TP semaine 6 : Influence de la taille de la fenêtre sur les embeddings distributionnels

L'objectif du TP est d'observer l'influence de la taille de la fenêtre utilisée lors de l'élaboration d'embeddings statiques par comptage de co-occurrences. L'hypothèse est que les mots qui partagent des contextes étroits (de taille 3 par exemple) sont plutôt des mots qui partagent des propriétés morpho-syntaxiques, alors que des mots qui partagent des contextes plus larges vont sans doute plutôt partager des propriétés sémantiques.

Pour mener cette étude, on va passer par les étapes suivantes :

1. Choix d'un corpus de taille réduite, segmentation et tokenisation du texte.
2. Calcul d'une matrice terme-terme (demi-matrice carrée) pour une taille de fenêtre donnée.
3. Choix de 10 mots variés apparaissant dans le corpus.
4. Pour chacun des mots choisis, identification des 10 mots les plus voisins par similarité cosinus.

En réalisant cette étude pour deux valeurs distinctes de taille de fenêtre (par exemple $k = 2$ et $k = 6$), et en comparant les 10×10 mots obtenus dans les deux cas, on devrait pouvoir confirmer ou infirmer l'hypothèse initiale.

Convention : un voisinage à $k = 2$ d'un mot cible est constitué du sac de mots comprenant les 2 mots précédant la cible et les deux mots qui la suivent (*cbow = continuous bag of words*).

(1) a. Chaque homme porte la forme entière de l'humaine condition.
b. $k = 2$: cible : *forme* ; cbow = { *de*, *entière*, *la*, *porte* }
c. $k = 6$: cible : *forme* ;
cbow = { *chaque*, *condition*, *de*, *entière*, *homme*, *humaine*, *l'*, *la*, *porte*, *.* }