Codage du texte

Les octets sont constitués de 0 et de 1. Ils ont été enregistrés en suivant un encodage. Pour les lire il faut utiliser le même encodage. Si l'on utilise un encodage différent, on obtient n'importe quoi.

I Le code ASCII (American Standard Code for Information Interchange)

Ce code est né dans les années 1960. Il était nécessaire d'échanger du texte entre ordinateur. Pour cela, il a fallu faire le choix d'une convention de codage standardisée, le code ASCII.

C'est une table de 128 caractères numérotés de 0 à 127 (sur 7 bits)

Code	48	49	50	51	52	53	54	55	56	57	58	59
Car	0	1	2	3	4	5	6	7	8	9	:	;

Vous pouvez avoir accès à ces différents caractères en maintenant appuyé la touche Alt et en tapant le code correspondant.

Quel caractère obtenez-vous en tapant Alt 65 ?

La table complète contient les lettres majuscules et minuscules, les chiffres, plusieurs caractères de ponctuation, et des caractères invisibles comme l'espace, la tabulation ou le retour à la ligne.

La table ASCII ne contient aucun caractère accentué.

II ASCII étendues

Cette table contient 256 caractères. Il existe une multitude de tables (latin 1, latin 9, windows-1252, ...)

Toutes ces tables contiennent les 128 premiers caractères de l'ASCII, ainsi que 128 caractères spécifiques selon la table utilisée : caractères latin, grecs, cyrilliques...

III Problème d'échange de fichier

Lorsqu'un fichier texte est enregistré en utilisant une table de caractères (un charset) particulier, il faut le lire en utilisant la même table, sans quoi certains caractères ne seront pas corrects.

Vous avez eu l'occasion d'observer ceci sur la séquence 1.

IV Unicode

Au début des années 1990, une norme est apparue : **Unicode**.

L'objectif était de régler définitivement le problème d'encodage. Cette table contient actuellement plus de 135 000 symboles.

Unicode ne peut pas être utilisé directement.

V Le codage UTF-8

Les 128 caractères du code ASCII sont codés sur 1 seul octet, et une séquence particulière permet d'accéder aux autres caractères de la table, qui sont codés sur 2 à 4 octets.

Python représente ses chaînes de caractères en Unicode.