IDA - Data Mining Quiz Questions

Question 1. Suppose Alice has a box of different coloured balls, containing nine red, six green and five blue. Every time she randomly picks a ball from the box she will relay the colour of the ball to her friend Bob using a communication channel that can transmit only binary values. We let X be a r.v of which colour ball Alice takes out.

i) Devise a simple coding scheme Alice can use to tell Bob what colour ball she picked out.

ii) Find the Entropy of X.

iii) Until recently, Alice has been picking balls with replacement. After doing this activity for a long time, She decided to go outside for a short rest and while she was away, her little brother found the box and decided to mischievously steal seven red balls. Alice returned not knowing that someone has tampered with her experiment and carried on as before. Now that the situation has changed, what is the 'penalty' of encoding the events using Alice's assumption in terms of the expected number of additional bits needed?

Question 2. Let P and Q be two discrete probability distributions over an event set A. Let $D_{KL}(P||Q)$ be the K-L divergence from P to Q as defined in the lectures.

i) Prove that $D_{KL}(P||Q) \geq 0$ and $D_{KL}(P||Q) = 0 \Leftrightarrow P = Q$.

ii) Show that $H(P) \leq \log_2 n$, where $n = |A|$.

Question 3. Suppose we have a 'mini-library' containing four documents d^1, d^2, d^3 and d^4. We have $|d^1| = |d^2| = 500$ and $|d^3| = |d^4| = 50$. Using the set of four terms[1] $\mathcal{T} = \{\text{Alpha, Bravo, Charlie, Delta}\}$, we want to transform each document into a vector of weights i.e.

$$d^i \mapsto \mathbf{x}^i = (x^i_A, x^i_B, x^i_C, x^i_D)^T.$$

The term frequency in each document is represented by the table below:

[1] the documents, of course, contain many other terms, but those were not selected for the term set $\mathcal{T}$

	Alpha	Bravo	Charlie	Delta
N_k^1	2	1	2	1
N_k^2	1	2	0	0
N_k^3	3	4	3	10
N_k^4	0	0	1	0

i) Which term do you expect to be of high importance? Explain your answer.

ii) Using TFIDF, transform each document into a vector of weights.

Question 4. We have a small library of seven documents $d^1, d^2, \ldots, d^7$, where the length of each of them is 500. Using the set of terms $\mathcal{T} = \{\text{Terminal, Linux}\}$, we find the term frequency in each document, represented by a table below:

	terminal	Linux
N_k^1	50	100
N_k^2	5	2
N_k^3	100	20
N_k^4	1	0
N_k^5	0	0
N_k^6	25	32
N_k^7	50	75

i) Perform Latent Semantic Analysis on the document set. First transform each document into a vector of weights and then apply PCA using these set of vectors by projecting onto the leading eigenvector.

ii) What concept could the leading eigenvector possibly represent?

iii) I want to search for a document that is a comprehensive guide to using the Linux operating system. Which document in the library closely matches what I'm looking for?