

Embedding model (représentations vectorielles)

Qu'est-ce qu'un embedding ?

Un **embedding** est une représentation vectorielle d'un mot, d'une phrase ou même d'un texte complet dans un espace mathématique de haute dimension. L'objectif est de transformer des éléments de texte en vecteurs (des suites de nombres) qui capturent leur signification.

Par exemple, prenons les mots suivants :

- "chat"
- "chien"
- "voiture"

Un modèle d'embedding va attribuer à ces mots des **vecteurs numériques**. Ces vecteurs seront placés dans un espace multidimensionnel, et la distance entre ces vecteurs peut refléter la similarité sémantique entre les mots. Par exemple :

- Les vecteurs de "chat" et "chien" devraient être proches dans l'espace vectoriel car ces mots sont sémantiquement similaires (ce sont des animaux domestiques).
- Le vecteur de "voiture" serait éloigné de ceux de "chat" et "chien", car ce sont des concepts différents.

Contexte d'un mot et fenêtre de tokens :

L'un des éléments clés d'un modèle d'**embedding** performant est sa capacité à comprendre le **contexte** dans lequel un mot apparaît. Par exemple, le mot "banc" peut signifier "un meuble" ou "une institution financière" en fonction du contexte.

Voici deux exemples avec la même **fenêtre de contexte** :

- **Phrase 1** : "Je suis allé au banc de touche."
- **Phrase 2** : "J'ai ouvert un compte au banc."

Un modèle avec une **fenêtre de contexte large** va prendre en compte non seulement le mot "banc" lui-même, mais aussi les mots qui l'entourent, comme "de touche" dans le premier cas et "compte" dans le second. Grâce à une grande fenêtre de contexte, il comprendra que "banc" se réfère à un **siège** dans le premier exemple, et à une **institution financière** dans le deuxième.

Pourquoi une grande fenêtre de contexte est importante ?

Un modèle avec une **fenêtre de contexte large** peut prendre en compte plus de mots autour du mot cible pour mieux comprendre son sens. Par exemple :

- Dans la phrase : "Le chat mange une souris."
- La fenêtre de contexte pour le mot "mange" pourrait inclure les mots "chat" et "souris", ce qui aide le modèle à comprendre que "mange" fait référence à une action de nourriture, et non à une autre signification possible (comme "manger un repas" dans un autre contexte).

Une **fenêtre de contexte large** signifie que le modèle peut analyser des sections plus longues du texte, ce qui améliore la compréhension du sens d'un mot, même dans des phrases complexes.

Exemples concrets avec "nomic-embed-text" :

Supposons que nous utilisions un modèle "**nomic-embed-text**" pour générer des embeddings pour ces deux phrases :

1. **Phrase 1** : "Le chat dort sur le canapé."
2. **Phrase 2** : "Le chat a attrapé une souris."

Le modèle va générer des embeddings pour chaque mot en prenant en compte le contexte autour d'eux.

- Pour le mot "chat", le modèle pourrait produire un vecteur similaire dans les deux phrases, car il est utilisé dans des contextes relativement similaires (un animal domestique). Cependant, le modèle prendra aussi en compte des mots comme "dort" dans la première phrase et "attrapé" dans la deuxième, ajustant l'embedding de "chat" pour capturer les différences de contexte.
- Le mot "dort" dans "Le chat dort sur le canapé" aura un contexte avec "chat" et "canapé", ce qui le placera près d'autres mots associés à des actions de repos ou de sommeil, comme "dormir" ou "repos".
- Le mot "attrapé" dans "Le chat a attrapé une souris" aura un contexte avec "chat" et "souris", plaçant ce mot plus près de mots liés à l'action de chasser ou de capturer.

Conclusion :

Le modèle "**nomic-embed-text**" est un modèle performant qui utilise une **fenêtre de contexte large** pour analyser les relations entre les mots dans un texte. Cela lui permet de produire des **embeddings** plus précis et contextuellement adaptés, ce qui est essentiel pour des tâches telles que la recherche sémantique, la traduction automatique ou l'analyse de texte, où il est important

de comprendre le sens global d'un mot en fonction de son contexte spécifique.

En résumé, une grande fenêtre de contexte permet au modèle de mieux "comprendre" le texte dans son ensemble et de produire des embeddings qui reflètent correctement le sens des mots dans chaque situation.

Revision #2

Created 6 December 2024 13:44:07 by Maxime

Updated 6 December 2024 13:46:12 by Maxime