LS4 - Travaux pratiques 7

7 avril 2014

Objectifs : introduire les expressions régulières.

1 Avant de commencer

Python gère les jeux de caractères comme UTF-8, qui permettent d'inclure des lettres accentuées. Faites quelques réglages nécessaires de votre environnement de travail :
1. Ajoutez la ligne suivante dans votre fichier .bashrc :

```
export LANG="fr_FR.UTF-8"
```

2. Réglez votre console sur le codage UTF-8. Vérifiez que cela fonctionne en affichange un fichier codé en UTF-8 (par exemple, vers-queneau.txt) sur votre console.
3. Vérifiez que votre éditeur de texte gère bien UTF-8 et sauvegarde les fichiers dans ce code.

2 Partie 1 : Syntaxe de expressions régulières de Python

Pour les expressions régulières suivantes, dire si chacune des chaînes correspond au motif.

Correspondance d'un symbole, juxtaposition

Le point "." peut être mis en correspondance avec n'importe quel symbole. Une juxtaposition de motifs correspond à la concaténation entre les chaînes qui sont en correspondance avec les motifs juxtaposés.
Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
ABC
```

q1 :
- ☐ abc
- ☐ ABC
- ☐ AXC
- ☐ A
- ☐ C

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
A.C
```

q2 :
- ☐ abc
- ☐ ABC
- ☐ AXC
- ☐ A
- ☐ C

Séquences spéciales

Certaines séquences spéciales sont utilisées pour représenter certains motifs. Elles sont en correspondance avec un symbole de la chaîne.

```
\d un chiffre
\D pas un chiffre
\s symbole blanc (espace, tabulation, retour chariot)
\S symbole non-blanc
\w symbole alphanumérique
\W symbole non-alphanumérique
```

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
A\d\w
```

q3 :

- ☐ A2abc
- ☐ A2a
- ☐ AA
- ☐ AaA2
- ☐ 2

Classes

Les expressions entre crochets représentent un ensemble de symboles possibles. Entre les crochets, les symboles spéciaux ne sont pas interprétés. (Excepté le '^' en première position dont on verra la signification dans la question suivante, ainsi que les séquences spàciales \d, \n, etc.)
Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[ABC]
```

q4 :

- ☐ A
- ☐ B
- ☐ AC
- ☐ ABC

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[\d?\w]
```

q5 :

- ☐ 2 ?n
- ☐ n
- ☐ \w
- ☐ \
- ☐ ?
- ☐ d

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[a-k]
```

q6 :

- ☐ -

- ☐ a-k
- ☐ a
- ☐ z
- ☐ 1
- ☐ e
- ☐ k

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[a-k1-5]
```

q7 :

- ☐ a2
- ☐ 2a
- ☐ e
- ☐ 3

Négation dans une classe

Lorsqu'un crochet ^ apparaît comme premier symbole dans un groupe, on l'interprète comme la négation de l'ensemble. Lorsqu'il se trouve en dehors d'une définition de classe ^ indique le début d'une ligne.
Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[^A]
```

q8 :

- ☐ A
- ☐ B
- ☐ C
- ☐ D
- ☐ ^

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[^ABC]
```

q9 :

- ☐ A
- ☐ B
- ☐ C
- ☐ D
- ☐ ^

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
^ABC
```

q10 :

- ☐ A
- ☐ B
- ☐ C
- ☐ D
- ☐ ^
- ☐ ABC

Alternative

L'alternative entre deux expressions est représentée par la barre verticale `|`.
Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB|C
```

q11 :
- ☐ AB
- ☐ C
- ☐ AC
- ☐ ABC
- ☐ —

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB.|A.C|.BC
```

q12 :
- ☐ ABC
- ☐ ADC
- ☐ ABAZZBC
- ☐ AB.

