Les expressions régulières

François Gannaz

9 décembre 2014

Recherche simple

- `vrai` est une expression régulière
- Cherche cet enchaînement de 4 caractères
- On l'écrit souvent `/vrai/`, où les `/` entourent l'**expression régulière** (regex, regexp)
- L'intérieur (ce qu'on cherche) est appelé *motif* (*pattern*)

Mais c'est insuffisant

- Comment chercher tous les chiffres d'un texte ?

Mais c'est insuffisant

- Comment chercher tous les chiffres d'un texte ?

- Pour chercher, non un caractère, mais un parmi plusieurs : les lister entre [et]
- Ici `/[0123456789]/`, ou plus simplement `/[0-9]/`
- C'est une **classe de caractères**

Où peut-on utiliser les expressions régulières ?

- Bureautique (LibreOffice, OpenOffice, Word...)
- Éditeurs de texte
- Les IDE pour développeurs
- En ligne de commande (Linux, Mac) avec grep, sed, perl

Où peut-on utiliser les expressions régulières ?

- Bureautique (LibreOffice, OpenOffice, Word…)
- Éditeurs de texte
- Les IDE pour développeurs
- En ligne de commande (Linux, Mac) avec grep, sed, perl

Passer en minuscules les balises de tous les fichiers XML dans le répertoire courant et en dessous :

```
perl -i -pe 's/<(\S+)/<\L$1/g' **/*.xml
```

Classes de caractères

- `[A-F0-9]` : classe formée de 2 intervalles de lettres
- `[0-9]` s'écrit aussi `\d` (digits)
- `\s` pour les espaces (tabulations, espaces insécables, etc)
- `.` : n'importe quel caractère
- `\.` : un point

Négation

- `[^0-9]` : tout sauf un chiffre
- `\S` : tout sauf un espace (idem avec `\D`)

Tricher aux mots croisés

- Dans une liste de mots, trouver ceux avec un “a”, un “k”, et une lettre entre les deux
- Trouver ceux avec un “h” à l’intérieur mais pas de la forme “ch” ni “ph”

Tricher aux mots croisés

- Dans une liste de mots, trouver ceux avec un “a”, un “k”, et une lettre entre les deux
- Trouver ceux avec un “h” à l’intérieur mais pas de la forme “ch” ni “ph”

```
/a.k/
/[^cp]h/
```

Tricher aux mots croisés

- Dans une liste de mots, trouver ceux avec un “a”, un “k”, et une lettre entre les deux
- Trouver ceux avec un “h” à l’intérieur mais pas de la forme “ch” ni “ph”

```
/a.k/
/[^cp]h/
```

- Comment trouver les mots commençant par “ph” ?

Ancres et bordures

- `^` marque le début de la ligne (ou du texte)
- `$` marque la fin de la ligne (ou du texte)
- `\b` marque le bord d'un mot (frontière avec un espace, une fin de ligne, une virgule, etc)
- Comment trouver les mots commençant par "ph" ? `/^ph/`

Ancres et bordures en pratique

- Comment trouver les mots de 6 lettres commençant par “ch” et finissant par “t” ?
- Comment trouver les mots composés avec “loup” ?

Ancres et bordures en pratique

- Comment trouver les mots de 6 lettres commençant par “ch” et finissant par “t” ?
- Comment trouver les mots composés avec “loup” ?

```
/^ch...t$/
/\bloup\b/
```

Répétitions

- `a+` : au moins une lettre “a”
- `a*` : 0, 1 ou plusieurs “a”
- `a?` : 0 ou 1 “a”
- `/^\s*$/` : une ligne vide (sauf les espaces)
- Comment convertir un fichier avec des colonnes d'espaces en fichier CSV avec des points-virgules comme séparateurs ?

Répétitions

- `a+` : au moins une lettre “a”
- `a*` : 0, 1 ou plusieurs “a”
- `a?` : 0 ou 1 “a”
- `/^\s*$/` : une ligne vide (sauf les espaces)
- Comment convertir un fichier avec des colonnes d'espaces en fichier CSV avec des points-virgules comme séparateurs ?

```
s/\s\s+/\t/g
```

Alternatives

- Comment trouver les passages traitant du nucléaire, y compris les termes annexes comme atome et fission ?

