Pratique des machines
TP5 : Expressions régulières

am@up8.edu

Octobre 2022

Dans ce TP :
— Découverte des expressions régulières

Avant de commencer :
— créez un répertoire TP5 dans le répertoire dédié à ce cours.

Avant de partir :
— déposez le fichier **tests_regexp.sh** (et le cas échéant le fichier chaines_de_test.txt) sur le moodle

Syntaxe des expressions régulières

Une expression régulière est une séquence de caractère qui décrit un ensemble de séquences de caractères.
La plus simple exemple est la séquence "abc" qui correspond à la séquence "abc".
Les expressions régulières utilisent certains caractères pour désigner des ensembles de caractères :

— **.** (le point), n'importe quel caractère : **.bc** désigne abc, bbc, cbc, etc.
— [ ] (les crochets) désignent un ensemble (ou un intervalle) de caractères possibles (*ou impossibles* avec un chapeau en début) :
  — [**ab**]**bc** désigne abc et bbc ;
  — [**^ab**]**bc** désigne cbc, dbc, ebc, etc. mais pas abc et bbc
  — [**a-c**]**bc** désigne abc, bbc, cbc
  ;
— | (le pipe), une alternative : a|b désigne a ou b
— **?** signifie : 0 ou une fois : abc ? désigne abc et ab
— ***** signifie : 0, une ou plusieurs fois : **a*bc** désigne bc, abc, aabc, aaabc,aaabc, etc.
— **+** signifie : une ou plusieurs fois : **a+bc** désigne abc, aabc, aaabc, etc.
— **()** (les parenthèses) groupe une séquence, cela s'applique en combinaison avec les opérateurs précédents : **(ab)+c** désigne abc, ababc, abababc etc.
— ^ (le chapeau), désigne le début d'une ligne
— **$** (le dollar), désigne la fin d'une ligne

1 Exercice : Ecriture d'expressions régulières et test avec la commande grep

Dans cet exercice, je vous donne une suite de **motifs**. Pour chaque motif, vous devez :
— écrire l'expression régulière correspondante ;
— la tester sur une chaîne de caractère.

Par exemple, si on souhaite reconnaître le motif "ri" suivi d'un caractère, je peux écrire l'expression régulière : "**ri.**". Pour tester mon expression, je peux par exemple exécuter la commande suivante :

```
1 echo "ri, tri, rit, ribambelle, Paris, il a ri" | grep -E --color "ri."
```

Les occurrences du motif apparaissent en couleur, que remarque-t-on au passage ?

Pour faire un test sur plusieurs lignes, on peut utiliser "\n" dans une chaîne de caractères, par exemple la commande suivante permet de reconnaître la lettre « u » en début de ligne seulement :

```
1 echo "un\ndeux\ntrois" | grep -E --color "^u"
```

Vous pouvez également créer un fichier texte chaines_de_test.txt dans laquelle vous écrivez les chaînes à tester, et appliquer votre commande **grep** :

```
1 cat chaines_de_test.txt | grep -E --color "^u"
```

Faites de même en notant dans un fichier **tests_regexp.sh** vos commandes de test les unes à la suite des autres pour les motifs suivants.

Attention c'est bien TOUT le motif qui doit apparaître en couleur.

1. lettre K
2. séquence "cha"
3. lettre 't' suivie de la lettre 'r' ou de la lettre 'h'
4. chaînes de caractère "plop" et "plip"
5. en début de ligne, lettre 'j' suivie de la lettre 'e'
6. ligne qui se termine par ' :'
7. les chaînes de caractères commençant par "ah" avec un nombre inconnu de h : "ah", "ahh", "ahhh", "ahhhhh", etc.
8. n'importe quel chiffre entre 0 et 9
9. n'importe quel nombre à trois chiffres entouré de deux espaces
10. chaînes de caractères terminant par ".txt"
11. ligne qui commence par une majuscule et qui finit par un point
12. chaîne qui commence par un 3 et alterne avec n'importe quel autre chiffre sauf 4 (par exemple "353637", ou "36", ou "3131313131")
13. toute chaîne de la forme 0101..0101010 : séquence de 0 et de 1 terminant par 0
14. n'importe quel nombre terminant par 323
15. les numéros de téléphone portable (commençant par 06 ou 07)
16. les adresses e-mail valides (d'un point de vue formel, pas des noms de domaine existants)