LO17 : Indexation et Recherche d'information
TD3 : Anti-dictionnaire

Printemps 2025

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 le TD précédent. 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. 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 cet anti-dictionnaire va s'appuyer sur le calcul du coefficient tf-idf qui a été étudié en cours.

1 Choix de l'unité documentaire

Le calcul de ce coefficient s'appuie sur la fréquence d'un mot (le nombre de *tokens* identiques) dans un document et sur le nombre de documents qui contiennent ce mot. 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 textes de différents articles de différentes rubriques.

— **un document = un article :**
Dans ce cas, il faut calculer la fréquence des mots qui apparaissent dans le titre ou le texte de chaque article.

Vous devrez réfléchir aux conséquences des choix ci-dessus sur le mode de calcul du coefficient tf-idf selon l'unité documentaire choisie, sa signification et les résultats obtenus pour différents types de requêtes. **Le résultat de votre réflexion sera argumenté dans le rapport à rendre**.

1. Vous écrirez un script `segmente.py` qui découpe le corpus (les titres et les textes) en tokens. Le format du résultat est un mot par ligne séparé par une tabulation de son document d'origine (nom du fichier ou numéro du bulletin selon votre choix).
2. Vous écrirez un script `substitue.py` qui permet d'éliminer ou de remplacer un mot par autre dans un texte. Il prend en entrée le texte ainsi qu'un fichier de deux colonnes de mots et de leur substitution (`""` si le mot est à éliminer) séparées par une tabulation.

2 Détermination de l'anti-dictionnaire (stop words)

Vous devez calculer le coefficient tf-idf pour chaque mot du corpus et fixer un seuil au delà duquel les mots seront affectés à l'anti-dictionnaire.

1. Pour cela, nous vous recommandons de commencer par construire le fichier des coefficients $tf_{t,d}$ de chaque mot t dans chaque article d. Vous construirez donc un fichier qui contient trois colonnes : une colonne $identifiant\_du\_document$ (c.à.d. le nom du fichier ou le numéro du bulletin), une colonne mot_t et une colonne $tf_{t,d}$.
2. Ensuite, vous construirez le fichier des coefficients $idf_t = \log_{10} \frac{N}{\mathrm{df}_t}$, où N est ne nombre total de documents et df_t est le nombre de documents dans lesquels le mot t apparaît. Ce sera un fichier à deux colonnes mot_t, idf_t.

3. Finalement, vous construirez un fichier à trois colonnes : une colonne $identifiant\_du\_document$, une colonne mot_t et une colonne $tf \times idf_{t,d}$.

A l'issue de cette analyse, vous devrez déterminer une règle d'extraction des mots non significatifs qui seront stockés dans l'anti-dictionnaire. 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 pourrez utiliser le script `segmente.py` pour commencer et `substitue.py` pour créer le fichier XML filtré.