Alternatives

- Comment trouver les passages traitant du nucléaire, y compris les termes annexes comme atome et fission ?
- | sépare les alternatives

```
/nucléaire|atome|fission/
```

Groupes

- Comment trouver les passages traitant de publicité (ou de pub, des pubs) ? (mais pas de public ni de république)

Groupes

- Comment trouver les passages traitant de publicité (ou de pub, des pubs) ? (mais pas de public ni de république)

- `/\bpub\b|\bpubs\b|\bpublicité\b|\bpublicités\b|/`

Groupes

- Comment trouver les passages traitant de publicité (ou de pub, des pubs) ? (mais pas de public ni de république)

- `/\bpub\b|\bpubs\b|\bpublicité\b|\bpublicités\b|/`
- `/\bpubs?\b|\bpublicités?\b|/`

Groupes

- Comment trouver les passages traitant de publicité (ou de pub, des pubs) ? (mais pas de public ni de république)

- `/\bpub\b|\bpubs\b|\bpublicité\b|\bpublicités\b|/`
- `/\bpubs?\b|\bpublicités?\b|/`

Groupes

- Comment trouver les passages traitant de publicité (ou de pub, des pubs) ? (mais pas de public ni de république)

- `/\bpub\b|\bpubs\b|\bpublicité\b|\bpublicités\b|/`
- `/\bpubs?\b|\bpublicités?\b|/`

- Utiliser `(` et `)` pour former un groupe

```
/\b(pub|publicité)s?\b/
/\bpub(licité)?s?\b/
```

Captures et remplacement

```
Donuts for 10 $ and muffins for 5 £.
```

- Comment inverser chiffres et monnaies ?

Captures et remplacement

```
Donuts for 10 $ and muffins for 5 £.
```

- Comment inverser chiffres et monnaies ?

- Chaque groupe peut être réutilisé lors du remplacement
- \1 ou $1 est remplacé par le premier groupe
- \2 ou $2est remplacé par le deuxième groupe, etc.

Captures et remplacement

```
Donuts for 10 $ and muffins for 5 £.
```

- Comment inverser chiffres et monnaies ?

- Chaque groupe peut être réutilisé lors du remplacement
- \1 ou $1 est remplacé par le premier groupe
- \2 ou $2est remplacé par le deuxième groupe, etc.

```
(\d+) ([$£])
$2 $1
s/(\d+) ([$£])/$2 $1/g
```

Remplacements (exemples 1)

- Comment insérer (si nécessaire) l'espace requise par la ponctuation française ?

Remplacements (exemples 1)

- Comment insérer (si nécessaire) l'espace requise par la ponctuation française ?

```
(\S)([;:!?])
$2 $1
s/(\S)([;:!?])/$2 $1/g
```

Remplacements (exemples 1)

- Comment insérer (si nécessaire) l'espace requise par la ponctuation française ?

```
(\S)([;:!?])
$2 $1
s/(\S)([;:!?])/$2 $1/g
```

- Comment remplacer les adresses “bibi@labas.org” en “bibi CHEZ labas POINT org” ?

Remplacements (exemples 1)

- Comment insérer (si nécessaire) l'espace requise par la ponctuation française ?

```
(\S)([;:!?])
$2 $1
s/(\S)([;:!?])/$2 $1/g
```

- Comment remplacer les adresses “bibi@labas.org” en “bibi CHEZ labas POINT org” ?

```
s/(\S)@([a-z0-9]+)\.([a-z]+)\b/$1 CHEZ $2 POINT $3/g
```

Remplacements (exemples 2)

- Comment remplacer les virgules décimales par des points ?

Remplacements (exemples 2)

- Comment remplacer les virgules décimales par des points ?

```
s/(\d),(\d)/$1.$2/g
```

- Comment passer sur deux chiffres 01, 02 ... 20 une énumération 1, 2 ... 20 ?

Remplacements (exemples 2)

- Comment remplacer les virgules décimales par des points ?

```
s/(\d),(\d)/$1.$2/g
```

- Comment passer sur deux chiffres 01, 02 ... 20 une énumération 1, 2 ... 20 ?

```
s/\b(\d)\b/0$1/g
```

Pour aller plus loin

- Regex Golf `http://regex.alf.nu/`
- `http://perldoc.perl.org/perlretut.html` Perl regular expressions tutorial
- `http://perldoc.perl.org/perlreref.html` Aide-mémoire des expressions régulières