NSI compléments Le codage des caractères et le codage de Huffman (8 points)

La méthode de codage de Huffman est une méthode de compression de données inventée par David Albert Huffman en 1952, qui permet de réduire la longueur du codage d'un alphabet et qui repose sur la création d'un arbre binaire.

On appelle alphabet l'ensemble des symboles (caractères) composant la donnée de départ à compresser.

Codage unicode

Pour un alphabet reduit à une seule langue, le nombre de bits est à peu près le même

La *longueur* du nombre binaire est alors *variable*. Un caractère peut nécessiter 8 (ascii), 16 bits, ou plus (jusqu'à 32). Une information dans le code numérique va préciser cette longueur (correspond à un caractère spécial comme le Ã). Cela va permettre d'afficher tous les caractères. Cela génère un fichier dont le poids sera inférieur à l'utf-32.

Pour le système d'exploitation Windows, l'encodage par defaut est l'utf-16. MacOs utilise un encodage utf-8.

0-> 127 : 1 octet

128 -> 2048 (2^11): 2 octets

2049 -> 1 048 335

Caractères codés	Représentation binaire UTF-8
U+0000 à U+007F	0bbb·bbbb
U+0080 à U+07FF	110b·bbbb 10bb·bbbb
U+0800 à U+FFFF	1110·bbbb 10bb·bbbb 10bb·bbbb
U+10000 à U+10FFFF	1111·0bbb 10bb·bbbb 10bb·bbbb 10bb·bbbb
	1111·0100 1000·bbbb 10bb·bbbb 10bb·bbbb

Code	Caractère	Code	Caractère	Code	Caractère	Code	Caractère	Code	Caractère
48	0	64	@	80	P	96	`	112	p
49	1	65	A	81	Q	97	a	113	q
50	2	66	B	82	R	98	b	114	r
51	3	67	C	83	S	99	c	115	s
52	4	68	D	84	T	100	d	116	t
53	5	69	E	85	U	101	e	117	u
54	6	70	F	86	V	102	f	118	v
55	7	71	G	87	W	103	g	119	w
56	8	72	H	88	X	104	h	120	x
57	9	73	I	89	Y	105	i	121	y
58	:	74	J	90	Z	106	j	122	z
59	;	75	K	91	[	107	k	123	{
60	<	76	L	92	\	108	l	124	|
61	=	77	M	93	]	109	m	125	}
62	>	78	N	94	^	110	n	126	~
63	?	79	O	95	_	111	o	127	DEL

ASCII