En esta hoja código significa código binario únicamente descifrable

1) Sean $0 \le p \le 1$ y $q = 1 - p$. Comprueba que la entropía de $\{a, b\}$ con $p_1 = p$, $p_2 = q$ es la mitad que la de $\{A, B, C, D\}$ con $p_1 = p^2$, $p_2 = p_3 = pq$, $p_4 = q^2$. ¿Cómo se generaliza este hecho?

2) Tenemos una moneda trucada en la que sale cara sólo el 30 % de las veces. ¿Qué tiene más entropía (cantidad de información), lanzar un dado una vez o lanzar la moneda trucada tres veces?

3) Prueba que la entropía de un conjunto con probabilidades p_i, cumple $H \le -\sum p_i \log q_i$ cualesquiera que sean $q_i \in \mathbb{R}^+$ con $\sum q_i \le 1$. Prueba también que la entropía es máxima en el caso de equidistribución (todas las probabilidades p_i iguales).

4) Calcula la entropía de una distribución binomial $B(4, 0.5)$. Explica cómo la definirías para una distribución continua de probabilidad y calcúlala para la normal $N(0, 1)$.

5) Consideramos el número de caras al tirar dos monedas. Encuentra el número medio de bits mínimo para almacenar los resultados de este experimento repetido cien veces y diseña un código que lo alcance.

6) En informática normalmente un carácter se codifica con un byte a través del código ASCII, usándose aproximadamente 40 posibilidades cuando se desprecian las mayúsculas y acentos y algunos símbolos poco comunes. En español, las vocales constituyen estadísticamente el 45.07 % de los caracteres de un texto. Usando sólo este dato, ¿hasta cuánto se debería poder comprimir un fichero de un mega de texto en minúsculas sin acentos?

7) Mejora el resultado del problema anterior sabiendo que la estadística de las probabilidades de las diferentes vocales es: $\mathtt{a} \to 0.1253$, $\mathtt{e} \to 0.1368$, $\mathtt{i} \to 0.0625$, $\mathtt{o} \to 0.0868$, $\mathtt{u} \to 0.0393$.

8) Escribe un ejemplo de un código definido en $S = \{s_1, s_2, s_3, s_4\}$ que no tenga la propiedad de prefijo.

9) Dibuja el árbol que corresponde a un código con $\text{Im}(C) = \{\mathtt{11}, \mathtt{000}, \mathtt{001}, \mathtt{101}\}$.

10) Supongamos cierto tipo de ficheros en los byte representan números entre 1 y 4 con probabilidades respectivas $p_1 = 0.35$, $p_2 = 0.34$, $p_3 = 0.30$, $p_4 = 0.01$. Halla su codificación de Huffman y el tamaño medio en bytes que alcanzará un fichero de 10^5 bytes tras la codificación. Compara el resultado con la cota inferior que da la entropía.

11) ¿Qué aspecto tiene el árbol que corresponde a la codificación de Huffman para un conjunto de 2^k elementos, todos ellos con la misma probabilidad? ¿Y qué aspecto tiene el árbol en la codificación de Huffman si las probabilidades p_i son proporcionales a los números de Fibonacci F_i? En este último caso, intenta dar una prueba de lo que te sugieren los ejemplos.

12) Da un ejemplo de un código sobre un conjunto de cuatro elementos tal que las longitudes de las codificaciones sean $l_1 = 1$, $l_2 = 2$, $l_3 = l_4 = 3$. ¿Es posible añadir un quinto elemento al conjunto sin modificar la codificación de los otros?

13) Supongamos que queremos comprimir ficheros binarios que tiene un 80 % de unos. Calcula la tasa de compresión que se obtendría con la codificación de Huffman cuando los bits se agrupan de dos en dos, es decir $S = \{00, 01, 10, 11\}$ con probabilidades respectivas 0.2^2, 0.16, 0.16, 0.8^2, y cuando se agrupan de tres en tres.

14) Escribe la lista de pares que corresponden a la codificación de la frase `tres tristes tigres` en LZ78. Obtén también la codificación con LZW de esta frase suprimiendo los espacios y empleando un diccionario inicial de seis letras: $0 \to$ `e`, $1 \to$ `g`, $2 \to$ `i`, $3 \to$ `r`, $4 \to$ `s`, $5 \to$ `t`.

15) ¿Que dice `(0,m)`, `(0,i)`, `(0,␣)`, `(1,a)`, `(4,␣)`, `(1,e)`, `(3,a)`, `(4,.)` codificado en LZ78?

16) Supongamos el algoritmo LZW funcionando bit a bit con el diccionario inicial formado por `0` y `1`. Calcula cuántos elementos tendría al final el diccionario cuando se comprime un fichero formado por 8000 bits, todos ellos iguales a cero.

17) Consideremos la cadena bits c_k obtenida al concatenar todas las de longitud 1, todas las de longitud 2 y así hasta de longitud k, todas ellas ordenadas. Por ejemplo, para $k = 3$ sería $c_3 =$ `01 00011011 000001010011100101110111`. Intuitivamente éste es el peor caso para la compresión LZ78. Calcula la longitud de c_k en bits y el número de pares de que constaría su codificación.

18) Para elegir al azar un número de 1 a $2n$ podemos decidir primero aleatoriamente si queremos que esté en el subconjunto de los pares o de los impares y después escoger un número al azar del subconjunto. Por tanto, si queremos que la entropía mida la cantidad de información es razonable exigir que cumpla $H\big(\frac{1}{2n}, \overset{2n \text{ veces}}{\dots}, \frac{1}{2n}\big) = H\big(\frac{1}{2}, \frac{1}{2}\big) + H\big(\frac{1}{n}, \overset{n \text{ veces}}{\dots}, \frac{1}{n}\big)$. Utiliza un argumento similar para justificar la tercera propiedad exigida por Shannon.

19) Un código óptimo (en cuanto a la longitud media) sobre un conjunto de n elementos equiprobables, cuando n no es una potencia de dos, necesariamente da codificaciones de longitud $[\log_2 n]$ para algunos elementos y de longitud $1 + [\log_2 n]$ para otros, donde $[\,\cdot\,]$ indica la parte entera. Prueba que los codificados con longitud mayor son exactamente $2\big(n - 2^{[\log_2 n]}\big)$.