Projet - cours "Outils informatiques"

DESS TEXTE

Recueil de corpus textuels sur le web

Le projet se fait seul ou à deux. Le programme est à réaliser au moyen d'ActivePerl.

Sujet :

(1) Sélectionner un moteur de recherche à partir duquel récupérer un ensemble de documents HTML liés à un mot-clé donné. Etudier éventuellement le format d'une requête menant à un tel document, et la façon de paramétrer une telle requête.

(2) Ecrire le programme permettant de récupérer automatiquement l'ensemble des pages contenant ces documents HTML. Dans sa version finale, on pourra paramétrer le programme de façon à ce qu'il ne rapatrie qu'un nombre donné de pages, ce pour limiter les problèmes éventuels de stockage de données.
Cette partie du projet pourrait se dérouler en deux étapes : (a) extraction de l'ensemble des liens vers des pages HTML, puis (b) utilisation du fichier contenant ces liens pour récupérer les documents correspondants (attention au délai entre deux requêtes).

(3) Filtrer l'ensemble des documents obtenus pour ne garder que le texte. Veiller notamment :
- à supprimer automatiquement toute séquence répétitive (type "Retour à la page d'accueil") qui pourrait apparaître sur chaque document,
- à recoder automatiquement tous les accents et autres caractères pour lesquels `HTML : :Entities : :decode_entities` serait impuissant.

L'utilisateur final du programme ne devra avoir qu'une ligne de commande à taper, qui appelle le programme avec comme argument le mot-clé de recherche.

Pièces attendues :

- un dossier décrivant :
  - votre démarche,
  - les stratégies de vos programmes,
  - vos choix, lorsque plusieurs stratégies seront envisageables,

  - les spécifications de chacunes des fonctions que vous utiliserez (pré et post-conditions),
  - une note sur la réutilisabilité de votre programme (avez-vous dû vous conformer à une certaine url ?),
- le code source **commenté** de votre programme,
- une disquette (ou CDR, ou autre) contenant votre programme accompagné d'une notice d'utilisation, et du résultat d'une exécution précédente, i.e. un corpus textuel récupéré automatiquement.