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[A|B]
```

q13 :
- ☐ A
- ☐ B
- ☐ AB
- ☐ —

Répétition

Les motifs peuvent être répétés avec l'étoile `*` (0 ou plusieurs fois), le plus `+` (1 ou plusieurs fois), le point d'interrogation `?` (0 ou 1 fois).

On peut aussi préciser le nombre de répétitions avec `{n,m}` (entre n et m fois), `{n}` (n fois), `{n,}` (au moins n fois), `{,m}` (au plus m fois).

L'ordre de priorité des opérateurs est comme suit :

```
répétition > concaténation > alternative
```

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
A*
```

q14 :
- ☐
- ☐ A
- ☐ AAAA

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[ABC]*
```

q15 :

- ☐ ABC
- ☐ [ABC]]]
- ☐ ABCCC
- ☐ AABBCC
- ☐ ABCABC

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
ABC*
```

q16 :

- ☐ ABC
- ☐ ABCCC
- ☐ AABBCC
- ☐ ABCABC

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB?
```

q17 :

- ☐ ABC
- ☐ ABAB
- ☐ A
- ☐ AB
- ☐ ABB

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB{0,1}
```

q18 :

- ☐ ABC
- ☐ ABAB
- ☐ A
- ☐ AB
- ☐ ABB

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB{3}
```

q19 :

- ☐ AB
- ☐ ABAB
- ☐ ABABAB
- ☐ ABB
- ☐ ABBB

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB{1,3}
```

q20 :

- ☐ A
- ☐ AB
- ☐ ABAB
- ☐ ABABAB
- ☐ ABB

☐ ABBB

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB{,3}
```

q21 :

☐ A
☐ AB
☐ ABAB
☐ ABABAB
☐ ABB
☐ ABBB

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
[0-9A-F]{3}
```

q22 :

☐ 0C0C0C
☐ 000
☐ FFF
☐ FF0

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
AB*|CD
```

q23 :

☐ ABBBD
☐ ABB
☐ ABBBB—CD
☐ CD

Groupes : sous-expressions (...), motifs répétés

On peut grouper des sous-expressions pour les réutiliser plus tard. On peut soit faire référence à un groupe à l'intérieur même d'une expression régulière pour exprimer la répétition d'un motif en utilisant `\1`, `\2`, `\3` selon la position du groupe dans l'expression. On verra aussi que le module `re` de Python permet de récupérer des groupes lorsque l'on veut découper du texte suivant des motifs complexes.

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
A(.*)B(.*)C\1\2
```

q24 :

☐ ABC
☐ AXXBZZCZZXX
☐ AXXBZZCXXZZ
☐ ABZZZCZZZ

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
A(.)(.)(.)C\3\2\1
```

q25 :

☐ AabcCcba
☐ AabcCabc
☐ AACCCCCA

☐ AC

Parmi les chaînes ci-dessous, quelles sont celles qui sont en correspondance avec le motif suivant ?

```
<B>(.*)</B>\1
```

q26 :

☐ ¡B¿¡/B¿
☐ ¡B¿uuu¡/B¿uuu
☐ ¡B¿uuu¡B¿uu

3 Quelques remarques additionnelles

3.1 Symboles réservés, antislash

Les expressions régulières de Python utilisent les symboles suivants :

```
. ^ $ * + ? { } [ ] \ | ( )
```

On utilise l'antislash pour signifier que le symbole doit être interprété litéralement.

3.2 Marqueurs de position

Lorsque l'on fait des recherches de motifs dans des chaînes, on peut préciser l'endroit où le motif se trouve : en début de ligne (avec `^`) ou en fin de ligne (avec `$`).

3.3 Répétition gourmande ou non-gourmande

Par défaut, les opérateurs de repétition sont gourmands, c'est-à-dire qu'ils se mettent en correspondance avec la sous-chaîne la plus longue parmi toutes les possibilités. Si on souhaite modifier ce comportement, on ajoute `?` après l'opérateur de répétition.

4 Partie 2. Expressions régulières en Python

Le module `re` contient (entre autres) les fonctions suivantes :

`re.match (motif, chaine)` retourne la correspondance du motif avec une sous-chaîne commençant au début de la chaîne.

`re.search (motif, chaine)` retourne la correspondance du motif avec une sous-chaîne quelconque de la chaîne.

Si ces fonctions trouvent une correspondance, celle-ci est retournée sous la forme d'une instance de la classe `match`.

Dans le cas contraire, elle retourne `None`.

On peut obtenir des informations sur un objet `m` de la classe `match` à l'aide de la méthode `m.group ()` :

`m.group (0)` retourne la sous-chaîne en correspondance avec le motif.

`m.group (1)` retourne la sous-chaîne en correspondance avec le premier sous-motif `\1`.

Plus généralement, `m.group (e)` pour e s'évaluant en un entier i supérieur à 1 retourne la sous-chaîne en correspondance avec le sous-motif `\i`.

Remarque : Dans les chaînes de Python (str), certaines séquences de caractères `\n`, `\t`,... ont des significations spéciales. Pour éviter que Python n'interprète ces symboles, on préfixera les chaînes litérales par un "r" qui signifie que la chaîne est une *raw string*.

Question 1

Que valent a, b, c et d après l'exécution du programme suivant ?

