Outils système
TP 5

Vos professeurs ont une vie en dehors des cours, et généralement même un autre travail : celui de chercheur. Aujourd'hui vous allez être curieux et regarder leurs travaux de recherche.

1 Harvesting d'une page

Vous avez à votre disposition sur Arche une liste contenant les noms de vos professeurs faisant leurs recherche au LORIA et ayant accepté qu'on regarde leur site internet. Dans cette liste, choisissez votre professeur préféré.

1. Écrivez un script attendant en argument l'url du site de votre professeur préféré et créant un fichier `listePdf.txt`. Ce fichier contiendra l'adresse des documents pdf dont un lien existe sur la page parcourue.
2. Complétez ce script pour qu'il récupère aussi, dans un fichier texte *ad-hoc*, les liens vers les autres pages du professeur passé en argument.
3. Complétez ce script pour qu'il parcours toutes les pages de votre professeur hébergées par le LORIA, et pas les autres.

Indices

— *Utilisez la commande* `wget -O fichierResultat.txt url` *(c'est un O majuscule comme dans Ornithorynque).*

— Attention aux références croisées (pageA a un lien vers pageB ; pageB a un lien vers pageA). Pour éviter de vous lancer dans des boucles infinies vous pouvez :
  — Créez un fichier `pileAVisiter.txt` ayant la liste des liens des pages a visiter
  — `pileVisites.txt` ayant la liste des liens des pages visitées.

  Votre boucle prendra la première adresse de `pileAVisiter.txt`, la mettra dans `pileVisites.txt` puis récupérera toutes les adresses dans la page visitée, regardera si elles sont dans `pileVisites.txt`. Si ce n'est pas le cas, elle les ajoutera (sauf si elles existent déjà) dans `pileAVisiter.txt`.
  C'est un parcours en profondeur du graphe des hyperliens.
  Vous pouvez utiliser la commande `uniq`.

2 Harvesting complet

Écrivez un script récupérant les fichiers pdf de tous les professeurs listés dans le fichier arche. Il doit écrire sur la sortie standard un rapport donnant par professeur le nombre de pages visitées et le nombre de pdf trouvés.

Indices :

1. Vous devez commencer par, grâce à un script, modifier les noms de vos professeurs pour trouver les adresses. Attention aux noms composés.

2. Vous pouvez : soit créer un nouveau script appelant celui de l'exo1 ; soit compléter le précédent. *Faites en sorte que votre code soit lisible !*
3. Ne visitez que les pages du LORIA des professeurs de la liste. Les autres n'ont pas donné leurs accords et la légalité du harversting dans ces cas là n'est pas claire.
4. Ne soyez pas méchant avec la bande passante svp.
5. N'oubliez pas de supprimer tous les fichiers auxilliaires à la fin de votre programme.

3 Finalisation

1. Créez une commande associé à votre script qui attendra le nom d'un fichier texte contenant les noms des professeurs à glaner. Ce script informera l'utilisateur si les noms ne sont pas bien formatés puis créera le fichier des liens pdf et affichera le rapport.
2. Sous l'option `-dl`, la commande téléchargera les pdf en vrac en modifiant le nom : ils doivent commencer par le nom du professeur.