Création manuelle de vecteurs distributionnels

L'objectif global de ce TP est de réaliser de façon « manuelle » toutes les étapes menant à la création de vecteurs distributionnels à partir d'un corpus, jusqu'à une première évaluation de la qualité des vecteurs produits.

On pourra prendre en entrée l'un des deux corpus fournis : `Voyage-Verne-utf8.txt` (environ $82k$ tokens et $16k$ types), ou `Emile-Rousseau-utf8.txt` ($236k$ tokens et $28k$ types).

1. Indexation et fréquence
   (a) Après avoir extrait le vocabulaire du corpus, créer un index qui associe à chaque type un numéro entier.
   (b) Implémenter une fonction de recherche dichotomique et trier l'index. Cette fonction permet maintenant de créer une version du corpus où chaque token est remplacé par son numéro.
   (c) Calculer la fréquence de chaque type dans le corpus.
   (d) On vérifiera que la fréquence calculée avec le corpus indexé est nettement plus rapide qu'un calcul de fréquences par la méthode habituelle (création d'un dictionnaire dont les clés sont les types et les valeurs leur fréquence).
2. Calcul des vecteurs distributionnels
   (a) Sélectionner les D mots les plus fréquents (on prendra pour commencer $D = 100$), et créer un nouvel index où ces D mots sont numérotés de 0 à $D - 1$: ce seront les dimensions de nos vecteurs.
   (b) On définit le contexte autour d'un mot (cible) comme les 5 tokens avant et les 5 tokens après ce mot. Pour un mot donné w (représenté par son index), pour chacun des D mots fréquents (d), compter le nombre de fois que w a le mot d dans son contexte. Cette valeur est la coordonnée de w pour la dimension s. On obtient donc un vecteur de D entiers.
   (c) Sélectionner les D' mots les plus fréquents, et remplir une matrice de $D \times D'$ entiers correspondant aux vecteurs distributionnels de ces D' mots ($D' \geq D$).
3. Mesure des distances
   (a) Implémenter la mesure de distance euclidienne et la mesure de distance cosinus.
   (b) Rechercher les n paires de mots les plus proches et les n paires de mots les plus éloignées.
   (c) Evaluer manuellement/intuitivement la qualité des résultats trouvés.
   (d) [Piste 1] Evaluer l'impact de prétraitements du corpus (nettoyages et normalisation, suppression de stop words...).
   (e) [Piste 2] Comparer les similarités obtenues avec d'autres mesures de similarité lexicales (en anglais, on pourrait prendre `nltk`...)