Indexation et recherche d'information
Préparation du Corpus

LO 17

Travail à réaliser

On souhaite, créer les index des pages LCI-monde, à partir du fichier XML que vous avez réalisé dans les TD précédents. Le principe est de créer une série de fichiers (dits fichiers inverses) permettant de décrire les documents à l'aide de tout ou partie des éléments contenus dans ce fichier XML. Le résultat de l'indexation sera donc un ensemble de fichiers inverses, chaque fichier correspondant à une balise ou un ensemble de balises (date page, rubriques, titre, titre+résumé, thème, source, date article, ...).

La partie la plus délicate concerne la réalisation des fichiers inverses à partir des mots des titres et des résumés. On souhaite en particulier représenter par un même mot de référence (un lemme) toutes les dérivations d'un même mot (par exemple, mise au féminin, pluriel des noms et adjectifs, déclinaisons des verbes). D'autre part on souhaite pouvoir s'affranchir des mots qui ne sont pas porteurs de sens, tels les articles, les pronoms, les adverbes, etc, et de ceux qui n'apportent pas d'information, tels les verbes auxiliaires, les mots très généraux.

1 Etude préliminaire

La réalisation des fichiers inverses à partir des mots des titres et des résumés demande donc au préalable une étude détaillée du vocabulaire, telle :

- la liste de *petits* mots (1, 2, 3 ou 4 lettres).
- la liste de mots apparaissant dans un nombre important de pages et n'ayant pas de pouvoir discriminant (c'est à vous de définir le nombre de pages approprié).
- A l'issue de cette analyse vous devrez créer un script permettant d'éliminer ces mots du corpus à partir de ces listes.

Vous avez à votre disposition une série de scripts :
NOTE : Il est indispensable que vous ayez parfaitement compris le contenu de chaque script avant de l'utiliser.

- `newsegmente.pl` : Ce script découpe le corpus (les titres et les résumés) en mots. Le format du résultat est un mot par ligne. L'option `-f` permet d'afficher en face de chaque mot sa page de provenance séparé par une tabulation, l'option `-t` permet d'afficher en face de chaque mot la rubrique

dans laquelle il est apparu, séparé par une tabulation et l'option `-a` permet d'afficher l'URL de l'article dans lequel il est apparu. Il est nécessaire que vous modifiiez ce script en fonction des noms que vous avez attribué aux différentes balises.
- `newcreeFiltre.pl` : Ce script permet de créer des filtres *i.e.* des scripts permettant d'éliminer ou de remplacer des mots. Il prend en entrée une liste de mots (qui peut être sur deux colonnes) et crée un script perl.

2 Création des lemmes

Une fois que vous aurez filtré votre fichier XML initial de façon appropriée, vous pourrez construire, à partir du fichier filtré, une liste à deux colonnes contenant, en première colonne, un mot de titre ou de résumé et, en seconde colonne, son lemme. Vous disposez pour cela des scripts suivants ;
- `tronc.pl` : Ce script permet de générer la liste des successeurs pour chaque lettre des mots d'une liste de mots (algorithme vu en cours).
- `filtronc.pl` : À partir des résultat de `tronc.pl` ce script crée (avec l'option `-v`) un fichier à deux colonnes associant un mot à un lemme.

3 Création d'un corpus pour construire les tableaux inverses

Vous allez maintenant créer deux filtres : le filtre des mots qui ne sont pas significatifs et celui qui associe les autres mots à leur lemme.

Une fois ces filtres réalisés, filtrez votre fichier XML pour constituer le fichier XML qui servira à construire les tableaux inverses.

4 Création des fichiers inverses

Vous allez pouvoir maintenant réaliser des fichiers inverses contenant en première colonne un identifiant (un mot, une date, un e-mail, ...) et dans les colonnes suivantes le nom du fichier html dans lequel il apparaît et, par exemple, la rubrique (une, gros titre, focus, ...), le thème, etc.Vous disposez pour cela des scripts suivants ;
- `index.pl` : Ce script permet de créer, à partir du corpus, un fichier inverse sur une balise donnée en argument.
- `newindexMot.pl` : Ce script permet de créer un fichier inverse à partir d'un flux de données de la forme « mot page rubrique urlArticle »

NOTE : Ces scripts doivent être éventuellement édités si vous travaillez sur un fichier corpus XML différent de celui proposé au téléchargement.