate the postings list size for each of these words: positional and nonpositional.

- 5 words: a, b, c, d, e
- Word frequencies according to Zipf's law:
  - a: 100,000
  - b: 50,000
  - c: 33,333
  - d: 25,000
  - e: 20,000
- Total tokens in collection: 100,000

Necessary Assumptions:

1. Number of documents (N): Assuming average document length of 100 words
   N = 100,000/100 = 1,000 documents

2. Document ID size: 10 bits (to represent 1,000 documents)
3. Position index size: 8 bits (to represent word positions within documents)

1. Non-positional Postings Lists:

Each entry contains:

- Only document ID (10 bits)

Estimating number of documents containing each term (df):

- Word a: Appears in nearly all documents (df ≈ 1,000)
- Word b: Appears in ≈ 500 documents
- Word c: Appears in ≈ 333 documents
- Word d: Appears in ≈ 250 documents
- Word e: Appears in ≈ 200 documents

List sizes:

List size = df × 10 bits

- a: 1,000 × 10 = 10,000 bits ≈ 1.25 KB
- b: 500 × 10 = 5,000 bits ≈ 0.625 KB
- c: 333 × 10 ≈ 3,330 bits ≈ 0.416 KB
- d: 250 × 10 = 2,500 bits ≈ 0.312 KB
- e: 200 × 10 = 2,000 bits ≈ 0.25 KB

2. Positional Postings Lists:

Each entry contains:

- Document ID (10 bits)
- Number of positions (variable)
- Actual positions (8 bits per position)

Estimating positions per word:

- Assuming each word appears once on average in documents where it occurs

List sizes:

List size ≈ df × (10 + 8 × average frequency per document)

- a: 1,000 × (10 + 8×100) ≈ 1,000 × 810 = 810,000 bits ≈ 101 KB
- b: 500 × (10 + 8×50) = 500 × 410 = 205,000 bits ≈ 25.6 KB
- c: 333 × (10 + 8×33) ≈ 333 × 274 ≈ 91,242 bits ≈ 11.4 KB
- d: 250 × (10 + 8×25) = 250 × 210 = 52,500 bits ≈ 6.56 KB
- e: 200 × (10 + 8×20) = 200 × 170 = 34,000 bits ≈ 4.25 KB

Problem 6.

