LO17 : Indexation et Recherche d'information
TD4 : Indexation du Corpus

Printemps 2025

Travail à réaliser

On souhaite créer les index des articles des bulletins électroniques de l'ADIT, à partir du fichier XML filtré que vous avez réalisé dans le TD précédent. 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+texte de l'article, texte de l'article, url des images, légende des images, ...).

La partie la plus délicate concerne la réalisation des fichiers inverses à partir des mots des titres et des textes des articles. Nous allons utiliser des outils modernes de traitement du langage pour effectuer la lemmatisation (spaCy ou Snowball).

1 Création des lemmes

A partir du fichier XML filtré ne contenant pas les mots de l'anti-dictionnaire, vous devrez construire une liste à deux colonnes contenant, en première colonne, un mot de titre ou de texte et, en seconde colonne, son lemme. À cette étape, vous devrez utiliser :

- `spaCy` : Une bibliothèque de traitement du langage naturel en Python, avec un modèle de lemmatisation français (`fr_core_news_sm`).
- `Snowball` : Contenu dans NTLK, une autre bibliothèque de traitement du langage naturel en Python, utilisé pour une lemmatisation basée sur des règles de racinisation, et adapté depuis le Stemmer de Porter à la langue française.

1.1 Avec SpaCy

1. Installez la librairie SpaCy et le modèle français (`fr_core_news_sm`).
2. Écrivez un script Python qui :
   - Pour chaque mot du corpus, récupérez son lemme.
   - Créez un fichier à deux colonnes : mot → lemme.
3. Analysez les résultats : quels types de variations sont bien gérés ? Quelles sont les limites ?

1.2 Avec Snowball (NLTK)

1. Installez NLTK et le stemmer Snowball pour le français.
2. Écrivez un script Python qui :
   - Appliquez le stemmer Snowball à chaque mot.
   - Créez un fichier à deux colonnes : mot → racine.
3. Analysez les résultats : quels types de variations sont bien gérés ? Quelles sont les limites ?

1.3 Analyse comparative et choix

- Pour un échantillon de mots, comparez les résultats des deux méthodes SpaCy et Snowball.
- Calculez des statistiques (nombre de lemmes uniques, distribution, etc.).
- Quel système semble le plus adapté pour ce corpus ? Justifiez.

Filtrez le fichier XML que vous avez créé dans l'étape précédente selon votre choix (SpaCy et Snowball) et en vous aidant de `substitue.py`. Sauvegardez le résultat dans un nouveau fichier XML qui servira à construire les fichiers inverses.

2 Création des fichiers inverses

Vous allez pouvoir maintenant réaliser des fichiers inverses contenant en première colonne un terme (un lemme, une date, une rubrique, ...) et dans la colonne suivante les identifiants des documents dans lesquels il apparaît (bulletin ou article) ainsi que la fréquence d'apparition.

— A partir du corpus XML lemmatisé, créez un fichier inverse pour chaque balise importante, comme le titre, le résumé, la rubrique, la date, etc.

— **Bonus :** Comment pourriez-vous améliorer la qualité de l'indexation de ce corpus ?