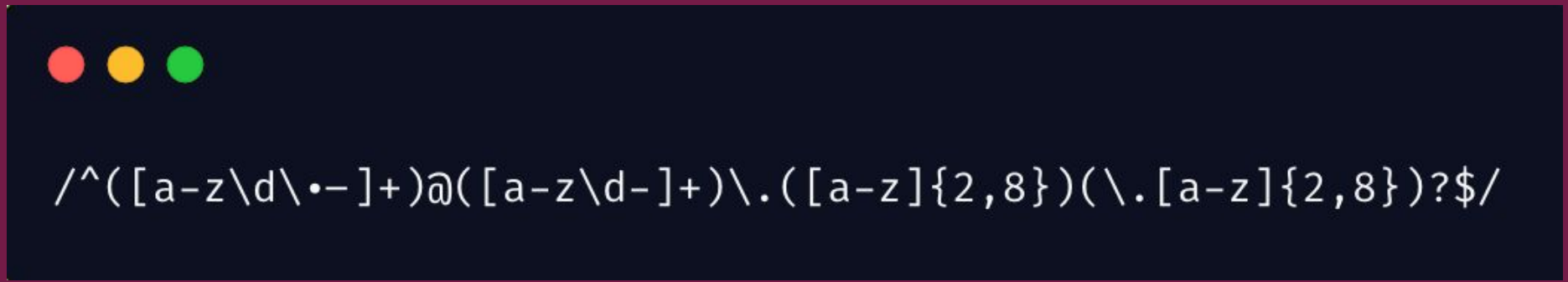
Petit guide sur les expressions régulières



Regex ?

REG = REGular

EX = EXpression

REGEX = expression régulière

"décrit un ensemble de chaînes de caractères selon une syntaxe précise"

http://fr.wikipedia.org/wiki/Regex

POSIX et PCRE

En PHP, il existe deux familles de RegEx. Elle font toutes la même chose : chercher des motifs au sein d'une chaîne de caractères. Bien qu'elles partagent des fois les mêmes caractères spéciaux au sein des models, il existe quand même de nettes différences entre les deux. Ces 2 familles sont:

- **POSIX**: il s'agit d'une famille de RegEx utilisée en PHP. POSIX est l'acronyme de Portable Operating System Interface.
- **PCRE**: signifie Perl Compatible Regular Expressions. C'est une famille de RegEx qui vient du langage PERL. En JavaScript, seules les regex PCRE sont supportées

La principale différence entre les deux familles c'est le temps d'exécution. En effet, les PCRE sont plus rapides que les POSIX, surtout quand il s'agit d'expressions régulières longues et complexes. Une autre différence se manifeste dans la syntaxe. Bien que la plupart des caractères spéciaux des expressions peuvent être utilisés pour les deux familles. Il en existe qui sont propres à chaque famille.

Utilisation des RegEx

Si vous êtes développeur, j'espère que vous en voyez l'utilité (réécriture d'url, extraire des numéro de téléphone d'une page web, ou vérifier que l'email saisi dans un formulaire ressemble bien à un email, etc …)

Si vous n'êtes pas développeur ou webmaster, les expressions régulières peuvent quand même vous servir

Ce sont des outils très puissants et très utilisés : on peut les retrouver dans de nombreux langages comme le PHP, MySQL, Javascript... ou encore dans des logiciels d'édition de code !

Les crochets [ ]

Désigne un élément parmi une liste

[abcd] = a ou b ou c **ou** d
[a-z] = une lettre en minuscule, entre a et z
[A-Z] = une lettre en majuscule, entre A et Z
[0-9] = un chiffre entre 0 et 9
[123] = 1 ou 2 ou 3
[a-zA-Z0-9] = une lettre en minuscule ou en majuscule ou un chiffre
- un caractère alphanumérique

Exemples [ ]

[a-c] correspond à a, b ou **c**
[a-cA-D] correspond à une lettre minuscule entre a et c, ou une lettre majuscule entre A et D

Les réponses possibles sont a, b, c, A, B, C, D

Les accolades { }

Indique combien de fois est répété un élément

{3} est répété exactement 3 fois
{3,} est répété au moins 3 fois
{3,5} est répété entre 3 et 5 fois