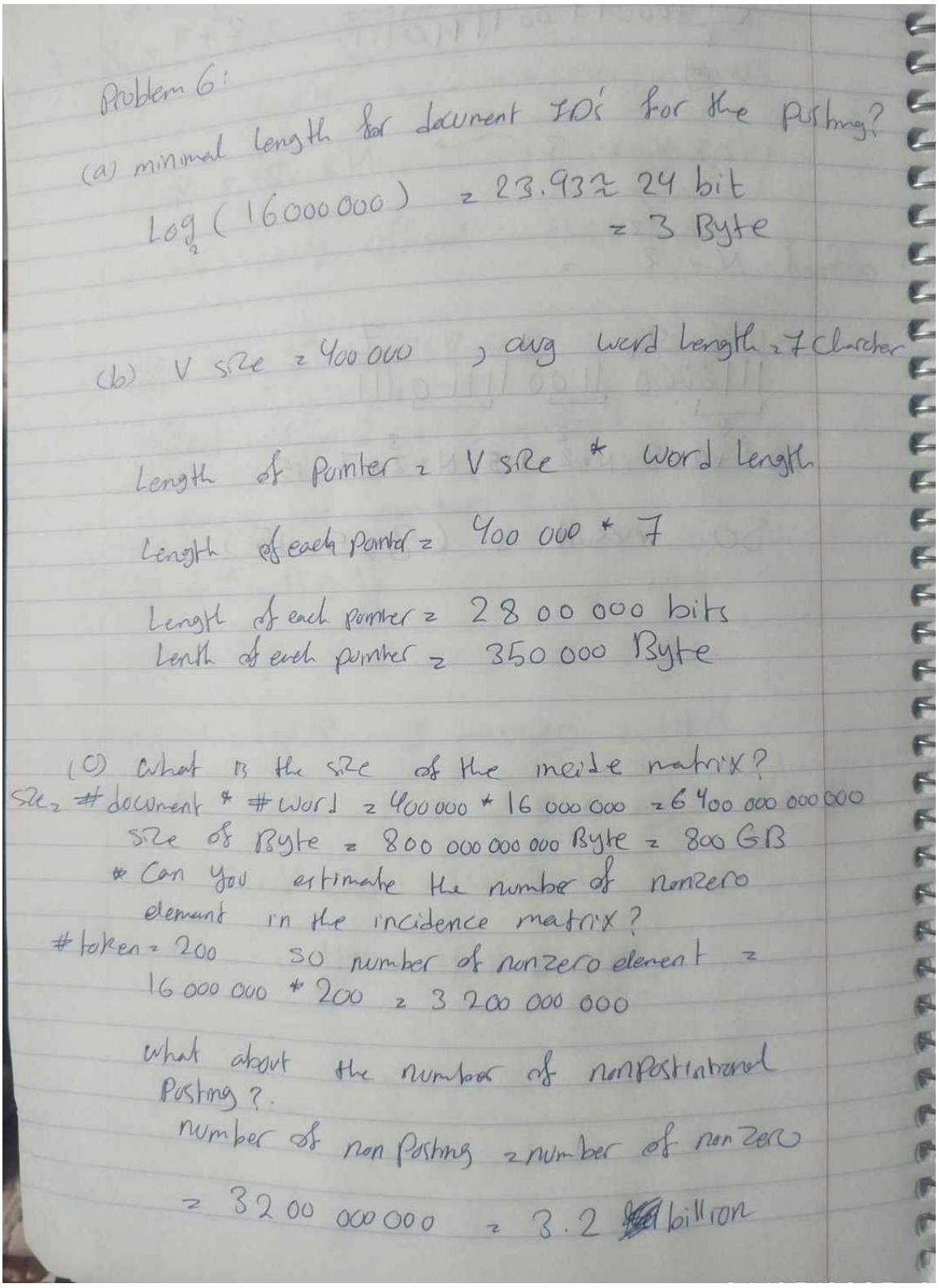
2-a. A search engine has a collection of 16,000,000 pages (documents) with 200 tokens per page, on average.

What is the minimal length for document IDs for the postings? In bits and in full bytes.

2-b. If the vocabulary size is 400,000, and the average dictionary word length is 7 characters.

How many bits do you need for pointers if one is to store the dictionary as a single string with pointers to the start of each word (what is the length of each pointer).

2-c. What is the size of the incidence matrix? Can you estimate the number of nonzero elements in the incidence matrix? What about the number of nonpositional postings?



Problem 6:

(a) minimal length for document IDs for the Posting?

$\log_2(16000000) = 23.93 \approx 24$ bit

= 3 Byte

(b) V size = 400 000 , avg word length = 7 character

Length of Pointer = V size * word length

Length of each Pointer = 400 000 * 7

Length of each pointer = 2800 000 bits

Lenth of each pointer = 350 000 Byte

(c) what is the size of the meide matrix?

Size = # document * # word = 400 000 * 16 000 000 = 6 400 000 000 000

size of Byte = 800 000 000 000 Byte = 800 GB

* Can you estimate the number of nonzero element in the incidence matrix?

