Encodage du texte

•ASCII

i En 1960, le développement de l'informatique est essentiellement anglophone. A cette époque, le codage `ASCII`, pour American Standard Code for Information Interchange est crée pour écrire des textes en anglais.
Cette norme ne définissait que $128 = 2^7$ codes.
Ils sont en fait représentés par mots de huit bits (un octet), le premier étant toujours un zéro.

La table `ASCII` fournit la correspondance entre $128 = 2^7$ caractères et leur représentation binaire. 95 caractères sont imprimables :

- les chiffres de 0 à 9,
- les lettres minuscules de a à z et les majuscules de A à Z,
- des symboles mathématiques et de ponctuation

Exercice 1

Trouver la représentation binaire en ASCII du texte

`SNT !`

.....

.....

.....

Exercice 2 ★

Décoder le texte représenté en ASCII binaire par la suite de bits :

01000011001001110110010101110011011101000000011001100110000010110001101101001011011000110010

.....

.....

.....

Exercice 3

Peut-on coder le message suivant ? Justifier.

`«Bonjour à tous, c'est très facile de coder en ASCII.»`

.....

.....

•Norme ISO-8859

La nécessité de représenter des textes comportant des caractères non présents dans la table ASCII tels ceux de l'alphabet latin utilisés en français comme le 'à', le 'é' ou le 'ç' impose l'utilisation d'un autre codage que l'ASCII.
Afin de faciliter les choses, ces propositions sont des extensions du codage ASCII :

1. le codage des caractères présents dans la table ASCII est conservé ;
2. le principe du codage de chacun des caractères sur un octet est conservé.

Mais les 8 bits de l'octet vont être utilisés. Cela permet de coder $2^8 = 256$ caractères, soit 128 caractères supplémentaires.
L'ISO, organisation internationale de normalisation, propose de son côté plusieurs variantes de codages adaptées aux différentes langues. La plus utilisée concerne les langues européennes occidentales. Il s'agit de l'ISO-8859-1
Mais il existe d'autres variantes adaptées à d'autres langues :

— ISO-8859-2 pour les pays d'Europe de l'est,

— ISO-8859-3 pour les pays du sud est de l'Europe...

•UTF-8

UTF = UCS Transformation Format *(UCS = Universal Character Set, norme ISO-10646)*
Le numéro de chaque caractère est donné par le standard Unicode.
Les caractères de numéro 0 à 127 sont codés sur un octet dont le bit de poids fort est toujours nul.
Les caractères de numéro supérieur à 127 sont codés sur plusieurs octets. Dans ce cas, les bits de poids fort du premier octet forment une suite de 1 de longueur égale au nombre d'octets utilisés pour coder le caractère, les octets suivants ayant 10 comme bits de poids fort.
Définition dun nombre d'octets utilisés

Représentation binaire UTF-8	Signification
0xxxxxxx	1 octet codant 1 à 7 bits
110xxxxx 10xxxxxx	2 octets codant 8 à 11 bits
1110xxxx 10xxxxxx 10xxxxxx	3 octets codant 12 à 16 bits
11110xxx 10xxxxxx 10xxxxxx 10xxxxxx	4 octets codant 17 à 21 bits

Caractère	Numéro du caractère	Codage binaire UTF-8
A	65	01000001
é	233	11000011 10101001
€	8364	11100010 10000010 10101100
𝄞	119070	11110000 10011101 10000100 10011110

Voir la table des codes https ://unicode-table.com/fr/cjk-unified-ideographs-extension-a

? Quel encodage choisir ?

Sans hésitation il faut choisir systématiquement l'encodage UTF-8.
UTF-8 est en effet un codage de caractères conçu pour coder l'ensemble des caractères Unicode, tout en restant compatible avec la norme ASCII.
C'est devenu l'encodage par défaut sur Python3 et c'est aussi l'encodage le plus utilisé sur le Web.