

SÉANCE 10

EXPRESSIONS RÉGULIÈRES PRATIQUES

SUJETS

Notations pratiques qui sont souvent utilisées dans les expressions régulières

Quelques exercices de pratique

TRUCS POUR LES EXPRESSIONS RÉGULIÈRES PRATIQUES

Nous verrons des trucs pour les expressions régulières pratiques qui sont supportées par la plupart des librairies de regex

Souvenez-vous, les expressions régulières ne sont pas utilisées uniquement dans le contexte des compilateurs

- Nous les utilisons souvent pour extraire l'information d'un texte

Exemple: Imaginez que vous cherchez dans un fichier log qui accumule des entrées depuis deux mois pour un motif d'erreur particulier

- Sans les expressions régulières, ceci sera une tâche ennuyante

MATCHER LES CHIFFRES

Afin de matcher un chiffre simple, comme on a vu précédemment, on peut utiliser l'expression régulière suivante:

[0-9]

Cependant, puisque l'opération de matcher un chiffre est une opération très commune, on peut utiliser la notation suivante:

\d

Un slash est un caractère d'échappement utilisé pour le distinguer de la lettre **d**

Similairement, afin de matcher un caractère non-chiffre, on peut utiliser la notation:

\D

CARACTÈRES ALPHANUMÉRIQUES

Afin de matcher un caractère alphanumérique, on peut utiliser la notation:

[a-zA-Z0-9]

Ou on peut utiliser le raccourci suivant:

\w

Similairement, on peut représenter n'importe quel caractère non-alphanumérique comme suit:

\W

CARACTÈRE GÉNÉRIQUE (WILDCARD)

Un caractère générique est défini pour matcher n'importe quel caractère simple (lettre, chiffre, espace blanc ...)

Il est représenté par le caractère **.** (point)

Alors, afin de matcher un point, vous devez utiliser le caractère d'échappement: **\.**

EXCLUSION

Nous avons déjà vu que **[abc]** est équivalent à **(a | b | c)**

Mais, parfois on veut matcher tout sauf un ensemble de caractères

Pour accomplir ceci, on peut utiliser la notation: **[^abc]**

- Ceci matche n'importe quel caractère simple sauf a, b ou c

Cette notation peut aussi être utilisée avec les classes de caractères abrégés

- **[^a-z]** matche n'importe quel caractère sauf les lettres minuscules

RÉPÉTITIONS

Comment peut-on matcher une lettre ou une chaîne qui se répète plusieurs fois :

- Ex. ababab

Jusqu'à maintenant, nous avons implémenter des répétitions à travers trois mécanismes:

- Concaténation: simplement concaténer la chaîne ou le caractère avec lui-même (ne fonctionne pas si vous ne connaissez pas le nombre exact de répétitions)
- Fermeture étoile Kleene: pour matcher des lettres ou des chaînes répétés 0 ou plusieurs fois
- Fermeture positive: pour matcher des lettres ou des chaînes répétés 1 ou plusieurs fois

RÉPÉTITIONS

On peut aussi spécifier une gamme de combien de fois une lettre ou une chaîne peuvent être répétées

Exemple, si on veut matcher des chaînes de répétition de la lettre « a » entre 1 et 3 fois, on peut utiliser la notation: **a {1,3}**

- Alors, **a {1,3}** matche les chaînes
 - a
 - aa
 - aaa

On peut aussi spécifier un nombre exact de répétitions au lieu d'une gamme

- **(ab) {3}** matche la chaîne **ababab**

CARACTÈRES OPTIONNELS

Le concept du caractère optionnel est plus ou moins similaire à l'étoile kleene

- L'opérateur étoile matche **0 ou plus** cas de l'opérande
- L'opérateur optionnel, **représenté par ? (point d'interrogation)**, matche **0 ou un** cas de l'opérande

Exemple: le motif **ab?c** matchera soit les chaînes **"abc"** ou **"ac"** parce que **b** est considéré optionnel.

ESPACE BLANC

Souvent, on veut facilement identifier les espaces blancs

- Soit pour les enlever ou pour identifier le début ou la fin de mots

Les espaces blanc les plus utilisés avec les expressions régulières:

- Espace `_`, le caractère de tabulation `\t`, la nouvelle ligne `\n` et le retour chariot `\r`

Un caractère spécial d'espace blanc `\s` matchera n'importe quel espace blanc spécifiez ci-haut

Similairement, vous pouvez matcher n'importe quel autre caractère en utilisant la notation `\S`

MATCHER LE DÉBUT ET LA FIN D'UNE LIGNE

Vous pouvez spécifier qu'une expression régulière matche seulement le début ou à la fin de la ligne

Si un caret (^) est au début de toute l'expression régulière, il correspond au début d'une ligne

Si un signe dollar (\$) se trouve à la fin de toute l'expression régulière, il correspond à la fin d'une ligne

Si une expression régulière entière est entourée d'un caret et d'un signe dollar (^ c'est une phrase \$), elle correspond à une ligne entière

MATCHER LE LIMITE DES MOTS

Des fois, c'est important de matcher les limites des mots

- Cela nous permet de distinguer les chaînes qui apparaissent au début d'un mot de celles qui apparaissent au milieu

Le début d'un mot est matché en utilisant `\b`

- Par exemple: `\bart` matche la chaîne "art" mais non "start"

La fin d'un mot est aussi matché par `\b`

- Par exemple: `art\b` matche la chaîne "start" mais non "arts"

De même, `\B` matche le milieu d'un mot

- Par exemple: `\Bart` matche la chaîne "start" mais non "art"

QUELQUES EXERCICES - 1

Étant donné la phrase:

- Error, computer will now shut down...

Développez une expression régulière qui matchera tous les mots dans la phrase

Réponse:

[A-Za-z]+

or

\p{L}+

QUELQUES EXERCICES - 2

Étant donné la phrase:

- Error, computer will now shut down...

Développez une expression régulière qui matchera tous les caractères non-alphanumériques

Réponse:

\W+

QUELQUES EXERCICES - 3

Étant donné le fichier log:

- [Sunday Feb. 2 2017] Program starting up
- [Monday Feb. 3 2017] Entered initialization phase
- [Tuesday Feb. 4 2017] Error 5: cannot open XML file
- [Thursday Feb. 6 2017] Warning 5: response time is too slow
- [Friday Feb. 7 2017] Error 9: major error occurred, system will shut down

Matcher n'importe quel message d'erreur (Error) ou d'alerte (Warning) qui termine par le terme “shut down”

Réponse:

(Error|Warning).*(shut down)

QUELQUES EXERCICES - 4

Étant donné le fichier log:

- [Sunday Feb. 2 2017] Program starting up
- [Monday Feb. 3 2017] Entered initialization phase
- [Tuesday Feb. 4 2017] Error 5: cannot open XML file
- [Thursday Feb. 6 2017] Warning 5: response time is too slow
- [Friday Feb. 7 2017] Error 9: major error occurred, system will shut down

Matcher n'importe quel erreur ou alerte entre le 1 et 6 février

Réponse:

`\\w+ Feb\\. [1-6] 2017\\] (Error|Warning)`

QUELQUES EXERCICES - 4

Spécifiez l'expression régulière pour les mots de passe qui doivent commencer par une lettre suivie de caractères alphanumériques

Le mot de passe doit contenir au moins 8 caractères et un maximum de 15 caractères

`\b[a-zA-Z]\w{7,14}\b`

MERCI!

QUESTIONS?