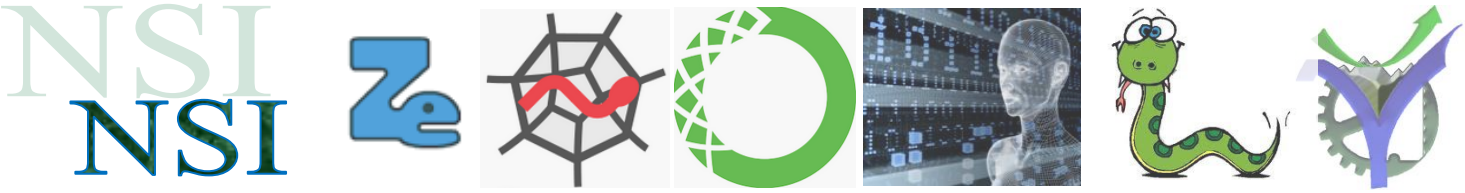


Découvrir et approfondir Unicode

Nom :

Unicode est un système de **codage** informatique qui a pour but d'unifier les échanges de texte au niveau international. Avec **Unicode**, chaque caractère informatique est décrit par un nom et un **code** l'identifiant de manière unique quelque soit le support informatique ou le logiciel utilisé.

Codage Unicode \uXXXX - Déchiffrer, Décoder, Encoder
https://www.dcode.fr › codage-unicode

Classe : **Note :**

1 Le codage ASCII

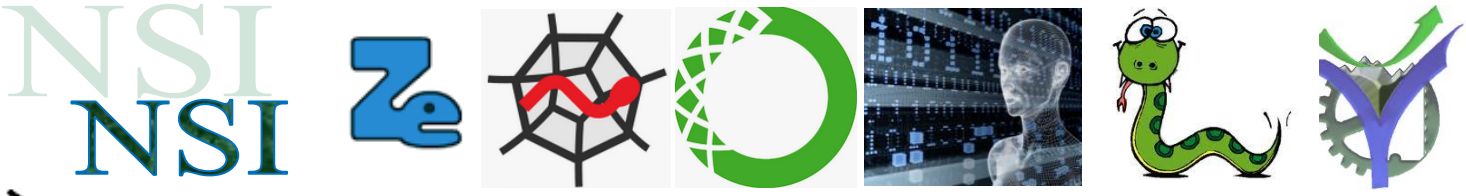
American Standard Code for Information Interchange

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

Le jeu de caractères ASCII a été créé dans les années 1960 pour, déjà, uniformiser les différents codages existants. Ce code est défini sur 7 bits.

Exercice 1. Donner le nombre de caractères disponibles :

La table de codage ASCII est rappelée ci-dessus, on peut y observer différentes catégories de caractères. Les lettres de l'alphabet latin en majuscule et minuscule, les chiffres de 0 à 9, des signes de ponctuations, des opérateurs mathématiques + * / -, des caractères spéciaux.



Exercice 2. **Remplir en le complétant le tableau ci-dessous :**

Caractère ASCII	Catégorie	Code ASCII en hexadécimal	Code ASCII en binaire (sur 7 bits)
a …… z	Lettres Latin minuscule		
A …… Z	Lettres Latin majuscule		
0 …… 9	Chiffres		
CR	Caractère spécial		
+	Opérateur d'addition		
Le caractère espace	Caractère spécial		

Exercice 3. **Combien de bits diffèrent entre le codage d'une lettre en minuscule et le codage de la même lettre en majuscule ?**

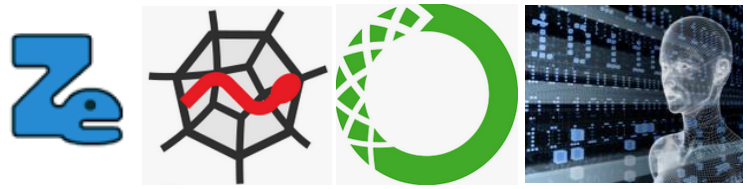
Exercice 4. **En utilisant une ressource Web donner la signification et le code des caractères spéciaux suivants :**

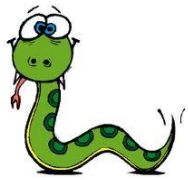
CR :

LF :

DEL :

Ces codes étaient utilisés pour piloter des téléscripteurs et donc on y trouve des commandes d'avancement du chariot (la tête d'impression), d'avancement d'une ligne du papier, d'avancement à la page suivante. Certains codes étaient utilisés dans des protocoles de transmission ou des contrôles de périphériques.





2 Le codage Unicode

2.1 Identification des caractères avec un numéro et un nom unique

Avec 7 bits il est bien évident qu'il est impossible de coder 'tous les caractères du monde'. Il a bien fallu normaliser de nouveau les tables de codage. Une tentative a eu lieu avec la norme ISO 8859 https://fr.wikipedia.org/wiki/ISO/CEI_8859

Le standard de codage Unicode s'est imposé, il propose une représentation unique de chaque caractère **par son nom** et **un code appelé point code**. Le code est noté U+xxxx où chaque x représente une valeur hexadécimale.

Exemple : le caractère ñ a pour code Unicode U+00F1 et pour nom LETTRE MINUSCULE LATINE N TILDE (trouvé dans Word Supplément Latin-1) :

Ce codage permet de mélanger tous caractères dans un même document. Il est compatible avec le code ASCII pour les 128 premiers caractères.

