Mémo Expressions Régulières

Les **expressions régulières** sont un moyen puissant de recherche de chaîne de caractères proposé dans de nombreux instruments de linguistique de corpus. Elles permettent de rechercher non pas une séquence de caractères figée, telle que la séquence « pomme », mais un motif regroupant plusieurs formes, tel que le **motif** « pommes? » permettant de repérer les formes « pomme » et « pommes » en une seule requête.

La syntaxe des expressions régulières peut varier d'un langage de programmation ou d'un logiciel à l'autre. Lorsqu'elle n'est pas spécifiée dans la documentation du langage ou du logiciel, elle correspond à celle présentée ci-dessous.

Dans cette syntaxe, on distingue les **caractères littéraux** de **caractères spéciaux**, aussi appelés **méta-caractères**. Par caractère littéral, on entend le caractère tel qu'on souhaite le voir apparaître dans le texte. Par exemple, si l'on recherche des mots contenant la lettre « N », on dira que le motif doit comporter le caractère littéral « N ».

<table>
<tr><th>Code</th><th>Description</th></tr>
<tr><td>\</td><td>Marque le caractère suivant comme caractère spécial ou littéral.
Par exemple, "n" correspond au caractère "n". "\n" correspond à un caractère de saut de ligne. La séquence "\\" correspond à "\", tandis que "\(" correspond à "(".</td></tr>
<tr><td>^</td><td>Permet de spécifier la position début de la saisie. (bien souvent début de ligne)</td></tr>
<tr><td>$</td><td>Permet de spécifier la position fin de la saisie. (bien souvent fin de ligne)</td></tr>
<tr><td>( )</td><td>Permet de délimiter un groupe de caractères.</td></tr>
<tr><td>*</td><td>Permet de rechercher une chaîne contenant zéro ou plusieurs fois le caractère, ou groupe de caractères, qui précède.
Ainsi, "zo*" permet de trouver "z" et "zoo", "(zo)*" permet de trouver "zozo".</td></tr>
<tr><td>+</td><td>Permet de rechercher une chaîne contenant une ou plusieurs fois le caractère, ou groupe de caractères, qui précède.
Ainsi, "zo+" permet de trouver "zoo", mais pas "z".</td></tr>
<tr><td>?</td><td>Permet de rechercher une chaîne contenant zéro ou une fois le caractère, ou groupe de caractères, qui précède.
Ainsi, "a?(ve)?" permet de trouver ve dans "lever", mais rien dans "lèvre".</td></tr>
<tr><td>.</td><td>Permet de rechercher n'importe quel caractère unique, sauf le caractère de nouvelle ligne.</td></tr>
<tr><td>x|y</td><td>Permet de rechercher soit x soit y.
Par exemple, "z|foot" permet de trouver "z" et "foot". "(z|f)oot?" permet de trouver "zoo" et "foot".</td></tr>
<tr><td>{n}</td><td>n est un nombre entier non négatif. Permet de rechercher un chaîne contenant exactement n fois le caractère, ou groupe de caractères, qui précède .
Par exemple, "o{2}" ne permet pas de trouver "Bob", qui ne contient qu'un seul "o" mais correspond aux deux "o" de "foot" et "zoo".</td></tr>
</table>

Code	Description
{n,}	n est un entier non négatif. Permet de rechercher un chaîne contenant **au moins n fois** le caractère, ou groupe de caractères, qui précède. Par exemple, "o{2,}" ne correspond pas à "o" dans "Bob", mais à tous les "o" dans "fooooot". "o{1,}" équivaut à "o+" et "o{0,}" équivaut à "o*".
{n,m}	m et n sont des entiers non négatifs. Permet de rechercher un chaîne contenant **au moins n et au plus m fois** le caractère, ou groupe de caractères, qui précède. Par exemple, "fo{1,3}t" permet de trouver "fot" , "foot" et "fooot". "o{0,1}" équivaut à "o?".
[xyz]	Jeu de caractères. Permet de rechercher un chaîne contenant **l'un des caractères indiqués**. Par exemple, "[abc]" correspond à "a" dans "plat".
[^xyz]	Jeu de caractères négatif. Permet de rechercher un chaîne contenant **tout caractère non indiqué**. Par exemple, "[^abc]" correspond à "p", "l" et "t" dans "plat".
[a-z]	Série de caractères. Permet de rechercher un chaîne contenant **un caractère de la série spécifiée**. Par exemple, "[a-z]" correspond à tout caractère alphabétique minuscule compris entre "a" et "z", [0-9] à tout caractère numérique compris entre 0 et 9.
[^m-z]	Série de caractères négative. Permet de rechercher une chaîne contenant **un caractère ne se trouvant pas dans la série spécifiée**. Par exemple, "[^m-z]" correspond à n'importe quel caractère sauf les caractères alphabétiques qui se trouvent entre "m" et "z" dans l'ordre alphabétique.
\b	Permet de spécifier la position **limite de mot**, autrement dit, la position entre un mot et une espace, ou un signe de ponctuation. Par exemple, "er\b" correspond à "er" dans "lever", mais pas à "er" dans "verbe".
\B	Permet de spécifier la position **autre que limite de mot**. Par exemple, "en*t\B" correspond à "ent" dans "bien entendu", mais pas à "ent" dans "scient".
\d	Permet de rechercher **un caractère représentant un chiffre**. Équivaut à [0-9].
\D	Permet de rechercher **un caractère ne représentant pas un chiffre**. Équivaut à [^0-9].
\n	Correspond à **un caractère de saut de ligne**.
\s	Correspond à **tout espace blanc**, y compris l'espace, la tabulation, le saut de page, etc. Équivaut à "[\f\n\r\t\v]"
\S	Correspond à **tout caractère qui n'est pas un espace blanc**. Équivaut à "[^\f\n\r\t\v]"
\t	Correspond **au caractère de tabulation**.
\w	Correspond à **tout caractère permettant d'écrire un mot, y compris le trait de soulignement**. Équivaut à "[A-Za-z0-9_]".
\W	Correspond à **tout caractère autre que les caractères permettant d'écrire un mot**. Équivaut à "[^A-Za-z0-9_]".