Indexation et recherche d'information
Préparation du Corpus

LO 17

On veut réaliser une archive du site de la chaîne d'information LCI, et plus précisément des pages relatives aux informations internationales (rubrique "Monde" de la page d'accueil de LCI). Sur cette page (voir les exemples sur le site de lo17 : `http://www4.utc.fr/~lo17/LCI/`) figurent une "Une" (Le pape muet), un "Le Focus" (Bush/Putine : ambiance glaciale à Bratislava), des gros titres (Italie - Berlusconi ..., Procès - Un jury ...) et le rappel de certains titres de l'actualité internationale des pages antérieures (23 février - Rencontre, ..., 23 février - Vatican, ...). Dans cette archive on souhaite enregistrer pour chaque date de parution et pour chaque article de cette page : les titres, les résumés (s'ils existent), les imagettes (si elles existent), les liens http qui référencent les versions longues des articles et toute information permettant de repérer l'article (date, lieu, thème, ...). L'objectif est d'indexer cette archive et de stocker les index dans une base de façon à pouvoir l'interroger selon différents critères (date, thème, sujet, contenu, image, région, ville, ...). Voici quelques exemples de requêtes :

- Je veux tous les titres et les liens des articles qui parlent du Pape.
- Quelles sont les dates de parution des articles de Une parlant du procès de Michael Jackson.
- Combien d'articles de gros titres parus entre le 1 mars et le 18 juin font référence à l'enlèvement de Florence Aubenas.
- Tous les titres sur la Croatie, avec leurs liens.
- Qu'est-il arrivé au Caire en mars ?
- Qu'est-il arrivé au pape en avril ?
- ...

L'archive sera constituée uniquement de la partie informative de cette page (les "Brèves" ne seront pas incluses).

1 Travail à réaliser

L'objectif de ce TD est de préparer le corpus des pagesLCI-Monde en vue de leur indexation. Après avoir téléchargé le corpus disponible sur le site LO17 dans la rubrique Téléchargement/TDNETTOYAGE :

`http://www4.utc.fr/ lo17/TELECHARGE/TDNETTOYAGE/LCI.zip`,

vous l'installerez sur votre compte. Après l'avoir « dézipé » vous obtenez un répertoire de nom « `LCI` » .

Le résultat escompté est un seul fichier contenant pour chaque page LCI-Monde les informations relatives à la partie information sélectionnée, c'est à dire : la date, la Une, le focus, les gros titres et les actualités internationales. Ce fichier devra être structuré de façon à pouvoir retrouver :

- l'identifiant de la page (le nom et la position du fichier) ;
- la date ;
- les différentes rubriques énoncées ci-dessus, avec pour chacune d'elle les identifiants appropriés.

On a choisit de structurer ce fichier au format XML correspondant à l'arborescence donnée en annexe.

2 Tâches

2.1 Repérage et normalisation de la partie informative des pages LCI-Monde

La partie informative d'une page est contenue entre les expressions "`IBL_ID=27303`" ou "`Blc=27303`" et "`IBL_ID=27916 - Temps`" ou "`Blc=27916, [0-9]`".

1. Vérifiez que cela est vrai sur quelques fichiers

2. Ecrivez un script Perl qui, pour un fichier du répertoire initial dont le nom est entré en paramètre, génère dans un nouveau répertoire de sortie, une copie du fichier initial dont le nom est identique à celui lu en entrée mais qui n'en contient plus que la partie informative.
3. Intégrez à ce script, les lignes utiles du script de normalisation convert.pl disponible en téléchargement de façon à obtenir un fichier normalisé au format iso8859-1.
4. Ecrivez un script Perl qui répète l'opération sur l'ensemble des fichiers du répertoire d'entré.

2.2 Mise sur une seule ligne

1. De façon à faciliter les tâches de structuration qui vont suivre, écrivez un script Perl qui, pour un fichier du répertoire que vous venez de créer, supprime tous les caractères `LF ("line feed" : \n), CR ("carriage return" : \r), FF ("form feed" : \f)` et génère dans un nouveau répertoire de sortie un fichier, dont le nom est identique à celui lu en entrée.
2. Ecrivez un script Perl qui répète l'opération sur l'ensemble des fichiers du répertoire d'entré.

