1. You are using a boolean IR system with an inverted file index containing the following positional information:

cold, 2:	[1, 1 :< 6 >; 4, 1 :< 4 >]
days, 1:	[3, 1 :< 2 >]
eat, 1:	[6, 1 :< 1 >]
hot, 2:	[1, 1 :< 3 >; 4, 1 :< 8 >]
in, 3:	[2, 1 :< 3 >; 4, 2 :< 1,5 >]
lot, 1:	[6, 1 :< 3 >]
nine, 1:	[3, 1 :< 1 >]
old, 1:	[3, 1 :< 3 >]
peas, 5:	[1, 2 :< 1,4 >; 2, 1 :< 1 >; 5, 2 :< 1,3 >]
porridge, 5:	[1, 2 :< 2,5 >; 2, 1 :< 2 >; 5, 2 :< 2,4 >]
pot, 3:	[2, 1 :< 5 >; 4, 2 :< 3,7 >]
the, 4:	[2, 1 :< 4 >; 4, 2 :< 2,6 >; 6, 1 :< 2 >]

   (a) Which of the terms have multiple occurrences within the same document?

   in, peas, porridge, pot, the

   (b) Which documents would the following query return?
   eat OR (porridge /1 hot)

   6, 1

2. You are testing a program implementing the (Levenshtein) edit distance algorithm. It outputs the following table which has errors. Identify the cells in the table that contain wrong entries and provide the correct entries for them.

	‘’	l	e	t
‘’	0	1	2	4
c	1	1	2	3
a	2	2	2	3
t	3	2	2	1

	‘’	l	e	t
‘’	0	1	2	3
c	1	1	2	3
a	2	2	2	3
t	3	3	3	2

3. You want to compress the dictionary terms stored in the leaf pages of a B-tree. Apply front coding to the following list of terms:

   - Jalaun
   - Jalna
   - Jalpaiguri
   - Jamaica
   - Jamal
   - Jamb

word	front coding
Jalaun	0,6 Jalaun
Jalna	3,2 na
Jalpaigur	3,7 paiguri
Jamaica	2,5 maica
Jamal	4,1 l
Jamb	3,1 b

4. An IR system uses the γ-code to encode postings lists. The following bit-string is returned by an inverted file index as a gap encoded list of DocIDs.

   1110001111101011110001111100100

   Is this a sequence properly encoded in γ-code? If yes, write down the decoded list of DocIDs. If no, briefly explain why this bit-string is not a correct γ-code.

Calculate the offset:

[1110] 001	[11110] 1011	[110] 00	[111110] 0100
1110 = 3	11110 = 4	110 = 2	111110 = 5

Prefix with one:

1110 [001]	11110 [1011]	110 [00]	111110 [0100?]
1→001	1→1011	1→00	1→0100?
1001 = 9	11011 = 35	100 = 4	10100? = ?

The last γ-code offset was larger than the number of 0's or 1's.
This sequence is not properly encoded in γ-code.

5. Compute the cosine similarity between each pair of the following vectors, assuming that the elements of these vectors have already been weighted using TF-IDF.

$$d_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 1 \end{pmatrix} \qquad d_2 = \begin{pmatrix} 6 \\ 0 \\ 0 \\ 6 \\ 3 \end{pmatrix} \qquad d_3 = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 2 \end{pmatrix}$$

$$\begin{aligned} cos(d_1, d_2) &\approx \frac{(1 \cdot 6) + (2 \cdot 0) + (3 \cdot 0) + (1 \cdot 6) + (1 \cdot 3)}{\sqrt{1^2 + 2^2 + 3^2 + 1^2 + 1^2} \cdot \sqrt{6^2 + 0^2 + 0^2 + 6^2 + 3^2}} \\ &\approx \frac{6 + 0 + 0 + 6 + 3}{\sqrt{1 + 4 + 9 + 1 + 1} \cdot \sqrt{36 + 0 + 0 + 36 + 9}} \\ &\approx 0.417 = \frac{5}{12} = \frac{15}{36} = \frac{15}{4 \cdot 9} = \frac{15}{\sqrt{16} \cdot \sqrt{81}} \end{aligned}$$

$$\begin{aligned} cos(d_1, d_3) &\approx \frac{(1 \cdot 2) + (2 \cdot 1) + (3 \cdot 0) + (1 \cdot 0) + (1 \cdot 2)}{\sqrt{1^2 + 2^2 + 3^2 + 1^2 + 1^2} \cdot \sqrt{2^2 + 1^2 + 0^2 + 0^2 + 1^2}} \\ &\approx \frac{2 + 2 + 0 + 0 + 2}{\sqrt{1 + 4 + 9 + 1 + 1} \cdot \sqrt{4 + 1 + 0 + 0 + 0}} \\ &\approx 0.5 = \frac{1}{2} = \frac{6}{12} = \frac{6}{4 \cdot 3} = \frac{6}{\sqrt{16} \cdot \sqrt{9}} \end{aligned}$$

$$\begin{aligned} cos(d_2, d_3) &\approx \frac{(6 \cdot 2) + (0 \cdot 1) + (0 \cdot 0) + (6 \cdot 0) + (3 \cdot 2)}{\sqrt{6^2 + 0^2 + 0^2 + 6^2 + 3^2} \cdot \sqrt{2^2 + 1^2 + 0^2 + 0^2 + 1^2}} \\ &\approx \frac{12 + 0 + 0 + 0 + 6}{\sqrt{36 + 0 + 0 + 36 + 9} \cdot \sqrt{4 + 1 + 0 + 0 + 0}} \\ &\approx 0.66 = \frac{2}{3} = \frac{18}{27} = \frac{18}{9 \cdot 3} = \frac{18}{\sqrt{81} \cdot \sqrt{9}} \end{aligned}$$