token = 200 so number of nonzero element =

16 000 000 * 200 = 3 200 000 000

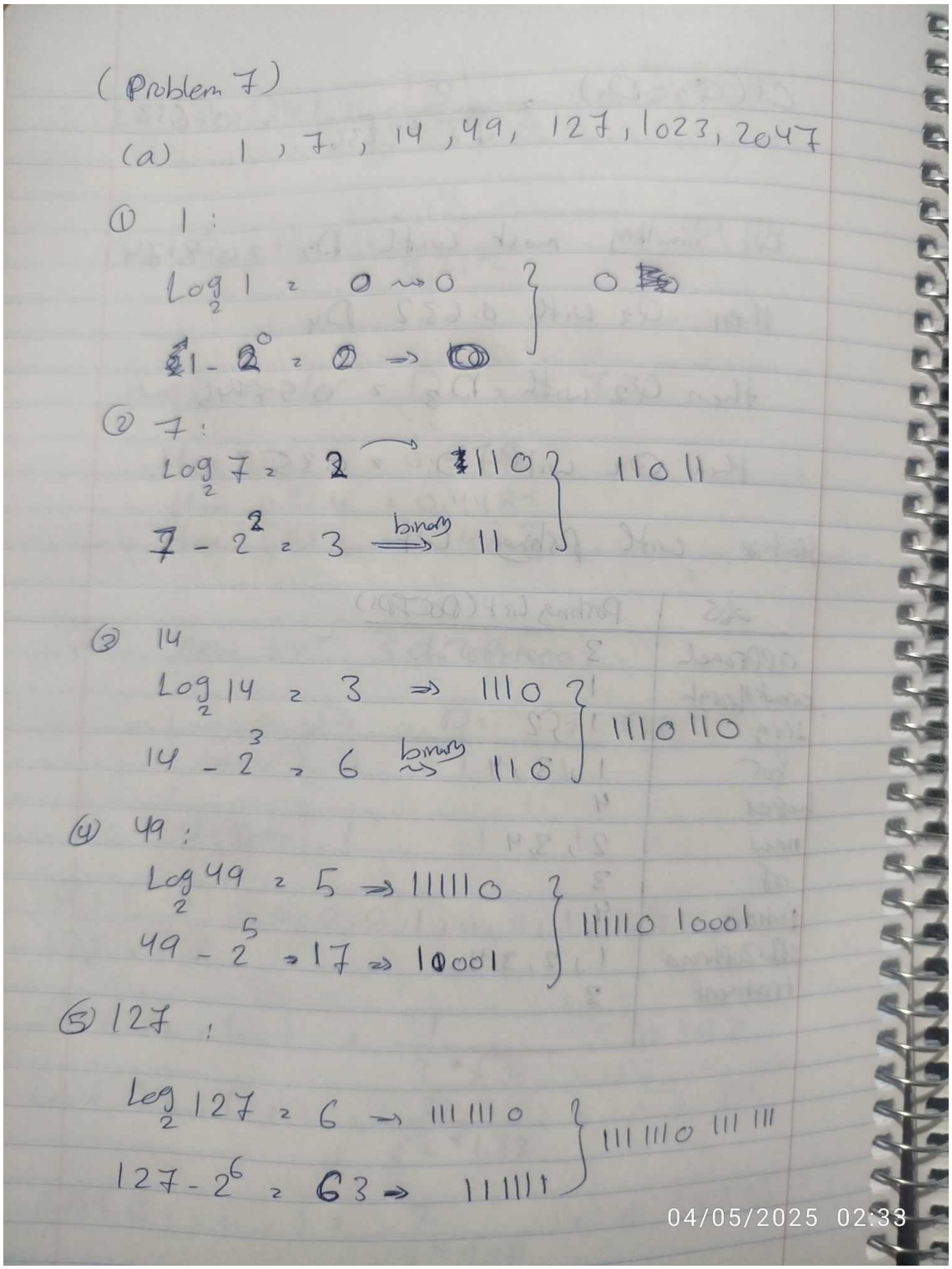
what about the number of nonPostintrand Posting?.

