Techniques de Vectorisation Courantes en Traitement du Langage Naturel

1 Introduction

La vectorisation des textes est une étape essentielle dans le traitement du langage naturel (NLP), permettant de convertir des textes en représentations numériques exploitables par les modèles de machine learning. Ce document présente certaines des techniques de vectorisation les plus courantes.

2 Techniques de Vectorisation

2.1 Sac de Mots (Bag of Words)

Le modèle Sac de Mots (Bag of Words, BoW) est l'une des techniques les plus simples de vectorisation. Il consiste à représenter un texte par un vecteur de comptage des occurrences de chaque mot dans le texte, sans tenir compte de l'ordre des mots.

$$\mathbf{v} = [f_1, f_2, \ldots, f_n]$$

où f_i est le nombre d'occurrences du mot i dans le texte.

2.2 TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF est une technique de vectorisation qui mesure l'importance d'un mot dans un document en fonction de sa fréquence dans ce document et de sa rareté dans un corpus de documents. La formule TF-IDF pour un mot t dans un document d est donnée par :

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

où

$$\text{TF}(t, d) = \frac{\text{Nombre d'occurrences de } t \text{ dans } d}{\text{Nombre total de mots dans } d}$$

et

$$\text{IDF}(t) = \log\left(\frac{N}{\text{Nombre de documents contenant } t}\right)$$

avec N étant le nombre total de documents dans le corpus.

2.3 Word2Vec

Word2Vec est une technique de vectorisation basée sur des réseaux de neurones, qui représente chaque mot par un vecteur dense de taille fixe. Il existe deux variantes principales de Word2Vec : *Continuous Bag of Words* (CBOW) et *Skip-gram*. Les vecteurs de mots appris capturent des relations sémantiques et syntaxiques entre les mots.

2.4 GloVe (Global Vectors for Word Representation)

GloVe est une méthode de vectorisation qui apprend des représentations de mots en fonction de leur co-occurrence globale dans un corpus de textes. Le modèle GloVe factorise une matrice de co-occurrence mot-mot pour produire des vecteurs de mots qui capturent des régularités sémantiques.

2.5 FastText

FastText, développé par Facebook, est une extension de Word2Vec qui prend en compte les sous-mots (ngrams) dans la représentation des mots. Cette approche permet de mieux gérer les mots rares ou inconnus et d'apprendre des représentations plus riches.

2.6 BERT (Bidirectional Encoder Representations from Transformers)

BERT est un modèle de transformer pré-entraîné qui produit des représentations contextuelles des mots. Contrairement aux techniques de vectorisation traditionnelles, BERT considère le contexte bidirectionnel de chaque mot dans une phrase, ce qui permet de capturer des relations contextuelles complexes.

3 Conclusion

Ces techniques de vectorisation, allant des méthodes simples comme le Sac de Mots et TF-IDF aux modèles avancés comme Word2Vec, GloVe, FastText et BERT, sont essentielles pour convertir des textes en représentations numériques exploitables dans diverses tâches de traitement du langage naturel. Chaque technique a ses avantages et ses inconvénients, et le choix de la méthode dépend souvent des spécificités de la tâche et des données disponibles.