Les chaînes de caractères - généralités

Capacités attendues

- ✓ Identifier l'intérêt des différents systèmes d'encodage.
- ✓ Convertir un fichier texte dans différents formats d'encodage.

1 Représentation des caractères

Les caractères sont les représentations informatiques des lettres (minuscules, majuscules, accentuées ou non, etc.), des chiffres et de tous les symboles utilisés pour communiquer par écrit (espace, ponctuation, etc.).

Depuis les années 1960, la norme ASCII (American Standard Code for Information Interchange) définit 95 caractères imprimables codés sur 7 bits, à savoir le caractère **« espace »** et :

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

La norme ISO 8859-1 (aussi appelée Latin-1), conçue en 1986, a étendu la norme ASCII à 191 caractères, en utilisant un octet. Cette norme était celle utilisée dans les pays occidentaux par le système d'exploitation Windows, et tend aujourd'hui à disparaître au profit du standard Unicode, né en 1991, qui permet aujourd'hui de représenter plus de 130 000 caractères dans une centaine d'écritures.

Concrètement, la norme Unicode est une table associant à chaque caractère un nombre entier naturel, appelé point de code, souvent écrit en hexadécimal. Par exemple :

- le point de code correspondant à la lettre `'A'` est U+0041 soit, en décimal : $4 \times 16 + 1 = 65$;
- le point de code correspondant à la lettre `'a'` est U+0061 soit, en décimal : $6 \times 16 + 1 = 97$;
- le point de code correspondant au signe € est U+20AC soit 8364 en décimal.

La représentation en machine d'un caractère Unicode s'effectue en général selon le format UTF-8. Ce dernier présente en effet l'avantage d'être compatible avec la norme ASCII, au sens où les 128 premiers caractères sont représentés par les mêmes bits qu'en ASCII, selon le schéma $0\square\square\square\ \square\square\square\square$.

La lettre `'a'` est par exemple codé par l'octet $0011\ 0001$.

Pour les points de code compris entre 128 et 2047, le format UTF-8 utilise 2 octets, selon le schéma

$$110\square\ \square\square\square\square\ \ 10\square\square\ \square\square\square\square$$

Par exemple, le point de code du caractère `'é'` vaut 233, soit en binaire sur 11 bits : $000\ 1110\ 1001$.

En UTF-8, le caractère `'é'` est donc représenté par les octets $110\mathbf{0\ 0011}\ \ 10\mathbf{10\ 1001}$.

Les points de code supérieurs à 2047 sont codés en UTF-8 sur 3 ou 4 octets.

1.1 Problèmes d'affichage d'un texte en français

Pourquoi des caractère étranges tels que Ã© sont parfois affichés à la place des lettres accentués ?

Cela se produit lorsqu'un fichier texte est encodé selon la norme UTF-8, mais lu en Latin-1.
Prenons en effet l'exemple de la lettre `'é'` codée par les octets $1100\ 0011\ \ 1010\ 1001$ en UTF-8 : si l'éditeur de texte (ou le navigateur Internet, etc.) traite ces octets en Latin-1, il interprète chacun d'entre eux comme la représentation d'un caractère sur 8 bits, et produit l'affichage des deux caractères Ã et ©.

Il faut donc veiller à préciser le format d'encodage pour que l'affichage soit correct.

Par précaution, il est ainsi recommandé de ne pas utiliser de caractères accentués dans les noms de variables ou fonctions d'un programme informatique, ni dans les noms de fichiers et répertoires.

---

2 Les chaînes de caractères

Définitions

Une **chaîne de caractères** est un type de variable (`str` en Python, de l'anglais *string*) composé d'un nombre quelconque de caractères. Une chaîne de caractères se note en général entre doubles guillemets droits : `"abcd"`.

La **longueur** d'une chaîne de caractères est le nombre de caractères qu'elle contient. En Python, c'est la fonction `len` (de l'anglais *length*) qui renvoie la longueur de la chaîne de caractères passée en argument.

Une chaîne peut ne contenir aucun caractère : c'est la **chaîne vide** `""`, de longueur nulle.

Les opérateurs

- La **concaténation** "+" : la concaténation `s1+s2` de deux chaînes de caractères `s1` et `s2` est la chaîne obtenue en ajoutant à la fin de `s1` tous les caractères de `s2` (renvoie une **chaîne de caractères**).
  (La concaténation d'une chaîne `s` avec elle même peut s'obtenir avec la syntaxe `s*2`.) (au lieu de `s+s`).
- La **répétition** "`*`" : elle permet de répéter autant de fois qu'on le souhaite une chaîne (renvoie une **chaîne de caractères**).
- L'**inclusion** `in` : permet de déterminer si un chaîne est dans une autre chaîne (renvoie un **booléen**).

