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 ou les mots très généraux.

1 Création d'une stop-list

On va donc commencer par construire la liste des mots qui ne vont pas figurer dans l'index et que l'on doit supprimer du fichier XML avant de faire la lemmatisation du corpus. La construction de cette stop-list va s'appuyer sur le calcul du coefficient $tf \times idf$ qui a été étudié en cours.

1.1 Choix de l'unité documentaire

Le calcul de ce coefficient s'appuie sur la fréquence d'un mot dans un document et sur la fréquence de documents pour un mot donné, l'unité documentaire doit donc être clairement définie. Dans cette application, on a le choix suivant

- **un document = une page**
  Dans ce cas on s'intéressera à la fréquence des mots dans une page, même s'il apparaît dans différents titres ou résumés.

– **un document= un article**
  Dans ce cas il faut calculer au sein de chaque page, la fréquence des mots dans chacun des articles (titre et résumé) où il apparaît.

Vous devrez réfléchir aux conséquences des choix ci-dessus en termes de difficulté du calcul du coefficient selon l'unité documentaire choisie et des résultats obtenus pour différents types de requêtes.
Le résultat de votre réflexion sera argumenté dans le rapport intermédiaire que vous rendrez à la fin de la partie sur l'indexation (le 25 octobre).

1.2 Détermination de la stoplist

On prend l'option **un document = une page**. Vous devez calculer le coefficient $tf \times idf$ pour chaque mot du corpus et fixer un seuil au delà duquel les mots seront affectés à une stop list.

Pour cela on vous recommande de commencer par construire le fichier des coefficients $tf_{i,j}$ de chaque mot i dans chaque page j. Vous construirez donc un fichier qui contient trois colonnes : une colonne $nom\_de\_la\_page$ (c'est à dire le nom du fichier html), une colonne mot_i et une colonne $tf_{i,j}$.

Ensuite vous construirez le fichier des coefficient $idf_i = \log_{10} \frac{N}{\mathrm{df}_i}$ où df_i est la fréquence de documents dans laquelle le mot i apparaît. Ce sera un fichier à deux colonnes mot_i, idf_i

Finalement vous construirez un fichier à trois colonnes : une colonne $nom\_de\_la\_page$ (c'est à dire le nom du fichier html), une colonne mot_i et une colonne $tf \times idf_{i,j}$.

A l'issue de cette analyse vous devrez déterminer une règle d'extraction des mots non significatifs qui seront stockés dans une stop-list.

Vous pouvez alors générer le script permettant d'éliminer ces mots du corpus à partir de cette liste. Filtrez le fichier XML initial et sauvegardez le résultat dans un fichier XML différent.
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.
  ATTENTION : Il sera peut-être nécessaire que vous modifiiez ce script en fonction des noms que vous avez attribués 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 à en supprimer les mots de la stoplist ci-dessus, 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 ;

- `successeurs_2013.pl` : Ce script permet de générer la liste des successeurs pour chaque lettre des mots d'une liste de mots (optimisation de l'algorithme vu en cours).
- `filtronc.pl` : À partir des résultat obtenus avec `successeurs_2013.pl` ce script crée (avec l'option `-v`) un fichier à deux colonnes associant un mot à un lemme.

Créez ensuite le filtre qui associe les mots à leur lemme et filtrez le fichier XML que vous avez créé dans l'étape précédente. Sauvegardez le résultat dans un nouveau fichier XML qui servira à construire les tableaux inverses.

3 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.