



PREMIERE GENERALE

NUMERIQUE ET SCIENCES DU NUMERIQUE

CODAGE BINAIRE DU TEXTE

- ▷ Histoire de l'informatique
- ▷ Représentation des données : types et valeurs de base
- ▷ Représentation des données : types construits
- ▷ Traitement de données en tables
- ▷ Interactions entre l'homme et la machine sur le Web
- ▷ Architectures matérielles et systèmes d'exploitation
- ▷ Langages et programmation
- ▷ Algorithmique

Afin de représenter les caractères alphanumériques avec un ordinateur, le code **ASCII** (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange) est développé en **1963**. A ses débuts, le code ASCII permettait de représenter les caractères principaux de la langue anglaise et des caractères non imprimables comme le retour à ligne.

En 1986, apparition d'une extension de la norme américaine ASCII, la norme **ISO 8859-1** encore appelée **Latin-1**. Formée d'un ensemble de 191 caractères de l'alphabet latin, tous codés sur un unique octet, elle permet de représenter la très grande majorité des caractères de nombreuses langues d'Europe occidentale. Des extensions existent pour d'autres alphabets, comme le cyrillique ou le polonais.

En 1996, la norme **UTF-8** est adoptée. Cette norme permet un encodage de taille variable selon le caractère à coder, ce qui permet de limiter le coût en mémoire et d'assurer la compatibilité avec la norme ASCII, plus ancienne. Les caractères de la table ASCII restent ainsi codés sur 7 bits, tandis que les autres caractères sont représentés sur 2, 3 ou 4 octets. L'encodage UTF-8 est utilisé par plus de 95% des sites web en octobre 2020.

A	Ω	語	III
41	CE A9	E8 AA 9E	F0 90 8E 84

1 – PROBLEMATIQUE

En recherchant un extrait du « Seigneur des anneaux », nous sommes tombés sur la page WEB ci-contre.

Â« Trois Anneaux pour les Rois Elfes sous le ciel,
Sept pour les Seigneurs Nains dans leurs demeures de pierre,
Neuf pour les Hommes Mortels destinÃ©s au trÃ©pas,
Un pour le Seigneur des TÃ©nÃ©bres sur son sombre trÃ©ne
Dans le Pays de Mordor oÃ¹ sÃ©tÃ©ntendent les Ombres.
Un Anneau pour les gouverner tous, Un Anneau pour les
trouver,
Un Anneau pour les amener tous et dans les tÃ©nÃ©bres les
lier
Au Pays de Mordor oÃ¹ sÃ©tÃ©ntendent les Ombres. Â»

Â« La Compagnie de lâ€™Anneau sera de Neuf ; et les Neufs
Marcheurs seront opposÃ©s aux Neufs Cavaliers qui sont
mauvais. Gandalf ira avec vous et votre fidÃ©le serviteur ; car
ceci sera sa grande tÃ¢che et peut-Ãªtre la fin de ses labeurs.
Pour le reste, ils reprÃ©senteront les autres Gens Libres du
Monde : Elfes, Nains et Hommes. Legolas reprÃ©sentera les
Elfes, et Gimli, fils de GlÃ©in les Nains. [...] Pour les
Hommes, vous aurez Aragorn fils dÃ©tÃ©Arathorn ainsi que
Boromir du Gondor Â»



Exercice n°1

1. Indiquer pourquoi une partie du texte n'est pas compréhensible.
2. Préciser quel type de lettres pose problème.

2 – LE CODE ASCII

Le code **ASCII** est basé sur un tableau contenant les caractères les plus utilisés en **langue anglaise**. Chaque caractère est représenté sur **7 bits** (ce qui donne $2^7 = 128$ combinaisons possibles).

	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BELL	BS	HT	LF	VT	FF	CR	SO	SI
1x	DEL	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2x	SP	!	«	#	\$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	,	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	'	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

Exemples

Le caractère « A » est codé en ASCII par **0x41** en hexadécimal (**65** en décimal ou **0b1000001** en binaire).
 Le caractère « 1 » est codé en ASCII par **0x31** en hexadécimal (**49** en décimal ou **0b0110001** en binaire).