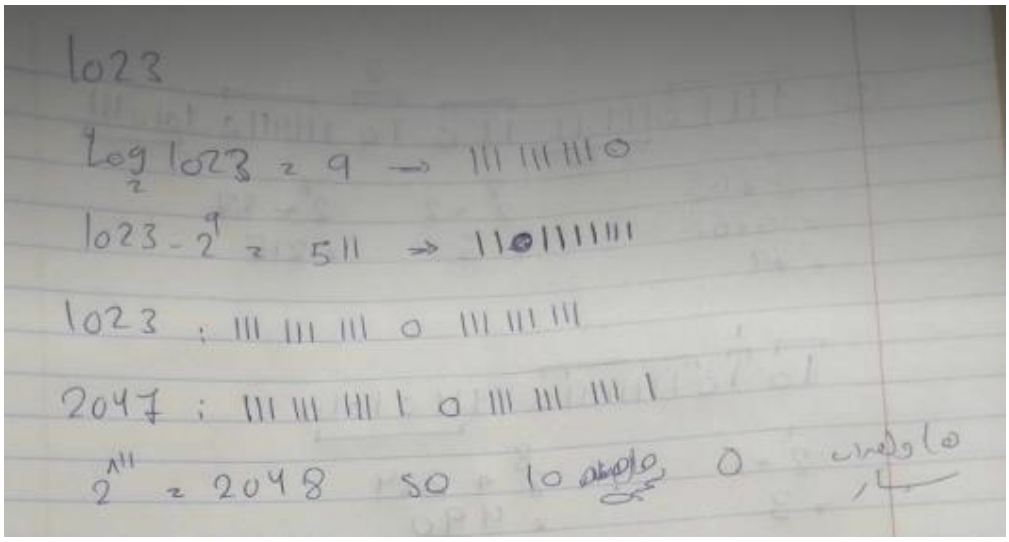
number of non Posting = number of non zero

= 3200 000 000 = 3.2 billion

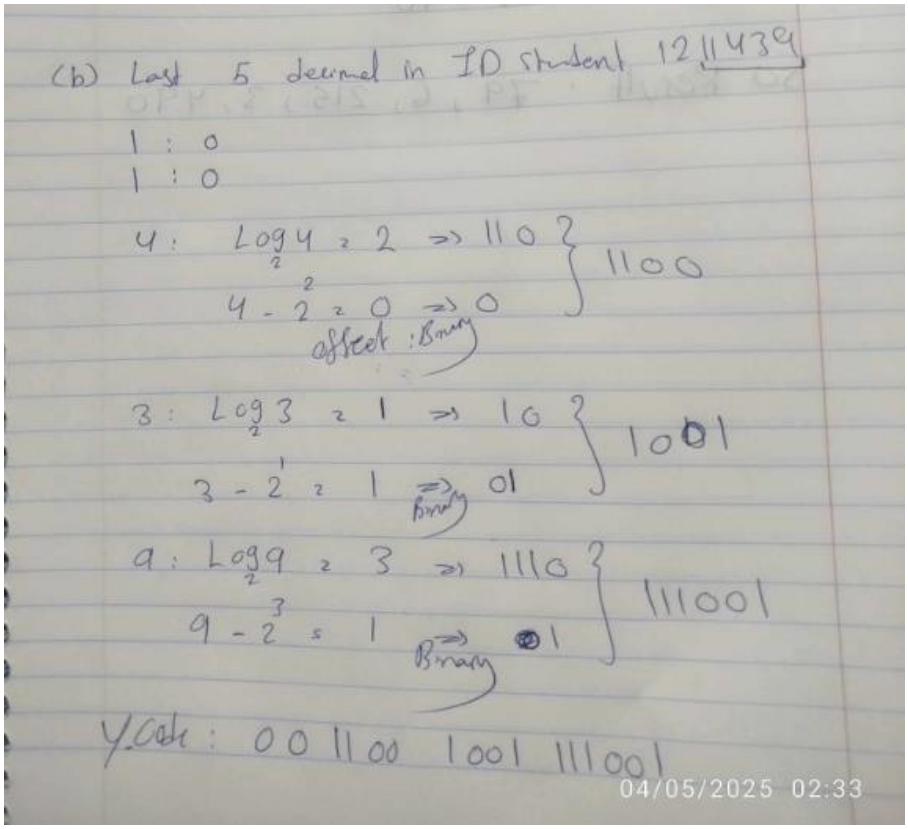
Problem 7.

3-a. Compute the γ-code for the sequence of decimals 1, 7, 14, 49, 127, 1023, 2047.



