TREINANDO EMBEDDINGS

Tópicos especiais em gestão de TI.

PROF. DIEGO ROSSI

EMBEDDINGS

- Embeddings são representações numéricas de objetos, conceitos ou entidades em um espaço vetorial de dimensões fixas.
- Eles são usados para representar informações complexas em uma forma que os algoritmos de aprendizado de máquina podem entender e processar de maneira eficaz.
- Os embeddings são amplamente utilizados em várias áreas, como PLN, visão computacional, sistemas de recomendação e muito mais.
- Para o PLN, a técnica de embbedings consiste em ter um valor fixo para cada palavra, onde esse valor pode representar a proximidade entre palavras. Sendo assim, é agrega valor semântico às palavras.

PORQUE USAR EMBEDDINGS

- **Redução de Dimensionalidade:** Permitem reduzir a dimensionalidade dos dados, tornando-os mais eficientes de processar, sem perder informações importantes.
- **Aprendizado Eficiente:** Algoritmos de aprendizado de máquina muitas vezes funcionam melhor com entradas numéricas. Embeddings transformam dados não numéricos em formatos adequados para esses algoritmos.
- **Generalização:** Embeddings capturam relações semânticas entre objetos, permitindo que os algoritmos generalizem melhor a partir dos dados de treinamento para novos exemplos.
- **Melhora da Performance:** Em tarefas de PLN, como análise de sentimentos, tradução automática e processamento de texto, embeddings pré-treinados podem melhorar significativamente o desempenho do modelo.

COMO UTILIZAR EMBEDDINGS?

- Podemos encontrar embeddings pré-treinados.
- Um dos exemplos mais conhecidos é o Word2Vec, que mapeia palavras em vetores contínuos. Isso é usado em tarefas de PLN, como análise de sentimentos e classificação de texto.
- Porém podemos criar o nosso embeddings, pois o Keras disponibiliza uma camada do tipo embeddings.

CAMADA EMBEDDINGS NO KERAS

- Para usar a camada embedding temos alguns parâmetros:
  - input_dim: Tamanho do vocabulário
  - output_dim: Tamanho do vetor denso
  - input_length: Tamanho máximo da sequência

- Camada Embedding deve ser conectada a uma camada Flatten, por sua vez a camanda Flatten conectada a uma Densa.

SERÁ QUE É POSSÍVEL MELHORAR O CLASSIFICADOR DE E-MAILS?