Le mot `paix` de par le monde

arabe : السلم	français : paix	grec : ειρήνη
hébreu : שלום	japonais : へいわ	russe : мир
symbole : ☮	tchèque : mír	thai : ความสงบสุข

[1]

Quelques codes Unicode :

	Codage Unicode	
Caractère	**Nom du caractère**	**Point Code**
☮	**Symbole de paix**	**U+262E**
♫	**Beamed eighth notes**	**U+262B**
€	**Symbole Euro**	**U+20AC**
A	**Latin majuscule A**	**U+0041**
è	**Latin e minuscule accent grave**	**U+00E8**

[1] Source http://www.fil.univ-lille1.fr/~wegrzyno/portail/Info/Doc/HTML/seq7_codage_caracteres.html consulté le 11 septembre 2019.

2.2 Encodage du caractère en binaire

Une fois connu le point code du caractère noté U+xxxx pour déterminer son code binaire effectif il faut tout d'abord choisir une méthode de codage UTF-8, UTF-16, UTF-32. La méthode la plus couramment utilisée aujourd'hui est l'UTF-8.

L'UTF-8 encode les caractères avec un mot de 1 à 4 octets organisés comme suit :

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Le premier octet indique le nombre d'octet total du mot. Si ce premier octet commence par un 0 le code est écrit sur un seul octet, compatible asccii. Si le premier octet commence par un 1 alors il faut compter le nombre de 1 se trouvant à la suite avant le premier 0 pour connaitre le nombre total d'octets du code. Ensuite tous les octets commencent par 10 de la sorte il n'y a aucune confusion possible pour le décodage.

Exemple pour un codage sur 3 octets :

1110xxxx	10xxxxxx	10xxxxxx

Le premier octet, encadré en jaune, commence par : 1 1 1 0 x x x x il y a trois valeurs 1 avant le premier 0 donc nous avons trois octets au total. Les x indiquent le nombre de bits disponibles pour écrire en binaire la valeur du point code ici 4 + 6 + 6 = 16.

Exemple codage du caractère è en UTF-8 :

Le point code vaut : U+00E8 soit 1110 1000 en binaire.

Ce code a besoin de 8 bits pour être écrit.

En UTF-8 il faut deux octets. Le codage est de la forme :

110xxxxx	10xxxxxx

Il ne reste plus qu'à remplir les x en commençant par la droite, les poids faibles, et en complétant par des 0 le résultat est donné ci-dessous :

1	1	0	0	0	0	1	1	1	0	1	0	1	0	0	0

En hexadécimal cela nous donne : 0x**C3A8**

Exercice 5. En utilisant les données et la méthode explicitée ci-dessus déterminer les codages UTF-8 des caractères ci-dessous en détaillant les calculs :

☮ U+262E

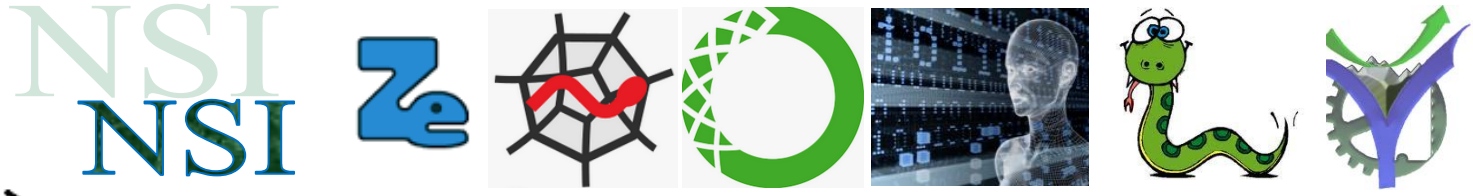
€ U+20AC

A U+0041

Exercice 6. Expliquer brièvement les raisons de la réception d'un email comme l'exemple ci-dessous :

rÃ©active, crÃ©ative et proche de vous, spÃ©cialiste pour :

- La rÃ©alisation de vos campagnes de Marketing Direct (mailing, emailing, multi-canal),
- La production de vos documents commerciaux (plaquette, fiche produit) et de votre signalÃ©tique (affiche, annonce-presse),
- Lâ€™animation de votre site web (enewsletter, bandeaux et banniÃ¨res)
- Le conseil en plan fichier et les bases de donnÃ©es marketing.



Exercice 7. Décodage de caractères Unicode

Vous recevez le flux de bits suivants composés de caractères Unicode : CF 81 CE A9 40 C3 A8

Pour décoder le message suivre les étapes ci-dessous :

a) Écrire le message en binaire

b) Repérer le découpage du codage UTF-8 en 1,2 ou 3 octets, (il n'y a pas de caractère encodé sur 4 octets dans cet exemple), pour cela après avoir recopié vos valeurs précédentes dans le tableau ci-dessous entourez les différents groupes avec de la couleur :

0							
1	1	0					
1	1	1	0				

Les trois premiers formats de codage de l'UTF-8

c) Dans chacun des groupes identifiés **isoler l'information point code** du caractère reçu, pour vous aider vous devez trouver la séquence sous la forme ci-dessous :

1	1	0					
1	1	0					
0							
1	1	0					

La suite des caractères reçus est donc :