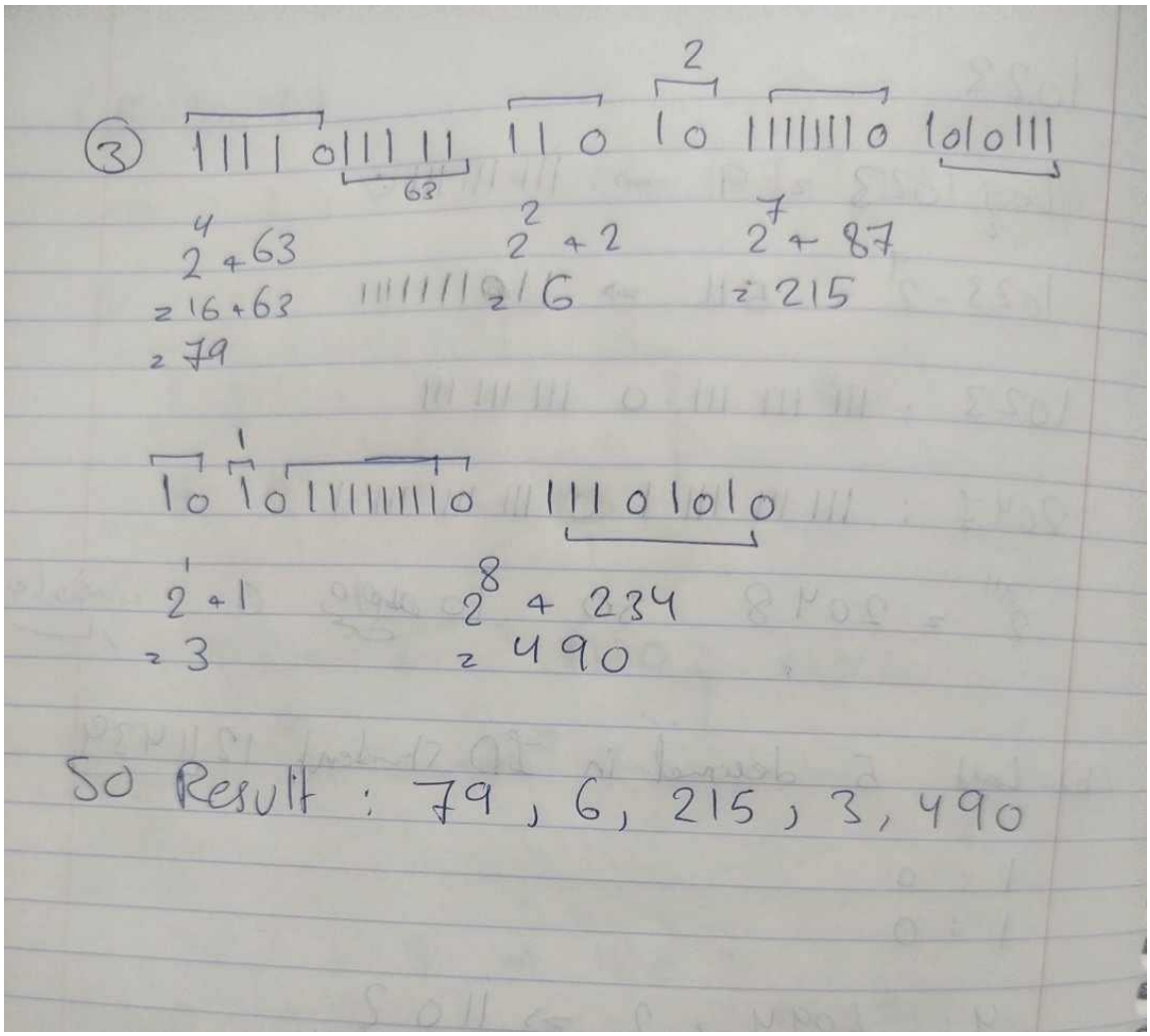


3-b. Represent the least significant 5 decimal digits of your ID number in γ-code.



3-c. Recover the gap value in decimal for the following string representing a sequence of gaps in a posting list, if possible.

1111011111110101111111010101111010111111111011101010



Good Luck