CH5 LE CODAGE DU TEXTE

I. LE CODE ASCII

Aujourd'hui, les ordinateurs ne sont plus, loin de là, de simples calculateurs. Il est donc important de pouvoir également coder facilement du texte.

Pour cela, on utilise notamment le code ASCII (American Standard Code for Information Interchange) ; une norme informatique du codage de caractère apparue dans les années 1960.La norme ASCII définit 128 codes (numérotés de 0 à 127 et codés sur 7 bits).Ceci comprend 95 caractères imprimables (chiffres de 0 à 9, lettres minuscules et majuscules de A à Z, des symboles mathématiques et de la ponctuation) : table ASCII

Même si 7 bits suffisent pour représenter 128 caractères, en pratique chaque caractère occupe un octet en mémoire. Le bit de poids fort est utilisé pour une somme de contrôle afin de détecter d'éventuelles erreurs de transmission : sa valeur est fixée de façon à ce que le nombre de bits à 1 dans l'octet soit toujours pair. On parle de bit de parité.

Exemple

On veut trouver l'écriture binaire en norme ASCII du mot « Info ».

On lit dans la table que I correspond à 73, soit 01001001 en binaire.

De même n correspond à 01101110, f correspond à 01100110, et o correspond à 01101111.

Info s'écrit donc sur 4 octets : 01001001 01101110 01100110 01101111

Applications

1) *A l'aide de la table ASCII, coder en binaire la phrase suivante : « Je m'amuse en NSI »*
2) *Voici maintenant une exclamation codée en binaire :*

01000010 01110010 01100001 01110110 01101111 00100001

Retrouver cette exclamation.

3) *Peut-on coder en binaire la phrase « Un âne est passé par là. » à l'aide de la table ASCII ? Justifier.*

Nous avons tous reçu un jour un courrier étrange ou lu une page web telle que celle-ci :

> Prenons lâ€™exemple typique de la lumiÃ¨re Ã©mise par un phare maritime : elle est dâ€™abord indivisible, son coÃ»t de production Ã©tant alors indÃ©pendant du nombre d'utilisateurs ; elle possÃ¨de une propriÃ©tÃ© de non-rivalitÃ© (elle ne se dÃ©truit pas dans l'usage et peut donc Ãªtre adoptÃ©e par un nombre illimitÃ© d'utilisateurs) ; elle est Ã©galement non excluable car il est impossible dâ€™exclure de lâ€™usage un utilisateur, mÃªme si ce dernier ne contribue pas Ã son financement.

Comment cela se fait-il ? Pourquoi peut-on comprendre mais pas complètement ?

ASCII et Python

La fonction ord de Python renvoie le code ASCII correspondant à un caractère. L'entier est renvoyé en base 10, que l'on peut convertir en hexadécimal avec la fonction hex.

```
>>> ord('a')
97
>>>
>>> hex(ord('a'))
'0x61'
```

Inversement, la fonction chr renvoie le caractère correspondant à un entier.

```
>>> chr(0x26)
'&'
```

II. Les normes ISO-8859 ET UNICODE

Il a donc fallu étendre la table ASCII pour pouvoir coder de nouveaux caractères. Les mémoires devenant plus fiables, on a utilisé le 8ème bit pour coder plus de caractères (il était utilisé auparavant pour contrôle). Elle permet l'affichage des caractères accentués de plusieurs langues et de symboles courants. La norme ISO 8859 permet ainsi de coder tous les caractères des langues européennes. Elle utilise 8 bits et permet donc de coder 256 caractères. Elle compte 16 tables.

Cette norme ne permet pas de coder d'autres langues que les langues européennes, par exemple le chinois ou l'arabe.

Il est donc devenu essentiel de proposer une version unifiée internationale des différents encodages de caractères. Ceci a été permis en complétant l'ASCII et en permettant l'encodage de caractères autres que ceux de l'alphabet latin. Unicode définit à cet effet des dizaines de milliers de codes, Les 128 premiers codes restent compatibles avec ASCII.

L'UTF-8 est l'encodage le plus utilisé. Les navigateurs Internet utilisent ce codage et les concepteurs de sites utilisent à présent presque tous cette même norme. C'est pourquoi il y a de moins en moins de problèmes de compatibilité.

Exemple

Coder le mot « défi », en UTF-8.

On pourra utiliser le site https://www.utf8-chartable.de/

La norme HTML/CSS prévoit la possibilité d'indiquer le codage utilisé pour les caractères d'un fichier donné. Le codage utilisé est précisé par une valeur du tag <meta>, comme dans l'exemple :

```
<!DOCTYPE html>
<html>
<head>
<meta charset="UTF-8" />
</head>

<body>
<p>J'écris € en UTF-8.</p>
</body>
</html>
```

Expérimentez par exemple avec :

http://www.w3schools.com/tags/tryit.asp?filename=tryhtml_charsets

Dans le code HTML précédent, changez le "UTF-8" en "ISO-8859-1" et observez ce qu'il se passe lors de l'affichage du fichier dans votre navigateur.

Lors de la définition d'une page HTML, il est important de préciser l'encodage des caractères utilisés. À défaut, le navigateur web utilisera un encodage par défaut qui ne correspondra pas nécessairement à l'encodage utilisé pour écrire le fichier.

Exercices d'application

Exercice 1 :

Donner le code ASCII des deux chaînes de caractères Python ci-dessous :

a. 'bonjour tout le monde'

b. ' « programmer en Python »'

Exercice 2 :

L'objectif de cet exercice est d'étudier la conversion de format d'un fichier texte et l'impact sur la taille et sur le contenu du fichier.

1. Créer un fichier Word (ou openoffice) dans lequel vous tapez les mots « le petit ». Enregistrer le fichier au format .docx et noter la taille du fichier obtenu.

2. Enregistrer ce même fichier au format texte brut (.txt) en sélectionnant autre codage, UTF-8. Comparer la taille des deux fichiers. Comment expliquer cette différence ?

3. Copier-coller un texte contenant des caractères spéciaux dans un fichier Word. L'enregistrer en .docx et en .txt et comparer les tailles de fichiers.

4. Enregistrer le fichier précédent en choisissant le codage ASCII. Fermer et ouvrir ce fichier. Que s'est-il produit ?