Codage des caractères : à savoir

Il existe différentes tables de codage des caractères ces tables associent à chaque caractère d'un ensemble de caractères un…….

Ces ……. sont ensuite transformés en binaires pour encoder les caractères (le binaire obtenu n'est pas forcément la simple écriture du point de code en binaire : voir Utf-8)

Tout d'abord la table…….. qui compte ……. caractères. On code ceci sur …….bits. En fait on utilise ……. octet avec 1 bit servant à …….

Dans cette table il manque …….. de la langue française et bien d'autres caractères des langues du monde entier.

On a donc créé des tables avec des compléments en occupant le dernier bit non utilisé.

Les tables ASCII ……..qui comptent …… caractères que l'on code sur …….et qui permettent de coder $2^{\cdots}$ = …… caractères.

Ces tables sont entièrement …… avec la table ASCII.

Dans le but d'……. tous les codages un consortium a créé …… Cette norme attribue à chaque caractère un identifiant numérique appelé …… Elle couvre tous les caractères existants et permet le codage de millions de caractères.

Cette table est entièrement compatible avec la table …… mais seulement partiellement avec les tables ……..

Il y a différentes normes pour encoder les caractères du jeu de caractères UNICODE.

La plus courante est ……Cet encodage n'utilise pas le même nombre …….. pour chaque caractère. En effet si 2^{21} caractères étaient codés sur le même nombre d'octets il faudrait ……. octets pour chaque caractère et les textes seraient …..plus lourds qu'en ASCII étendu.

En Utf-8, on utilise … octet pour les caractères les plus fréquents (ceux de la table …….).

Pour les autres caractères on utilise ……..octets.

C'est pour cela que par exemple "mathÃ©matique" apparaît avec deux caractères à la place du "é" quand le codage est mal détecté. Le "é" est codé sur deux octets : les binaires qui correspondent à chacun des deux octets sont les encodages de "Ã" et à "©" dans des tables ASCII étendu.

Pour reconnaître combien d'octets doivent être lus pour un caractère on regarde les bits de …….

Remarque : conversions en Python :

pour convertir un binaire (ex : 101010) en décimal : int('101010',2)
pour convertir un hexa (ex : C3) en décimal : int('C3',16)
pour convertir un décimal (ex : 195) en binaire : hex(195) et bin (195) en binaire
pour avoir le point de code Unicode d'un caractère : ord("A") par ex (donne 65)
pour avoir le caractère correspondant à un point de code : chr(65) qui donne "A"