Notes sur la librairie Word2vec

Word2vec est un algorithme non supervisé de *Word Embedding* (plongement de mots) utilisant une structure de réseaux de neurones à 2 couches afin d'attribuer une représentation vectorielle à des mots. Cette méthode diffère du comptage direct souvent utilisé, notamment via la technique TF-IDF.

Le principal intérêt de Word2vec est la prise en considération du contexte d'un mot. En effet, après avoir entraîné un modèle de ce type sur un jeu de données conséquent, on obtiendra les vecteurs représentant les mots. Cette génération de vecteurs est le résultat de l'utilisation de perceptrons linéaires, avec une seule couche cachée.

Deux techniques existent pour saisir le contexte d'un mot :

- CBOW (« Continuous Bag of Words »), qui entraîne le modèle pour prédire le mot, compte tenu du contexte environnant (i.e les mots qui se trouvent avant ou après le mot à prédire dans une phrase).
- « Skip-gram », prédisant le contexte en fonction du mot. Skip-gram agit, en quelque sorte, dans le sens inverse de celui de CBOW.

Dans les deux cas, l'objectif est d'obtenir un dictionnaire un dictionnaire recensant les vecteurs de mots dans une dimension inférieure (compression), en utilisant une des deux techniques présentées ci-dessus.

Idée pour la généralisation à la similarité entre documents

Afin d'obtenir la similarité d'un document avec d'autres, on pourrait calculer la similarité de chaque phrase ou paragraphe du document d'entrée avec les phrases ou paragraphes des autres documents puis calculer la moyenne de ces scores de similarités.