TP programmation shell et awk

Exercice 1- Reprendre l'exercice sur l'arobrescence (exercice 5 du TD shell). Vous pourrez vous aider de l'arborescence contenue dans l'archive `TPshellAwk.tar`.

Exercice 2- (Examen Janvier 2004)

L'administrateur d'un système UNIX souhaite gagner de la place sur le disque en cherchant tous les fichiers réguliers qui sont présents en plusieurs exemplaires à différent endroits de l'arborescence et pour chacun d'entre eux, en conservant un exemplaire et en remplaçant chaque doublon par un lien symbolique vers le fichier conservé.

On suppose que l'on dispose d'un fichier de nom *liste* qui contient la liste de tous les fichiers de type régulier (un nom par ligne, correspondant au chemin absolu du fichier). Un tel fichier peut être généré par la commande `find / -type f -print 1>liste`.

i) Écrire un script shell `doublons` qui prend en argument un nom (absolu) de fichier et affiche sur la sortie standard la liste des noms de fichier contenus dans le fichier *liste* ayant le même nom mais ne se trouvant pas dans le même répertoire.

Note: on pourra utiliser les commandes `basename` et `dirname` qui prennent en argument le chemin absolu d'un fichier et rendent respectivement le nom "court" du fichier et le répertoire contenant ce fichier. Ex:

```
$ nom=‘basename /usr/include/sys/socket.h‘
$ rep=‘dirname /usr/include/sys/socket.h‘
$ echo $nom est dans $rep
socket.h est dans /usr/include/sys
```

ii) Écrire un script shell `supprime-doublons` qui appelle le script `doublons` sur tout le fichier *liste*, qui récupère la liste des possibles doublons pour chaque fichier courant f, qui fait une comparaison entre ce fichier et ses possibles doublons, et qui met un lien symbolique (`ln -s`) à la place des doublons si ceux-ci sont identiques au fichier f.

Note: une comparaison entre deux fichiers se fait à l'aide de la commande `cmp` (compare). Ex:

```
if ‘cmp -s $fichier1 $fichier2‘
then echo "les deux fichiers sont identiques"
else echo "les deux fichiers sont differents"
fi
```

iii) Combien de parcours du fichier *liste* va entraîner l'exécution de la commande `supprime-doublons`? Décrire une méthode qui permette de diminuer ce nombre de parcours (on ne demande pas des programmes, seulement une description claire et précise de la méthode).

Exercice 3- Tester tous les scripts de l'exercice 7 du TD shellAwk.

Exercice 4- *BioInformatique (M.-N. Terrasse)* Cet exercice utilise des données de PDB (Protein Data Bank, http://www.rcsb.org/pdb/home/home.do)

1. Écrire un script AWK affichant le nombre de lignes de remarques de niveau 0, 3 et 300 (le premier champ est REMARK) ainsi que le nombre total de lignes de remarques et le nombre total de lignes d'un fichier PDB. Tester avec le fichier `3HHB.ent`.

```
REMARK 0
 identifies entries in which a re-refinement has been performed ...
REMARK 3
 presents information on refinement program(s) used and related statistics.
Remark 300
 describes the biologically functional molecule (biomolecule) in free text.
```

2. Afficher le plus haut niveau de remarques dans un fichier PDB. Test avec le fichier `3HHB.ent`.

3. Afficher les lignes correspondantes à chaque premier atome d'une chaîne (le premier champ doit être ATOM et le 5ème champ doit être différent du 5ème champ de la ligne précédente . Tester avec le fichier `3HHB.ent`.

4. Afficher la liste des auteurs avec un seul nom par ligne (sans espaces ni lignes vides). Tester avec le fichier `1MSE.ent`.

```
AUTHOR    K.OGATA,S.MORIKAWA,H.NAKAMURA,A.SEKIKAWA,T.INOUE,H.KANAI,
AUTHOR   2 A.SARAI,S.ISHII,Y.NISHIMURA
```

5. À partir de l'exemple et de la description ci-dessous,

   - avec de simples compteurs, afficher le nombre de chaque famille d'atomes (carbone, nitrogène, oxygène);
   - avec un tableau associatif, afficher le nombre de chaque type d'atomes.

   Tester avec à le fichier `3HHB.ent`

```
The atom name is the third item in the record.
Notice that the first one or two characters of the atom name consists of the chemical symbol for the atom type.
All the atom names beginning with C are carbon atoms; N indicates a nitrogen and O indicates oxygen.
ATOM      9  N1   DC A   1     -45.079 -23.694  17.682  1.00  0.00           N
ATOM     10  C2   DC A   1     -46.351 -24.174  17.362  1.00  0.00           C
ATOM     11  O2   DC A   1     -46.633 -24.485  16.208  1.00  0.00           O
```