SVM Mac 158
Rubrique Pratique
Terminal

Nous continuons notre tour d'horizon des outils de manipulation de texte avec cette fois un éditeur de texte en mode non-interactif. Il est appelé éditeur de flux ou encore stream editor en anglais, ce qui lui vaut son nom abrégé : sed. (La plupart des exemples porteront sur des documents HTML mais tout fichier texte, TXT, RTF, CVS, …, sources de n'importe quel langage de programmation seront de la même façon manipulables.)
Yannick Cadin

Personnaliser un roman

```
$ sed -e 's/Blanche-Neige/Carole/g ; s/Grincheux/Yannick/g ; s/Simplet/Erwan/g' blanche-neige.txt > blanche-neige_revu.txt
```

La commande sed lit chaque ligne du fichier blanche-neige.txt et la reproduit sur la sortie standard (redirigée ici à l'aide du symbole > pour produire le fichier blanche-neige_revu.txt) en leur appliquant les actions soumises avec l'option -e. Ici l'on demande la substitution (action s/ancien texte/nouveau texte/) des occurrences 'Blanche-Neige' par 'Carole'. Le qualificatif g précise qu'il faut substituer chaque occurrence trouvée et pas uniquement la première de chaque ligne. À l'aide du séparateur point-virgule, on spécifie deux substitutions globales complémentaires. La distinction minuscules/majuscules est scrupuleusement respectée.

Réduire la taille d'un fichier HTML

```
$ sed -e '/^$/d ; s/^[[:blank:]]\{1,\}// ; s/[[:blank:]]\{2,\}/ /g' fichier.htm > fichier_reduit.htm
```

Plus compliqué. ^ et $ symbolisent respectivement les début et fin de ligne. L'action d signifie delete (à traduire en "effacer" ou "supprimer" en français). Les lignes vides seront donc éliminées du résultat final enregistré dans fichier_reduit.htm. La première substitution indique qu'il faut remplacer toute séquence d'au moins 1 "blanc" (représenté par un espace ou une tabulation) apparaissant en début de ligne par une chaîne vide. La seconde substitution indique quant à elle de remplacer globalement toute séquence d'au moins 2 blancs par un seul (avec l'indicateur de quantité {minimum, maximum}).
Au final, nous devrions obtenir un fichier plus petit sans aucune différence d'interprétation de la part des navigateurs.

Insertion du contenu d'un fichier après une ligne spécifique

```
$ sed -e '/<HEAD>/r balises-meta.txt' fichier.htm > fichier_avec_meta.htm
```

La plupart des actions de sed peuvent être précédées par une portée, c'est à dire une ou deux adresses qui définissent la portion de document pour laquelle l'action est à appliquer. Ces adresses peuvent être des combinaisons de numéros de lignes ET d'expressions correspondant à des morceaux de lignes traitées. Dans le cas présent, on demande de lire (action r pour "read") le fichier balises-meta.txt juste après la ligne contenant la chaîne de caractères <HEAD>.

Remplacement de balises

```
$ sed -e 's/<b>\([[:alnum:]]*\)<\/b>/<i>\1<\/i>/g' fichier.htm > fichier_italique.htm
```

Toujours plus complexe. sed interprète un certain nombre de symboles (qui composent ce que l'on nomme communément les expressions régulières) comme * qui signifie n'importe quelle quantité de l'entité qui le précède. Cela donne ici une chaîne de caractères alphanumériques ([[:alnum:]]) de n'importe quelle longueur (y compris la longueur nulle). L'emploi de parenthèses permet lors d'une substitution de désigner des portions de texte qui se retrouveront dans la partie résultat, à droite, à des emplacements précisés par des références : \1 pour le premier groupe de parenthèses, \2 pour le deuxième, etc. Jusqu'à 9 au maximum. Dans la commande ci-dessus, tout mot apparaissant entre et sera reproduit à l'identique mais délimité par et . L'usage d'antislash (\) sert le plus souvent à "protéger" le caractère qui le suit immédiatement d'une interprétation inopportune.

Extraction d'un bloc de lignes

```
$ sed -n -e '/<BODY>/,/<\/BODY>/p' fichier.htm > corps_seulement.htm
```

Notez la présence de l'option -n qui précise à sed de NE PAS afficher les lignes traitées sur la sortie standard comme il le fait par défaut puisque c'est en l'espèce l'action p (pour "print" justement, "imprimer" en français) qui s'en chargera. Le couple /<BODY>/,/<\/BODY>/ séparé par une virgule définit deux adresse, celle débutant à ligne contenant la chaîne <BODY> et celle finissant à la ligne contenant </BODY>.

Mise en forme de prix

```
$ sed -e '/Tarifs/,$s/\([[:digit:]]\)\([[:digit:]]\{3\}\)\([^[:digit:]]\)/\1\ \2\3/g' fichier.htm > fichier_formate.htm
```

Dans la définition d'une portée, $ sert à désigner la dernière ligne du fichier. Ici nous substituons à toute séquence de 1 chiffre ([:digit:]) suivi immédiatement de précisément 3 ({3}) autres que suit tout caractère différent d'un chiffre ([^[:digit:]]) l'expression composée de ce même chiffre (référence au premier groupe de parenthèses avec \1) suivi d'un espace insécable dans la notation HTML ( ) suivi du groupe de 3 chiffres (référencés avec \2) suivi enfin du caractère qui n'est pas un chiffre (associé à la troisième référence \3).
Méfiance, si une année ou toute autre valeur sur 4 chiffres se promène dans le bloc de lignes traité par la commande ci-dessus, elle sera transformée aussi.