TELECOM Bretagne

INF 424 : Des grammaires formelles aux documents Web
TP 1 : Les expressions régulières sous Python

Responsables : Éric Cousin & Yannis Haralambous, Département Informatique

Le but de ce TP est de vous donner un aperçu de la puissance des expressions régulières «à la Perl». Nous allons utiliser leur implémentation dans le langage Python.

1 *Introduction*

Pour utiliser les expressions régulières sous Python il faut importer le module `re` :

```
import re
```

Une expression régulière est d'abord compilée et le résultat est stocké dans un objet `RegexObject`. On écrit une expression régulière dans une «chaîne brute Python» : une chaîne délimitée par `r"` et `"`. Exemple : `r"[a-z]+"` est l'expression régulière qui correspond aux chaînes formées d'une ou plusieurs lettres entre `a` et `z`.

Pour compliler cette expression et créer un objet `RegexObject` on écrit :

```
r = re.compile(r"[a-z]+")
```

L'objet `r` a plusieurs méthodes, nous allons en utiliser trois :

1. `findall` qui sert à appliquer l'expression régulière et à récupérer les sous-chaînes trouvées sous forme de liste Python ;
2. `finditer` qui sert à appliquer l'expression régulière et à récupérer les sous-chaînes trouvées sous forme d'itérateur Python ;
3. `sub` qui sert à appliquer une expression régulière, à remplacer les sous-chaînes trouvées par d'autres chaînes.

Il s'agit donc de fonctionnalités similaires aux «chercher» et «chercher / remplacer» des éditeurs de texte.

Une méthode plus simple, du nom de `match()` va simplement tester si une chaîne satisfait les contraintes imposées par l'objet expression régulière auquel on l'applique.

2 *Syntaxe des expressions régulières «à la Perl»*

Avant de voir l'utilisation des expressions régulières sous Python, un rappel de la syntaxe des expressions régulières «à la Perl» :

- `toto` va trouver les sous-chaînes `toto` ;
- `.` est un caractère quelconque, mis à part le passage à la ligne `\n` et le retour chariot `\r` ;
- `[ax123Z]` signifie : «un caractère quelconque parim `a`, `x`, `1`, `2`, `3` et `Z`» ;
- `[A-Z]` signifie : «un caractère quelconque dans l'intervalle de `A` à `Z`» ;
- le trait d'union sert à indiquer les intervalles mais peut faire partie des caractères recherchés s'il est placé à la fin : `[AZ-]` signifie : «un caractère quelconque parmi `A`, `Z` et `-`» ;
- on peut combiner à volonté les caractères énumérés et les intervalles : par exemple `[A-Za-z0-9. :?]` signifie «une lettre minuscule ou majuscule, un chiffre, un point, un deux-points, ou un point d'interrogation» ;
- les caractères `(`, `)`, `\`, `[`, `]` peuvent être recherchés, à condition de les protéger par un antislash : `\(`, `\)`, `\\`, `\[`, `\]` ;
- le symbole `^` placé après le crochet ouvrant indique que l'on va chercher le complémentaire de ce qui est placé entre les crochets. Exemple : `[^a-z]` va trouver un caractère quelconque *qui ne soit pas* une lettre entre `a` et `z` ;
- on dispose des quantificateurs suivants :
  1. `*` (zéro, une ou plusieurs fois),
  2. `+` (une ou plusieurs fois),
  3. `?` (zéro ou une fois),
  4. `{n,m}` (entre `n` et `m` fois),
  5. `{n,}` (plus de `n` fois) ;
- on dispose également des quantificateurs «non gourmands» suivants :
  1. `*?` (zéro, une ou plusieurs fois),
  2. `+?` (une ou plusieurs fois),
  3. `??` (zéro ou une fois),
  4. `{n,m}?` (entre `n` et `m` fois),
  5. `{n,}?` (plus de `n` fois) ;

