



Processamento de Linguagem Natural com Transformers

Processamento de Linguagem Natural com Transformers

Reforçando o Conceito de Embedding

"Embedding" é um termo geralmente usado em Processamento de Linguagem Natural (PLN) que se refere à representação de palavras, frases ou até mesmo documentos inteiros como vetores de números. A ideia por trás dos embeddings é converter o texto, que é uma forma de dados não estruturada, em uma forma estruturada que pode ser manipulada matematicamente.

Existem várias maneiras de calcular embeddings, mas em geral, a ideia é que palavras com significados semelhantes devem ser representadas por vetores que estão próximos uns dos outros no espaço vetorial, enquanto palavras com significados diferentes devem ser representadas por vetores que estão distantes uns dos outros.

O uso de embeddings permite que os modelos de aprendizado de máquina lidem com texto de uma maneira muito mais significativa, levando em consideração a semântica das palavras. Isso tornou possível o desenvolvimento de modelos de PLN altamente avançados, como o BERT, que podem entender a semântica complexa do texto em um grau muito elevado.

Os embeddings são aprendidos por meio de treinamento em grandes quantidades de texto. Modelos como Word2Vec ou GloVe aprendem embeddings de palavras, enquanto modelos mais avançados como BERT ou GPT aprendem embeddings contextuais, em que a representação vetorial de uma palavra pode variar dependendo do contexto em que a palavra é usada.