IDA - Data Mining Solutions

Question 1.

i) A possible coding scheme that can be used is:

$$\text{picking a red ball} - 00$$
$$\text{picking a green ball} - 10$$
$$\text{picking a blue ball} - 01$$

ii) The entropy of X is

$$H(X) = \frac{9}{20}\log_2\left(\frac{20}{9}\right) + \frac{6}{20}\log_2\left(\frac{20}{6}\right) + \frac{5}{20}\log_2\left(\frac{20}{5}\right)$$
$$= 1.539 \text{ (3 d.p)}$$

The value we calculated means that Bob should expect *on average* 1.539 bits of information when Alice transmits the realisation. In part i), we have shown that each possibility can be encoded with just two bits so this is not a surprise.

iii) Let P and Q be two discrete probability distributions on the r.v. X, where P is the correct distribution Alice uses (the new one, after some of the balls are taken) and Q the incorrect one Alice thinks is being used (not knowing that some balls have been removed). So,

$$P = \left\{\frac{2}{13}, \frac{6}{13}, \frac{5}{13}\right\},$$
$$Q = \left\{\frac{9}{20}, \frac{6}{20}, \frac{5}{20}\right\}.$$

We need compute the K-L divergence from P to Q to find the 'penalty'. So,

$$D_{KL}(P||Q) = \frac{2}{13}\log_2\left(\frac{\frac{2}{13}}{\frac{9}{20}}\right) + \frac{6}{13}\log_2\left(\frac{\frac{6}{13}}{\frac{6}{20}}\right) + \frac{5}{13}\log_2\left(\frac{\frac{5}{13}}{\frac{5}{20}}\right)$$
$$= 0.288 \text{ (3 d.p)}$$

Hence, on average the transmission will use 0.288 additional bits when using the "wrong" distribution.

Question 2.

i) For convenience, we let

$$D_{KL}(P||Q) = -\sum_{x\in A} P(x)\log_2\left(\frac{Q(x)}{P(x)}\right).$$

Also, we will work with the natural logarithm (log), which will not change anything as for $a \in \mathbb{R}$,

$$\log_2 a = \frac{\log a}{\log 2}.$$

Furthermore, it can be verified that for $a > 0$,

$$\log a \le a - 1, \tag{1}$$

with equality if (and only if) $a = 1$ (plot the graphs of $\log a$ and $(a-1)$!). Now,

$$D_{KL}(P||Q) \ge -\sum_{x\in A} P(x)\left(\frac{Q(x)}{P(x)} - 1\right) = -\sum_{x\in A} Q(x) + \sum_{x\in A} P(x) = -1+1 = 0.$$

If $D_{KL}(P||Q) = 0$, then by (1) this can only happen when

$$\log\frac{Q(x)}{P(x)} = \frac{Q(x)}{P(x)} - 1, \text{ for every } x \in A \Leftrightarrow Q(x) = P(x), \text{ for every } x \in A.$$

It follows that $D_{KL}(P||Q) = 0 \Leftrightarrow P = Q$ as the converse is trivial.

This inequality we have just proven has a special name called *Gibbs' inequality* and it has many applications in information theory.

ii) Let Q be a uniform distribution over X i.e. $Q(x) = \frac{1}{n}$ for every $x \in A$. Then by part i),

$$\begin{aligned}D_{KL}(P||Q) &= -\sum_{x\in A} P(x)\log_2\frac{1}{nP(x)}\\ &= -\left(\sum_{x\in A} P(x)\log_2\left(\frac{1}{P(x)}\right) + \sum_{x\in A} P(x)\log_2\left(\frac{1}{n}\right)\right)\\ &= -(H(P) - \log_2 n)\\ &\ge 0.\end{aligned}$$

It follows that $H(P) \le \log_2 n$.

Question 3.

i) Delta - it appears less frequently in the documents compared to the other terms. Also note that it shows up more frequently in d^3 which better 'represents' the document.

ii) The computed vector of weights for each document are[1]

$$\mathbf{x}^1 = (1.66 \times 10^{-3}, 8.3 \times 10^{-4}, 1.66 \times 10^{-3}, 2 \times 10^{-3})^T,$$
$$\mathbf{x}^2 = (8.3 \times 10^{-4}, 1.66 \times 10^{-3}, 0, 0)^T,$$
$$\mathbf{x}^3 = (0.0249, 0.0332, 0.0249, 0.2),$$
$$\mathbf{x}^4 = (0, 0, 8.3 \times 10^{-3}, 0)^T.$$

Question 4.

i) First, we transform each document into a set of weights using TFIDF:

$$\mathbf{x}^1 = (0.0222, 0.0971)^T,$$
$$\mathbf{x}^2 = (2.22 \times 10^{-3}, 1.94 \times 10^{-3})^T,$$
$$\mathbf{x}^3 = (0.0445, 0.0194)^T,$$
$$\mathbf{x}^4 = (4, 45 \times 10^{-4}, 0)^T,$$
$$\mathbf{x}^5 = (0, 0)^T,$$
$$\mathbf{x}^6 = (0.0111, 0.0311)^T,$$
$$\mathbf{x}^7 = (0.0222, 0.0728)^T.$$

Next, we find the co-variance matrix $\mathbf{C}$ of these data points and the eigenvalues and eigenvectors:

$$\mathbf{u}_1 = (-0.976, 0.216)^T, \ \lambda_1 = 1.72 \times 10^{-4}$$
$$\mathbf{u}_2 = (-0.216, -0.976)^T, \ \lambda_2 = 1.34 \times 10^{-3}$$

We will use the eigenvector that preserves the most variability in the data, which is $\mathbf{u}_2$. The projected data points are

$$\widetilde{\mathbf{x}}^1 = -0.0996,$$
$$\widetilde{\mathbf{x}}^2 = -0.00238,$$
$$\widetilde{\mathbf{x}}^3 = -0.0286,$$
$$\widetilde{\mathbf{x}}^4 = -9.62 \times 10^5,$$
$$\widetilde{\mathbf{x}}^5 = 0,$$
$$\widetilde{\mathbf{x}}^6 = -0.0327,$$
$$\widetilde{\mathbf{x}}^7 = -0.0759.$$

[1] logarithm is base 2

ii) One possible concept present in the documents can be '*learning how to use the command line in Linux*'.

iii) Let our query document be $\mathbf{d}_Q = (0, 1)^T$. Projecting $\mathbf{d}_Q$ onto the concept space, we get $\widetilde{\mathbf{d}}_Q = -0.976$. In fact, using cosine similarity, we see that, with the exception of d^5, every document appears to be a good match for what we want!