TD COLD #4

EXPRESSIONS REGULIERES

REGEX

Il arrive souvent en informatique que l'on recherche à l'intérieur de chaines de caractères des sous-chaines spécifiques. Parfois, ces sous-chaines sont précises (la chaine « 54000 »), parfois tout ce que l'on a est une spécification (on recherche un code postal, qui est composé exactement de 5 chiffres).

Les expressions régulières sont un moyen de définir cette spécification sous la forme d'un **motif** (**pattern** en anglais), et ensuite d'utiliser cette spécification pour faire de la **recherche de motifs** dans une chaine (**pattern matching**). En effet, il est plus pratique, voire indispensable, de travailler avec un motif plutot que de spécifier toutes les solutions possibles (imaginez la liste de tous les codes postaux possibles).

Par exemple, la spécification « exactement 5 chiffres » se traduit par : « \d{5} »

Le format des expressions régulières est assez standardisé de nos jours (bien qu'il y ai quelques différences de temps en temps). Certains caractères (« ()[]\^$.*? ») ont un sens particulier, qui est décrit ci-dessous. Ces caractères sont combinables entre eux pour obtenir le motif voulu. Pour matcher le caractère sans son sens spécial, il faut le « déspécialiser » en mettant un « \ » devant.

Ensembles

La première chose à spécifier est l'ensemble des caractères admis, ou à l'inverse ceux qui ne le sont pas.

a

Matche le caractère « a ». Fonctionne avec n'importe quel caractère, mais attention, il faut peut-etre déspécialiser certains caractères avec un « \ » (par exemple si on veut matcher le caractère « [»).

[*aef*]

Ne matche que les caractères présents entre les crochets.

[*a-f*]

Matche tous les caractères présents (alphabétiquement) entre le « a » et le « f ».

[^*aef*]

Matche tous les caractères SAUF ceux présents entre les crochets.

\s

Matche n'importe quel caractère d'espacement.

\S

Matche n'importe quel caractère SAUF ceux d'espacement.

\d

Matche n'importe quel caractère de chiffre.

\D

Matche n'importe quel caractère SAUF ceux de chiffre.

\w

Matche n'importe quel caractère pouvant faire parti d'un mot (lettre, nombre, underscore).

\W

Matche n'importe quel caractère sauf ceux pouvant faire parti d'un mot (lettre, nombre, underscore).

. (le caractère point)

Matche n'importe quel caractère.

Numération

La deuxième chose à spécifier est le nombre de fois où un élément peut ou doit être répété.

a?

Indique que le caractère « a » doit être présent 0 ou 1 fois.

*a**

Indique que le caractère « a » doit être présent 0 fois ou plus (pas de borne maximum).

a+

Indique que le caractère « a » doit être présent au moins une fois (pas de borne maximum).

a{5}

Indique que le caractère « a » doit être présent exactement 5 fois.

a{3,}

Indique que le caractère « a » doit être présent au moins 3 fois.

a{3,5}

Indique que le caractère « a » doit être présent au moins 3 fois mais au plus 5 fois.

Ancres

^a

Indique que le caractère « a » doit être recherchée à partir du début de la chaine.

a$

Indique que le caractère « a » doit être recherchée à partir de la fin de la chaine.

Choix

(aa|bb)

Matche soit « aa », soit « bb », mais pas « ab » ni « ba ».

Groupe

(*ab*)

Regroupe un pattern, par exemple pour lui appliquer une numération : (ab)+ matche « ab », « abab », mais pas « aa » ou « aab ».