Données numériques 1 - Exercices **Coder les caractères**

Ex 1 : l'UTF-32 est un encodage à taille fixe des caractères. Combien d'octets y-a-t-il pour chaque caractère?

Ex 2 : ASCII tableau 0-127

LSB \ MSB		0 000	1 001	2 010	3 011	4 100	5 101	6 110	7 111
0	0000	NUL	DLE	SP	0	@	P	`	p
1	0001	SOH	DC1	!	1	A	Q	a	q
2	0010	STX	DC2	"	2	B	R	b	r
3	0011	ETX	DC3	#	3	C	S	c	s
4	0100	EOT	DC4	$	4	D	T	d	t
5	0101	ENQ	NAK	%	5	E	U	e	u
6	0110	ACK	SYN	&	6	F	V	f	v
7	0111	BEL	ETB	'	7	G	W	g	w
8	1000	BS	CAN	(	8	H	X	h	x
9	1001	HT	EM	)	9	I	Y	i	y
A	1010	LF	SUB	*	:	J	Z	j	z
B	1011	VT	ESC	+	;	K	[	k	{
C	1100	FF	FS	,	<	L	\	l	|
D	1101	CR	GS	-	=	M	]	m	{
E	1110	SO	RS	.	>	N	^	n	~
F	1111	SI	US	/	?	O	_	o	DEL

1. Sachant que le caractère A est représenté en ASCII par la séquence binaire 100 0001, expliquer la méthode avec laquelle on utilise la table proposée ci-dessus.
2. Quels sont les points de code des lettres minuscules? Exprimer ces points de code en hexadecimal.

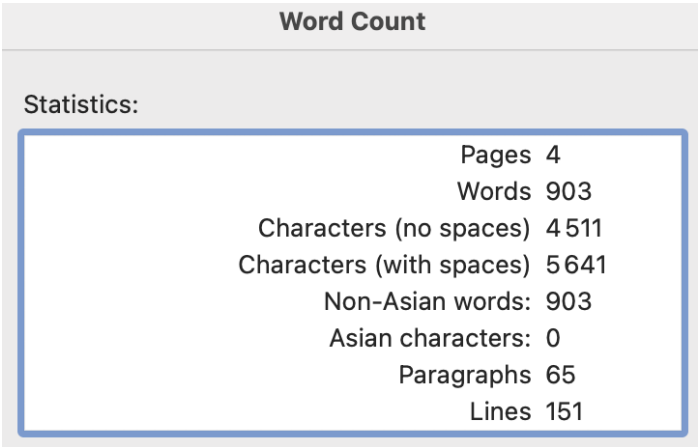
3. Donner la séquence binaire du mot : **lac**

4. En utf-8, pour rendre compatible les caractères ascii, on ajoute un 0 devant les 7 bits du caractère. Le A se code alors avec 0100 0001. Quels mots se cachent derrière les codes suivants?

```
01101000 01100101 01101100 01101100 01101111
```

Ex 3 : Word count

A partir des informations données par le logiciel word, évaluer la taille du fichier texte sur le disque dur (codage utf-8)



Ex 4 : ASCII étendu : tableau 128-255

La table suivante donne la suite des caractères en norme **ASCII-Latin1** (sur un octet), pour les caractères dont le point de code est supérieur à 127. Les encodages sur 1 octet ne sont pas compatibles avec ceux UNICODE utilisés par defaut par les navigateurs. Cela génère des erreurs d'affichage pour les caractères accentués par exemple.

ASCII	Caractère	ASCII	Caractère	ASCII	Caractère	ASCII	Caractère
128	€	160		192	À	224	à
129		161	¡	193	Á	225	á
130	‚	162	¢	194	Â	226	â
131	ƒ	163	£	195	Ã	227	ã
132	„	164	¤	196	Ä	228	ä
133	…	165	¥	197	Å	229	å
134	†	166	¦	198	Æ	230	æ
135	‡	167	§	199	Ç	231	ç
136	ˆ	168	¨	200	È	232	è
137	‰	169	©	201	É	233	é
138	Š	170	ª	202	Ê	234	ê
139	‹	171	«	203	Ë	235	ë
140	Œ	172	¬	204	Ì	236	ì
141		173		205	Í	237	í
142	Ž	174	®	206	Î	238	î
143		175	¯	207	Ï	239	ï
144		176	°	208	Ð	240	ð
145	‘	177	±	209	Ñ	241	ñ
146	’	178	²	210	Ò	242	ò
147	“	179	³	211	Ó	243	ó
148	”	180	´	212	Ô	244	ô
149	•	181	µ	213	Õ	245	õ
150	–	182	¶	214	Ö	246	ö
151	—	183	·	215	×	247	÷
152	˜	184	¸	216	Ø	248	ø
153	™	185	¹	217	Ù	249	ù
154	š	186	º	218	Ú	250	ú
155	›	187	»	219	Û	251	û
156	œ	188	¼	220	Ü	252	ü
157		189	½	221	Ý	253	ý
158	ž	190	¾	222	Þ	254	þ
159	Ÿ	191	¿	223	ß	255	ÿ

Avec l'encodage UNICODE utf-8, les 8 bits de ces caractères sont placés sur 2 octets. Ces bits codants (appelées Point de code) sont ajoutés à des bits non codants (le masque). Dans cet exercice, les bits codants sont mis en gras.

Une lettre en utf-8 a pour code binaire `110`**`00010`** `10`**`011011`**. Cette écriture mélange des bits qui servent de masque, et des bits qui forment le point de code (en gras).

1. Convertir chacun des 2 octets, `11000010 et 10011011` en caractères ascii. Ce seront les caractères affichés si la page HTML contient la balise **`<meta charset="ascii">`**?

2. Reconstruire l'octet correspondant au Point de code. (les 11 bits en **gras**). Donner alors sa représentation avec le réglage **`<meta charset="utf-8">`**

3. Même question avec le caractère à (a accent aigü), dont le code binaire est: `110`**`00011`** `10`**`100000`**