  La différence entre quantificateurs «gourmands» et «non gourmands» provient du fait que les premiers vont trouver la sous-chaîne la plus longue respectant les contraintes alors que les deuxièmes vont trouver la chaîne la plus courte.
  Exemple : l'expression `[a-z]+` appliquée à «`mon ami Pierrot`» va trouver `mon` alors que `[a-z]+?` va trouver `m` (ce qui n'a que peu d'intérêt). Autre exemple (qui montre l'utilité des quantificateurs non gourmands) : l'expression `\(.+\)` appliquée à «`Brest (29) et Aix (13)`» va retourner `29) et Aix (13` puisque c'est la plus longue sous-chaîne délimitée par une parenthèse ouvrante et une parenthèse fermante. Par contre `\(.+?\)` va retourner d'abord `29` et ensuite `13` ;
- les symboles `^` et `$` servent à indiquer le début et la fin d'une chaîne. Par exemple : `^a.+` va trouver toutes les chaînes qui commencent par un `a`, `toto$` va trouver toutes les chaînes qui finissent par `toto`, `^ $` va trouver toutes les chaînes égales à un blanc ;
- l'opérateur «ou» `|` sert à indiquer un choix entre deux expressions ;
- on peut utiliser les parenthèses pour deux raisons :

1. pour délimiter une expression qui sera utilisée par l'opérateur «ou» ou à laquelle on va appliquer un quantificateur (exemple : `abc(toto)+` signifie «`abc` suivi d'un ou plusieurs `toto`»);
2. pour délimiter une sous-chaîne que l'on va récupérer par la suite. On appelle cette sous-chaîne, un «groupe».

Ce double usage des parenthèses peut être gênant : en écrivant `abc(toto)+` on fait de `toto` un groupe, même si on n'a pas l'intention de le récupérer par la suite. En écrivant `abc(?:toto)+` les parenthèses ne servent qu'au premier usage, aucun groupe n'est formé.

3 *Utilisation des expressions régulières sous Python*

3.1 Recherche

Supposons que l'on veuille trouver tous les mots de la chaîne «Le bon chasseur sachant chasser sait chasser sans son chien» contenant un «s». On peut trouver un tel mot en écrivant `[a-rt-z]*s[a-z]*`. On peut donc déjà compiler une expression régulière :

```
import re
r = re.compile(r"[a-rt-z]*s[a-z]*")
```

Pour l'appliquer à la chaîne on écrira :

```
m = r.findall("Le bon chasseur sachant chasser sait chasser sans son chien")
print m
```

Le résultat sera une liste Python contenant tous les mots trouvés :

```
['chasseur', 'sachant', 'chasser', 'sait', 'chasser', 'sans', 'son']
```

On peut ré-écrire le code précédent en utilisant un *itérateur* Python :

```
import re
r = re.compile(r"[a-rt-z]*s[a-z]*")
for m in r.finditer("Le bon chasseur sachant chasser sait chasser sans son chien"):
    print m.group()
```

L'avantage de cette écriture est que l'on récupère non pas des simples chaînes de caractères mais des objets `MatchObject` qui ont leurs propres méthodes et attributs.

3.2 Recherche / remplacement

Maintenant nous allons essayer de rendre la phrase «Le bon chasseur sachant chasser sait chasser sans son chien» conforme au dialecte chti-mi. On peut commencer par remplacer tous les «s» par des «ch» :

```
import re
r = re.compile(r"s")
m = r.sub(r"ch","Le bon chasseur sachant chasser sait chasser sans son chien")
print m
```

Le résultat est «Le bon chachcheur chachant chachcher chait chachcher chanch chon chien», qui est relativement imprononçable. On peut rectifier le tir en évitant les doubles «ch». On va donc remplacer un *nombre quelconque de lettres s consécutives* par un seul «ch» :

```
import re
r = re.compile(r"s+")
m = r.sub(r"ch","Le bon chasseur sachant chasser sait chasser sans son chien")
print m
```