Exemple :

`"abc" + "def"` vaut `"abcdef"`

`"abc"*3` vaut `"abcabcabc"`

`"bc"` in `"abc"` vaut `True`

Caractères spéciaux

L'antislash `"\"` est utilisé pour définir ou pour échapper certains caractères spéciaux. On donne quelques exemples ci-dessous.

- `\n` : saut de ligne
- `\"` : double guillemet droit "
- `\\` : antislash \

Exercice 1

Quel est l'affichage produit par le script ci-dessous ?

```
s = "a b"*2 + "ba" + "b"*3
print(s, len(s))
```

> **Les fonctions Python `chr` et `ord`**
>
> En Python, deux fonctions permettent de convertir un point de code en caractère, et inversement :
>
> - `chr(ptc)` renvoie le caractère dont le point de code est `ptc` : l'appel 97 renvoie `"a"` ;
> - `ord(char)` renvoie le point de code du caractère `char` : l'appel `ord("A")` renvoie 65.

Exercice 2

Que renvoie la fonction ci-dessous lorsque l'argument passé en paramètre est une lettre minuscule ?

```
def convert(c):
    return chr(ord(c)-32)
```

Exercice 3

On associe à chaque lettre minuscule son rang selon le tableau suivant.

Lettre	a	b	c	d	...	z
Rang	0	1	2	3	...	25

1. Écrire une fonction `rang(l)` qui renvoie le rang d'une lettre `l`.
2. Écrire une fonction `lettre(r)` qui renvoie la lettre de rang `r`.

Exercice 4

Quel est l'affichage produit par le script ci-dessous ?

```
s = "abn\nn\\\""
print(len(s))
```

> **Accès à un caractère d'indice donné**
>
> Dans une chaîne de caractères, chacun des caractères est repéré par son indice : il s'agit de sa position, en partant de 0 pour le premier (le plus à gauche), comme illustré ci-dessous pour la chaîne `"abcd"`.
>
caractères	a	b	c	d
indices	0	1	2	3
>
> Étant donnée une chaîne de caractères `ch`, le caractère d'indice `idx` est fourni par la syntaxe `ch[idx]`

Exercice 5

Quels sont les affichages produits par le script ci-dessous ?

```
s = "abcdefghij"
n = len(s)
print(s[1], s[5])
s = 2*s
print(n)
print(s[12])
```

Exercice 6

Quelle instruction permet d'affecter à la variable `x` la valeur du dernier caractère d'une chaîne de caractères `ch` ?

Exercice 7

Sachant que les chiffres ont un point de code compris entre 48 et 57, écrire une fonction `est_numerique` qui prend en argument une chaîne de caractères de longueur 1, et renvoie `True` si le caractère est un chiffre, `False` sinon.

Exercice 8

Écrire une fonction `est_une_date_valide` qui prend en argument une chaîne de caractères et renvoie `True` si elle représente une date au format jj/mm/aaaa, et `False` sinon. (Penser à ce qui a été fait dans l'exercice précédent !)

Formatage d'une chaîne de caractères : l'opérateur %

Il est souvent utile d'afficher sur une même ligne la valeur de plusieurs variables. En Python, il existe au moins trois solutions.

- Passer plusieurs valeurs en argument de la fonction `print` :

```
jour = 25
mois = 10
annee = 2022
print("La prochaine éclipse aura lieu le", jour, "/", mois, "/", annee, ".")
# La prochaine éclipse aura lieu le 25 / 10 / 2022 .
```

- Concaténer plusieurs chaînes (nécessite la conversion des variables en chaînes de caractères) :

```
print("La prochaine éclipse aura lieu le " + str(jour)
      + "/" + str(mois) + "/" + str(annee) + ".")
```

- Utiliser le marqueur %s, qui signale la présence d'un élément à insérer, et l'opérateur % :

```
print("La prochaine éclipse aura lieu le %s/%s/%s."%(jour, mois, annee))
```

- Utiliser les expressions formatées (en n'oubliant pas le `f` avant d'ouvrir les guillemets) :

```
print(f"La prochaine éclipse aura lieu le {jour}/{mois}/{annee}.")
```

Exercice 9

Écrire une fonction `affiche_coordonnees(x, y)` qui affiche `"Ce point a pour coordonnées (x; y)."` avec les valeurs de `x` et `y` passées en argument.

Par exemple, l'appel `affiche_coordonnees(4, 5)` devra afficher `"Ce point a pour coordonnées (4; 5)."`