2.3 Création d'une ligne par rubrique

Ecrivez des expressions régulières qui caractérisent le début de chaque structure de niveau 1 (`UNE, LES_VOIRAUSSI, FOCUS, LES_GROSTITRES, LES_RAPPELS`).

1. Ecrivez un script Perl qui, pour un fichier du répertoire que vous venez de créer, le recopie dans un nouveau répertoire de sortie avec un nom identique à celui lu en entrée, de telle sorte que chaque rubrique se retrouve entièrement sur une ligne spécifique.
2. Ecrivez un script Perl qui répète l'opération sur l'ensemble des fichiers du répertoire d'entré.

2.4 Création du fichier structuré

1. Ecrivez des scripts Perl qui, pour un fichier du répertoire que vous venez de créer, et pour chacune des rubriques principales, génère un fichier structuré selon l'arbre XML donné en annexe.
2. Ecrivez un script Perl qui répète l'opération sur l'ensemble des fichiers du répertoire d'entré.

Notes :

- Il est recommandé d'écrire un script simple pour chacune des sous-structures que vous avez définies, pour en faire l'extraction. Après avoir validé votre script sur quelques pages, vous pourrez l'intégrer dans un script qui réalise l'extraction de chacune des sous-structures de la rubrique considérée.
- Il est absolument nécessaire de s'assurer à chaque étape de l'exhaustivité de l'information extraite sur toutes le pages et de contrôler les erreurs ou absences de rubriques.

3 Annexe

```
<CORPUS>
   <PAGE_LCI>
      <FICHIER>le nom du fichier</FICHIER>
      <DATE_PAGE>jj/mm/aaa</DATE_PAGE>
      <UNE>
         <urlArticle>l'url de l'article</urlArticle>
         <titreArticle>le titre de l'article</titreArticle>
         <dateArticle>jj/mm/aaa</dateArticle>
         <urlImage>l'url de l'imagette associée à l'article</urlImage>
         <resumeArticle>le texte résumé de l'article</resumeArticle>
         <mailto>l'adresse mail de l'auteur</mailto>
         <auteur>le nom et les informations sur l'auteur</auteur>
      </UNE>
      <LES_VOIRAUSSI>
         <VOIRAUSSI>
            <dateArticle>jj/mm/aaa</dateArticle>
            <urlArticle>l'url de l'article</urlArticle>
            <titreArticle>le titre de l'article</titreArticle>
         </VOIRAUSSI>
         <VOIRAUSSI>
            ...
         </VOIRAUSSI>
         ...
      </LES_VOIRAUSSI>
      <FOCUS>
         <urlArticle>l'url de l'article</urlArticle>
         <titreArticle>le titre de l'article</titreArticle>
         <dateArticle>jj/mm/aaa</dateArticle>
         <urlImage>l'url de l'imagette associée à l'article</urlImage>
         <resumeArticle>le texte résumé de l'article</resumeArticle>
         <mailto>l'adresse mail de l'auteur</mailto>
         <auteur>le nom et les informations sur l'auteur</auteur>
      </FOCUS>
      <LES_GROSTITRES>
         <GROSTITRE>
            <urlArticle>l'url de l'article</urlArticle>
            <themeArticle>le thème de l'article</themeArticle>
            <titreArticle>le titre de l'article</titreArticle>
            <dateArticle>jj/mm/aaa</dateArticle>
            <urlImage>l'url de l'imagette associée à l'article</urlImage>
            <resumeArticle>le texte résumé de l'article</resumeArticle>
            <mailto>l'adresse mail de l'auteur</mailto>
            <auteur>le nom et les informations sur l'auteur</auteur>
         </GROSTITRE>
         <GROSTITRE>
            ...
         </GROSTITRE>
         ...
      </LES_GROSTITRES>
      <LES_RAPPELS>
         <RAPPEL>
            <dateArticle>jj/mm/aaa</dateArticle>
            <themeArticle>le thème de l'article</themeArticle>
            <urlArticle>l'url de l'article</urlArticle>
            <titreArticle>le titre de l'article</titreArticle>
         </RAPPEL>
         <RAPPEL>
            ...
         </RAPPEL>
         ...
      <LES_RAPPELS>
   </PAGE_LCI>
   <PAGE_LCI>
      ...
   </PAGE_LCI>
   ...
<CORPUS>
```