```
s = "1907,37,15,Petronille"
m = re.match ("19(\d\d),(.*),.*,(.*)", s)
a = m.group (0)
b = m.group (1)
c = m.group (2)
d = m.group (3)
```

q126 :

a :
b :
c :
d :

Question 2

Que valent a, b et c après l'exécution du programme suivant ?

```
s = "1907,37,15,Petronille"
m = re.match("19(\d\d),(.*),.*",s)
a = m.group(0)
b = m.group(1)
c = m.group(2)
```

q127 :

a :
b :
c :

Question 3

Que valent a, b, c et d après l'exécution du programme suivant ?

```
s = "1907,37,15,Petronille"
m = re.match ("19(\d\d),(.*?),(.*)", s)
a = m.group (0)
b = m.group (1)
c = m.group (2)
d = m.group (3)
```

q128 :

a :
b :
c :
d :

Remarque : En mode MULTILINE, `$` correspond non seulement à la fin de la chaîne mais également à tout symbole `\n`. De même, `^` correspond non seulement au début de la chaîne mais également à tout début de ligne (symbole suivant `\n`).

Question 4

Écrire une fonction `double(s)` qui prend en entrée une chaîne `s` et retourne une chaîne `r` telle que `s = rr`. Par exemple :

```
double ("blabla") = "bla"
double ("") = ""
double ("ababc") = None
```

q129 :

Question 5

Écrire une fonction `contientdouble(s)` qui prend en entrée une chaîne `s` et retourne une chaîne `r` de longueur au moins 2 telle que `s` contient `rr`.

```
double ("blabla") = "bla"
double ("") = None
double ("cababc") = "ab"
```

q130 :

Question 6

Écrire une fonction `extraire_ligne(s)` qui prend en entrée une chaîne de la forme `annee,quantite,id,prenom` et retourne un 4-uplet de la forme `(annee,quantite,id,prenom)`. Vous devez utiliser uniquement la fonction `re.match` ainsi que la fonction `str.strip` pour supprimer les blancs (\t, \n, ...) non désirés.

q131 :

Question 7

Écrire une fonction `extraire_fichier(s)` qui prend en entrée un nom de fichier dont chaque ligne suit le format de la question précédente, et retourne un dictionnaire associant à chaque couple (prenom,annee) le nombre de fois où le prénom a été choisi pendant cette année.

Question 8

Écrire une fonction `extraire_prenoms(d)` qui prend en entrée un dictionnaire dont les clefs sont des paires `(prenom, annee)` et retourne la liste des prénoms sans doublons.

Question 9

Répondre aux questions suivantes. Elles portent sur les données du fichier suivant :

Voir prenoms.txt. [1]

Combien y a-t-il de prénoms doubles (au sens de la fonction `double(s)` ci-dessus) ? (N'oubliez pas que la première lettre est majuscule...)

qp1 :

Réponse :

Combien y a-t-il de prénoms contenant un double (au sens de la fonction `contientdouble(s)` ci-dessus) ?

qp2 :

Réponse :

Combien y a-t-il de prénoms qui contiennent 5 fois la même lettre ?

qp3 :

1. `Fichier fourni avec le sujet.`

Réponse :

Combien y a-t-il de prénoms qui contiennent au moins 4 voyelles consécutives ?

qp4 :

Réponse :

Combien y a-t-il de prénoms qui contiennent 4 consonnes consécutives ?

qp5 :

Réponse :

Combien y a-t-il de prénoms qui finissent par 4 consonnes consécutives ?

qp6 :

Réponse :