

Série de TD 3 : Représentation du texte

Exercice 01 :

Considérez les phrases suivantes :

"L'intelligence artificielle révolutionne les technologies modernes."

"Le machine learning permet d'automatiser les tâches complexes."

"Les réseaux neuronaux sont au cœur des avancées en deep learning."

1. Construisez manuellement un vocabulaire en utilisant la méthode Bag of Words (BoW).
2. Représentez chaque phrase sous forme d'un vecteur basé sur le vocabulaire construit.
3. Identifiez les limitations de la méthode BoW pour ces phrases (par exemple, perte d'ordre ou de contexte).

Exercice 02 :

En utilisant les mêmes phrases que l'exercice précédent :

1. Calculez la **fréquence de terme (TF)** pour chaque mot.
2. Calculez l'**inverse de la fréquence des documents (IDF)** pour chaque mot.
3. Multipliez les valeurs TF et IDF pour obtenir la matrice TF-IDF.
4. Comparez les résultats obtenus avec ceux du Bag of Words.
5. Expliquez pourquoi certains mots comme "les" ou "des" obtiennent une faible importance dans TF-IDF.

Exercice 03 :

1. À partir des représentations vectorielles obtenues (BoW ou TF-IDF), calculez les similarités entre les phrases en utilisant :
 - La distance cosinus.
 - La distance euclidienne.
2. Expliquez pourquoi certaines phrases apparaissent plus similaires que d'autres selon ces mesures.