Exemples { }

[a,b]{1} correspond à a ou b

[a,b]{1,2} correspond à a, b, aa, ab, ba, bb

[ab]{2,} correspond à a ou b répété au moins 2 fois

soit : aa, aaa, aaaaaaaaaa, bb, bbbbbbbbbb

Les parenthèses ( )

Elles ont la même utilité (ou inutilité) qu'en mathématiques
Elles désignent un groupe

(http://)(www)(monsite)(com)

L’accent circonflexe ^ (caret)

Indique le début d’une chaîne de caractère

^http = commence par http

Caret entre crochet

Mais entre crochets, il exclut des caractères

[^a] = un caractère qui n’est pas a

peut correspondre à b, c, d, 1, 2, 3, Z, K, etc …

Exemple

^[^a] : commence par un caractère qui n’est pas a

Attention

Dans [^(http)] , les parenthèses ne changent pas l'interprétation de l'expression régulière.

[^(http)] = ne correspond pas à h t ou p

Le signe dollar $

Indique la fin d’une chaîne de caractères

au revoir$ = se termine par au revoir

Le point .

Indique un caractère unique quel qu'il soit

c.t peut correspondre à cat, cbt, c3t, clt, …

Il y a un caractère présent entre c et t

Le signe +

Le signe + indique qu'un caractère est répété une fois ou plus (au moins 1 fois)

Si on prend l'écriture avec les accolades, cela équivaut à :

{1,}

chu+t = chu{1,}t = chut ou chuuuuuuuuuuuuuuuuuuut

Le point d’interrogation ?

Il indique qu’un caractère est optionnel, il est présent 0 ou 1 fois (au plus 1 fois)

Avec les accolades, il équivaut à {0,1}

Matt?hieu = Matt{0,1}hieu = Mathieu ou Matthieu

Astuce

Lorsqu’on utilise les parenthèses, tout le groupe encadré devient optionnel

I (don’t)? like RegEx = I don’t like RegEx ou I like RegEx

L’astérisque *

Il indique qu’un caractère est peut être présent, un nombre indéterminé de fois

Avec les accolades, cela équivaut à {0,}

chu*t = cht ou chut ou chuuuuuuuuuuuuuuuuuuuut, etc …

Le point et L’astérisque : .*

Cette combinaison permet d’englober beaucoup de cas :

On indique qu’un caractère unique est répété un nombre infini de fois, ou pas.

Globalement, cela encadre tout.

Le pipe |

Cette opérateur correspond à ou

a | b = a ou b

a | b | c | d = a ou b ou c ou d = [abcd] = [a-d]

Le backslash \

Il est utilisé de plusieurs façons.

Il permet d'annuler l'opérateur et de prendre en compte le signe de ponctuation

Exemple : www\.monsite\.fr

Le point n'est pas l'opérateur qui désigne un caractère unique, mais bel et bien un point.

\w \d et \s

Ils désignent des groupes de caractères.

\d = un chiffre (un caractère numérique) = [0-9]

\w = un caractère alphanumérique = [a-zA-Z0-9]

\s = un espace, retour à la ligne, etc…

Mnémotechnique

d = digit

w = word

s = space

\W \D \S

Mettre des majuscules permet de désigner l’inverse du groupe désigné en minuscule

\D = un caractère qui n’est pas un chiffre = [^0-9]

\W = un caractère qui n’est pas alphanumérique

\S = un caractère qui n’est pas un espace

En savoir plus sur les RegEx

Pour vous faire la main, je vous conseille des mots croisés :

https://regexcrossword.com/

De quoi tester les expressions régulières en ligne :

https://regex101.com/

https://rubular.com/

Exercices sur les RegEx

Exercice 1 :

Savoir si une chaîne de caractères est une couleur hexadécimale ou non.

Exercice 2 :

Savoir si une chaîne de caractères est un mail valide (TLD de 2 à 6 caractères)

Exercice 3 :

Savoir si une chaîne de caractères est un numéro de téléphone au format : 0652320478 ou +33652320478.

Exercices sur les RegEx

Exercice 4 :

Savoir si un mot de passe contient bien au moins 2 majuscules, 1 minuscule, 2 chiffres et un signe & ou # ou @, avec une longueur est de 8.

Pour vous aider :

lookahead-lookbehind

Conception - Stéphane PONTONNIER - 2022

Formateur Développeur Web & Web Mobile