Indexation et recherche d'information
Indexation du Corpus

LO 17

Printemps 2015

Travail à réaliser

On souhaite, créer les index des articles des bulletins électroniques de l'ADIT, 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 ou numéro du bulletin, rubrique, titre, titre+résumé, légende des images, ...).

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, 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 le nombre de documents qui contiennent un mot donné. L'unité documentaire doit donc être clairement définie. Dans cette application, on a le choix suivant :

- **un document = un bulletin**
  Dans ce cas on s'intéressera à la fréquence des mots qui apparaissent dans un bulletin, même s'ils figurent dans les titres ou résumés de différents articles.
- **un document= un article**
  Dans ce cas il faut calculer la fréquence des mots qui apparaissent dans le titre ou le résumé de chaque article.

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.

1.2 Détermination de la stoplist

Pour ce TD, on prend l'option **un document = un article**. 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.

1. Pour cela on vous recommande de commencer par construire le fichier des coefficients $tf_{i,j}$ de chaque mot i dans chaque article j. Vous construirez donc un fichier qui contient trois colonnes : une colonne $nom\_du\_fichier\_article$ (c.à.d. le nom du fichier html), une colonne mot_i et une colonne $tf_{i,j}$.

2. Ensuite vous construirez le fichier des coefficient $idf_i = \log_{10} \frac{N}{\mathrm{df}_i}$ où df_i est le nombre de documents dans laquelle le mot i apparaît. Ce sera un fichier à deux colonnes mot_i, idf_i
3. Finalement vous construirez un fichier à trois colonnes : une colonne $nom\_du\_fichier\_article$ (c.à.d. 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 les deux scripts suivants :
NOTE : Il est indispensable que vous ayez parfaitement compris le contenu de chaque script avant de l'utiliser.

- `segmente.pl` : Ce script découpe le corpus (les titres et les résumés) en mots. Le format du résultat est un mot par ligne. Ce script peut s'exécuter avec plusieurs options :
  - l'option `-f` permet d'afficher en face de chaque mot son fichier de provenance séparé par une tabulation,
  - l'option `-r` permet d'afficher en face de chaque mot la rubrique dans laquelle il est apparu, séparé par une tabulation,
  - l'option `-n` permet d'afficher le numéro de bulletin dans lequel l'article est apparu.

  Usage : segmente.pl [-f] [-r] [-n]

  ATTENTION : Il sera peut-être nécessaire que vous modifiiez ce script si vous utilisez un fichier XML avec des noms de balises différents.

- `newcreeFiltre.pl` : Ce script permet de créer des filtres *i.e.* des scripts permettant d'éliminer un mot ou de remplacer un mot par un autre mot. 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.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.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, une rubrique, ...) et dans les colonnes suivantes le nom du fichier html dans lequel il apparaît et, par exemple, la rubrique, 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.
- `indexText.pl` : Ce script permet de créer un fichier inverse à partir d'un flux de données de la forme « mot rubrique fichier numéro »

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