Le résultat «Le bon chacheur chachant chacher chait chacher chanch chon chien» est nettement plus chti-mi, mais il reste un cas problématique : le «s» muet du mot «sans» est devenu un «ch» prononcé dans «chanch». Il faut donc éviter de convertir les «s» en fin de mot :

```
import re
r = re.compile(r"s+([a-z]+)")
m = r.sub(r"ch\1","Le bon chasseur sachant chasser sait chasser sans son chien")
print m
```

Pour ce faire, on a créé un groupe `([a-z]+)` que l'on retrouve dans la chaîne de remplacement (`\1`). Le réchultat «Le bon chacheur chachant chacher chait chacher chans chon chien» est chans contechte parfaitement chti-mi.

3.3 Recherche / remplacement avec utilisation de fonction

Lorsqu'on remplace une sous-chaîne par une autre il peut être utile d'intercaler un traitement entre lecture de la sous-chaîne et écriture dans la nouvelle chaîne. Python nous permet d'appliquer une fonction à chacune des sous-chaînes trouvées. Imaginons que dans la chaîne `toto 123 blabla 456 titi` on veut représenter les nombres en hexadécimal. Ce calcul est trop compliqué pour être fait uniquement par des expressions régulières, on utilisera donc une fonction

```
def ecrire_en_hexa ( entree ):
    return hex( int( entree.group() ) )
```

et on écrira :

```
import re
r = re.compile(r"[0-9]+")
m = r.sub( ecrire_en_hexa,"toto 123 blabla 456 titi" )
print m
```

Le résultat est bien `toto 0x7b blabla 0x1c8 titi`. L'argument de la fonction est un objet de type `MatchObject`. La méthode `group()` fournit la chaîne tout entière, alors que `group(`n`)` fournira le n-ième groupe de la sous-chaîne.

4 Exercices

Récupérer sur Moodle le fichier `gen1551.csv`. Pour le lire ligne par ligne, utiliser le code Python suivant :

```
f = open("gen1551.csv",'r')
for ligne in f:
    #faire qqch avec la ligne ligne
f.close()
```

4.1 Exercice 1

Que fait le code suivant ?

```
import re
r = re.compile(r"^([0-9]+);[^;]*;PAUL;")
f = open("gen1551.csv",'r')
for ligne in f:
    for m in r.finditer(ligne):
        print m.group(1)+" OK"
f.close()
```

4.2 Exercice 2

Complétez ce programme afin qu'il sorte les identifiants des gens nés dans un village dont le nom commence par PLOU.

4.3 Exercice 3

Remplacez les lieux de naissance des personnes trouvées dans l'exercice 2 par des lieux qui commencent par LOC (par exemple : PLOUNEVEZ devient LOCNEVEZ).

Rappel : pour écrire dans un fichier on utilise le code suivant :

```
o = open("fichier_sortie",'w')
o.write("texte à écrire")
o.close()
```

4.4 Exercice 4

Incrémentez les dates de naissance des personnes trouvées dans l'exercice 2 de 10 ans.

Tuyaux :

1. définir une fonction `traiterdate` qui va gérer le remplacement de la chaîne qui nous intéresse ;
2. en Python on passe d'un objet nombre entier à un objet chaîne en utilisant `str`, pour l'opération inverse on dispose de la fonction `int`.

4.5 Exercice 5

Compter le nombre de fiches où le nom de la personne est ABALAIN.

4.6 Exercice 6

Calculez l'âge moyen lors des mariages, en considérant que tous les mois ont 30 jours. À noter que les dates de naissance et de mariage sont données par les champs DN et DM.

Tuyau : si a_m, m_m, d_m sont resp. l'année, le mois et le jour de mariage et a_n, m_n, d_n de même pour la naissance, l'âge d'une personne lors du mariage peut être exprimé par la formule

$$A = \left(a_m + \frac{m_m - 1}{12} + \frac{d_m - 1}{360}\right) - \left(a_n + \frac{m_n - 1}{12} + \frac{d_n - 1}{360}\right).$$