Indexation et recherche d'information
Préparation du Corpus

LO 17

On veut réaliser un système d'indexation et de recherche d'information sur une archive construite à partir du site de l'ADIT. L'ADIT diffuse, sous forme de bulletins électroniques, des informations internationales de veille technologique et scientifique. Vous étudierez pour cela un échantillon de cette archive correspondant à plus de 300 articles extraits de bulletins de veille sur la France entre 2011 et 2014. Cette archive est disponible sur le site de lo17 :

`http://www4.utc.fr/~lo17/TELECHARGEMENT/BULLETINS.zip`

Vous pouvez également consulter directement ces articles :

`http://www4.utc.fr/~lo17/TELECHARGEMENT/BULLETINS/`

Ces articles contiennent des méta-informations telles que le numéro du bulletin, sa date de parution, la rubrique (focus, événement, actualité innovation, en direct des laboratoires, ...) et des contenus (texte, images, légendes d'images, contacts, ...).

Après indexation, un moteur de recherche permettra d'interroger l'archive au moyen de questions en langage naturel qui porteront sur différents critères tels que une date de parution, un contenu d'article, une rubrique, des contacts ou un mélange quelconque de ceux-ci figurant dans la même requête.

Voici quelques exemples de requêtes :

- Je veux tous les articles qui parlent d'environnement.
- Quelles sont les dates de parution des focus sur l'innovation ?
- Combien d'articles de la rubrique événement parus entre le 1 mars et le 18 juin 2013 font référence à une conférence internationale ?
- Tous les titres sur la robotique ou l'automatique.
- Liste des bulletins contenant des images de matériaux.
- Les adresses mail de contacts sur les nanotechnologies.
- ...

1 Travail à réaliser

L'objectif de ce TD est de préparer ce corpus d'articles en vue de son indexation. Pour réaliser l'indexation, vous allez construire un fichier, qui rassemble tous les articles dans une structure XML qui sera facilement indexable. Ce fichier XML devra être structuré de façon à pouvoir retrouver (quand ils existent) :

- l'identifiant de la page (le nom et la position du fichier),
- le numéro du bulletin correspondant,
- la date,
- la rubrique,
- le titre de l'article,
- le texte de l'article
- la ou les images avec leur(s) URL(s) et leur(s) légende(s) respective(s),
- les informations de contact.

Après avoir téléchargé et « dézipé » ce corpus sur votre compte vous obtiendrez le répertoire « `BULLETINS` » .

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

2 Description de tâches

Note : les fichiers de l'archive sont au format UTF8, vous devrez vous assurer que tous les traitements et tous les fichiers produits restent compatibles avec ce format.

Naviguez dans quelques bulletins et repérez les éléments de structure HTML qui vous permettront d'identifier les différentes parties du fichier XML que vous allez constituer.

Procédez par étapes successives sur un seul fichier à la fois :

1. choisissez la partie de structure que vous souhaitez traiter (numéro d'identification, date, titre , ...),
2. repérez les éléments de structure HTML qui identifient cette structure de façon unique et complète,
3. écrivez un script perl qui extrait la partie souhaitée et l'écrit dans le fichier XML avec les balises appropriées,
4. testez et validez ce script sur plusieurs fichiers,
5. écrivez un script perl qui appelle le script précédent sur tous les fichiers de l'archive et écrit le résultat dans le fichier unique,
6. **pour chaque partie traitée, vous devez trouver une méthode qui vous permet de vous assurer que tous les fichiers ont été traités et que le traitement est correct et valide**,
7. répétez l'opération pour une autre partie de la structure.

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>
   <bulletin>
      <fichier>le nom du fichier</fichier>
      <numero>le numéro du bulletin</numero>
      <date>la date du bulletin au format jj/mm/aaa</date>
      <rubrique>la rubrique de la liste des rubriques que vous aurez identifiées</rubrique>
      <titre>le titre de l'article</titre>
      <texte>le texte de l'article</texte>
                           (le texte peut encore ici être découpé en paragraphes ou non)
      <images> (il peut y avoir 0, une ou plusieurs images dans un article)
         <image>
            <urlImage>l'URL de l'image</urlImage>
            <legendeImage>la légende de l'image</legendeImage>
         </image>
         <image>
            ...
         </image>
            ...
      </images>
      <contact>toute l'information de contact</contact>
   </bulletin>
   <bulletin>
      ...
   </bulletin>
      ...
</corpus>
```