Université de Lille 1 SCIENCES ET TECHNOLOGIES

TDM 2 - Automates et Langages

les expressions régulières avec egrep

septembre 2015

1 Introduction

`egrep` est un outil UNIX permettant une recherche de motif (expression régulière) dans un ou plusieurs fichiers. Dans son fonctionnement «normal» :

```
egrep motif fichiers
```

egrep affiche chaque ligne dans laquelle le motif a pu être trouvé. Attention il n'est pas nécessaire que la ligne complète corresponde au motif, il suffit que l'une de ses sous-chaînes corresponde. En termes formels cela veut dire que egrep affiche toutes les lignes dont un facteur appartient au langage dénoté par l'expression régulière.
Par exemple : `egrep [A-Z][0-9] *.txt` affichera toutes les lignes contenant une majuscule suivie d'un chiffre décimal (quoi qu'il se trouve avant ou après). Voici quelques options possibles pour egrep

Modifient la forme du résultat	
-c	affiche uniquement le nombre de lignes trouvées
-h	n'affiche pas le nom du fichier en début de ligne
-n	préfixe chaque ligne par son numéro
-o	n'affiche que la partie de la ligne qui correspond au motif
Modifient les critères de recherche	
-i	ignore la casse majuscules/minuscules
-x	ne cherche que les lignes qui correspondent en totalité (et pas les facteurs propres)
Modifient la syntaxe ou la source des motifs	
-P	utilise la syntaxe Perl pour les motifs
-F nom de fichier	lit le motif dans le fichier et non sur la ligne de commande

2 Syntaxe des expressions

Il existe des classes de caractères prédéfinies :

Nom symbolique	Classe correspondante
`[:alnum:]`	les caractères alpha-numériques
`[:alpha:]`	les caractères alphabétiques
`[:cntrl:]`	les caractères de contrôle
`[:digit:]`	les caractères chiffres
`[:lower:]`	les lettres minuscules
`[:punct:]`	les caractères de ponctuation
`[:space:]`	les caractères espace
`[:upper:]`	les lettres majuscules

Par exemple, pour désigner une lettre on pourra écrire `[[:alpha:]]` (attention les lettres accentuées ne sont pas prises en compte dans cette classe)
Pour désigner une lettre ou un @, ou pourra écrire `[[:alpha:]@]`
Pour egrep, `\s` ne désigne pas un espace (utiliser `[[:space:]]`)
Attention : quand le motif est entré sur la ligne de commande (donc dans la plupart des cas) les caractères spéciaux Unix doivent être échappés. Il est conseillé de mettre l'expression régulière entre deux signes `'`

3 Exercices

Exercice 1 :
Cyrano
Vous utiliserez le fichier Cyrano.txt
Q 1 . Affichez toutes les lignes du fichier contenant le mot «nez». En utilisant l'option `--color=auto` vous pourrez visualiser tous les facteurs du texte correspondant au motif cherché.
Q 2 . Affichez toutes les lignes du fichier contenant un mot ou un portion de phrase entre parenthèses.
Q 3 . On considèrera que les mots sont composés uniquement de lettres (malheureusement la classe `[:alpha:]` ne prend pas en compte les caractères accentués qu'il faudra donc ajouter explicitement : àâéàêîùûÀÂÉÈÊÎÔÙÛ). Un mot (au sens littéraire du terme) est une suite de ces lettres délimitée avant ou après par un autre caractère ou une extrémité de ligne.
Affichez toutes les lignes comportant un mot de longueur 4 exactement. Là aussi, vous pouvez utiliser l'option `--color` pour visualiser les mots trouvés.
Vérifiez que vous prenez bien en compte les mots de 4 lettres figurant en début ou en fin de lignes. Si ce n'est pas le cas, adaptez votre expression rationnelle.
Q 4 . Dans le résultat de la commande précédente, observez la ligne commençant par «Que paternellement vous vous ...». Seul le premier des 2 «vous» est affiché en couleur. Pourquoi ?
Q 5 . Toutes les exemples de style (agressif, amical, descriptif,...) de la célèbre «tirade du nez» commencent par le même motif. Observez le texte pour trouver ce motif et déduisez-en une commande grep qui affiche tous les vers de cette tirade commençant par un nom de style.
Puis, en utilisant l'option `-o` n'affichez que la première partie chacun de ces vers.

Exercice 2 :
Vous utiliserez les fichiers du sous-répertoire html.
Q 1 . En reprenant et adaptant ce que vous avez fait pour le premier TP, écrivez un script shell qui définit une variable `valeurAttribut` contenant le motif des valeurs d'attributs XML.
Ajoutez à votre script une commande egrep affichant (avec colorisation) toutes les lignes qui contiennent ce motif dans les fichiers du répertoire html
Q 2 . Définissez de même des variables `nomXML` et `refEntite` Vous pourrez ensuite définir une variable `baliseOuvrante` en utilisant les variables précédentes.
Testez cette expression en ajoutant une commande egrep affichant avec colorisation toutes les balises ouvrantes tenant sur une seule ligne dans les fichiers du dossier html.
Q 3 . Essayez d'extraire tous les numéros de téléphone apparaissant dans les documents du répertoire.

Exercice 3 :
Le fichier `bano-59009.csv` est un fichier CSV («comma separated values»). Un fichier CSV représente une table. Chaque ligne du fichier est une ligne de la table. Les données des différentes cellules (ou colonnes) d'une même ligne sont séparées par une virgule. Le fichier fourni représente une table à 8 colonnes, chaque ligne comporte exactement 7 virgules (il n'y a pas de virgule après la dernière cellule). Il contient une base d'adresses géolocalisées de Villeneuve d'Ascq, à raison d'une adresse par ligne.
Q 1 . La deuxième colonne contient le numéro (au sein de la voie). En utilisant egrep, affichez les adresses ayant un numéro BIS ou TER.
Q 2 . La troisième colonne contient le nom de la voie, affichez les adresses dont la voie est une «Ruelle»
Q 3 . Sélectionnez les adresses dont le nom de voie est écrit exclusivement en majuscules (attention il peut cependant y avoir des espaces, des chiffres ou de la ponctuation).
Q 4 . La première colonne est l'identifiant de la voie. Son format est normalisé :

- un numéro de commune à 5 chiffres lui même composé d'un code département suivi de 3 chiffres (pour la métropole) ou 2 chiffres (pour les DOM)
- vient en suite le code de la voie qui commence par une majuscule ou un chiffre suivi de 3 chiffres.
- un code de contrôle (lettre majuscule)
- après un tiret, vient le numéro au sein de la voie, format normalisé : au moins un chiffre éventuellement suivi d'une seule lettre (majuscule).

Vérifiez la validité des données en affichant les lignes dont la première colonne ne correspondrait pas à ce format (consultez le manuel de egrep pour trouver l'option adaptée)