

Syntaxe des expressions régulières en Python

Il y a 14 *méta-caractères* à précéder d'un backslash pour les capturer :

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

Classe de caractères

Représente un ensemble de caractères :

[abc]	a, b et c
[^abc]	tout sauf a, b et c
[0-9]	tous les caractères compris entre 0 et 9
[a-zA-Z]	toutes les lettres de l'alphabet romain

Il existe des classes prédéfinies :

.	n'importe quel caractère (sauf retour à la ligne)
\d	un chiffre décimal ([0-9])
\s	un caractère blanc ([\t\n\r\f\v])
\w	un caractère alphanumérique ([a-zA-Z0-9_])
\D, \S et \W	complémentaires des ensembles précédents

Répétition

Permet de matcher plusieurs fois une même expression régulière R :

$R\{n\}$	exactement n occurrences
$R\{m, n\}$	au moins m et au plus n occurrences
$R\{m, \}$	au moins m occurrences
$R\{, n\}$	au plus n occurrences

Il existe des répétitions prédéfinies :

R^*	zéro ou plusieurs occurrences ($\{0, \}$)
R^+	une ou plusieurs occurrences ($\{1, \}$)
$R^?$	zéro ou une occurrence ($\{0, 1\}$)

Par défaut, la chaîne analysée sera consommée au maximum possible avec les répétitions. Pour ne pas être glouton et consommer le moins possible, on ajoute un $?$ en suffixe :

$\{m, n\}^? \quad *^? \quad +^? \quad ??$

Frontière de recherche

\sim	début de la ligne
$\$$	fin de la ligne
$\backslash A$	début de la chaîne de caractères
$\backslash Z$	fin de la chaîne de caractères
$\backslash b$	limite d'un mot (début ou fin)
$\backslash B$	pas la limite d'un mot

Les frontières ne consomment aucun caractères dans la chaîne, elles représentent simplement une assertion.

Alternative

$R_1 | R_2$ choix entre les deux expressions régulières R_1 et R_2

Groupe

Permet de délimiter une sous-expression régulière R , pour par exemple la répéter ou pour former une alternative :

(R)	groupe capturant
$(?:R)$	groupe non-capturant
$(?P=name R)$	groupe capturant nommé $name$

Il est possible de faire référence à la valeur capturée par un groupe précédent :

$\backslash i$	valeur capturée par le groupe numéro i
$(?P=name)$	valeur capturée par le groupe nommé $name$

Lookahead

Permet de tester une assertion R , sans avancer dans la chaîne :

$(?=R)$	succès si R matche à la position courante
$(?!R)$	succès si R ne matche pas à la position courante