Les codes compris entre **0x00** et **0x1F** ne représentent pas des caractères, ils ne sont pas affichables. Ces codes, souvent nommés **caractères de contrôles** sont utilisés pour indiquer des actions comme passer à la ligne (**CR**, **LF**), émettre un bip sonore (**BEL**), etc...

Exercice n°2

1. Coder, à l'aide de la table ci-dessus, la phrase suivante: "La compagnie de l'anneau".
2. Retrouver le texte correspondant au code ASCII suivant : 0x46 72 6F 6E 64 6F 6E 20 6C 65 20 48 6F 62 62 69 74.
3. Justifier, pourquoi il n'est pas possible de coder correctement le texte "Mon précieux" à l'aide du code ASCII.

3 – CODAGE UNICODE

La généralisation d'Internet dans le monde a nécessité une prise en compte d'un nombre beaucoup plus important de caractères. La **norme Unicode** permet le codage de texte écrit en donnant à tout caractère de n'importe quel système d'écriture un **charset, constitué d'un nom d'un identifiant numérique**, et ce de manière unifiée, quelle que soit la plate-forme informatique ou le logiciel. Le répertoire Unicode peut contenir **plus d'un million de caractères**. La norme Unicode définit des **méthodes standardisées pour coder et stocker cet index sous forme de séquence d'octets** : **UTF-8**, **UTF-16**, **UTF-32** et leurs différentes variantes. L'UTF-8 est l'encodage (Encoding) le plus répandu. Les navigateurs Internet utilisent le codage UTF-8 par défaut. L'index des **caractères UNICODE** est disponible à l'adresse suivante : <http://www.unicode.org/fr/charts/charindex.html>.

L'encodage **UTF-8** utilise **1, 2, 3** ou **4 octets** en respectant certaines règles :

- Un caractère déjà codé en ASCII de base est **codé de manière identique en UTF-8**.

Caractère	Valeur Unicode		Code UTF-8	
	Décimale	Binaire	Binaire	Hexadécimal
A	65	0b01000001	0b01000001	0x41

- Les bits de poids fort du premier octet **forment une suite de 1 indiquant le nombre d'octets** utilisés pour coder le caractère **suivi d'un 0**. Les octets suivants commencent tous par le **bloc binaire 10**.

Caractère	Valeur Unicode		Code UTF-8	
	Décimale	Binaire	Binaire	Hexadécimal
é	233	0b 00011 101001	0b 11000011 10101001	0x C3 A9
€	8364	0b 0010 000010 101100	0b 11100010 10000010 10101100	0x E2 82 AC

Exercice n°3

1. **Déterminer** le code UTF-8 du caractère « é » sachant qu'il est codé sur 3 octets

2. **Justifier** pourquoi le caractère « é » (e accent aigu) a été remplacé par les caractères « Ã© » dans l'exemple ci-contre dans lequel l'encodage **ISO 8859-1** a été sélectionné alors que la page est encodée en UTF-8.

Un pour le Seigneur des TÃ©nÃ©bres sur son sombre trÃ©ne
Dans le Pays de Mordor oÃ¹ s'Ã©tendent les Ombres.

4 – REPRESENTATION DU TEXTE SOUS PYTHON

Pour représenter des **chaînes de caractères** (**str** : **string** en anglais) sous python on utilise des guillemets simples ou doubles.

'NSI' ou "NSI"

Les chaînes de caractères doivent tenir sur une seule ligne. Un texte sur **plusieurs lignes** soit être délimité par trois doubles guillemets (ce type de représentation est utilisé pour la documentation des fonctions dans un script).

""" Bienvenue en
NSI """

Les versions actuelles du langage Python, comme la plupart des langages modernes, ont intégré la norme de codage **Unicode** comme représentation à la fois du code source du programme et des chaînes de caractères du langage. Il est ainsi possible d'utiliser des caractères accentués, des alphabets non-latins, et même des émojis non seulement dans les commentaires, mais aussi (à l'exception des émojis), pour les noms de variables et de fonctions. Il vaut mieux cependant s'en tenir aux **caractères ASCII pour les noms de variables et de fonctions**.

Exercice n°4

Tester et expliquer les instructions suivantes :

```
>>> 'a'
>>> "a"
>>> 'é'
>>> ascii(456)
>>> str(456)
>>> int('456')
>>> ord('a')
>>> ord('é')
>>> chr(97)
>>> chr(30)
```