Guia de embeddings

O serviço de embedding na API Gemini gera embeddings de última geração para palavras, frases e sentenças. Os embeddings resultantes podem ser usados para tarefas de PLN, como pesquisa semântica, classificação de texto e clustering, entre muitos outros. Esta página descreve o que são embeddings e destaca alguns casos de uso importantes do serviço de embedding para ajudar você a começar.

O que são embeddings?

Embeddings de texto são uma técnica de processamento de linguagem natural (PLN) que converte texto em vetores numéricos. Os embeddings capturam significado semântico e contexto, resultando em um texto com significados semelhantes com embeddings mais próximos. Por exemplo, as frases "Levei meu cachorro ao veterinário" e "Levei meu gato ao veterinário" teriam embeddings próximos uns dos outros no espaço vetorial, já que ambas descrevem um contexto semelhante.

Isso é importante porque desbloqueia muitos algoritmos que podem operar em vetores, mas não diretamente no texto.

É possível usar esses embeddings ou vetores para comparar diferentes textos e entender como eles se relacionam. Por exemplo, se os embeddings do texto "gato" e "cão" estiverem próximos, você poderá inferir que essas palavras são semelhantes em significado, contexto ou ambos. Esse recurso permite vários casos de uso descritos na próxima seção.

Casos de uso

Os embeddings de texto servem para vários casos de uso de PLN. Exemplo:

- Recuperação de informações: o objetivo é recuperar um texto semanticamente semelhante considerando um texto de entrada. Uma variedade de aplicativos pode ser compatível com um sistema de recuperação de informações, como pesquisa semântica, resposta a perguntas ou resumo. Consulte o notebook de pesquisa de documentos (https://ai.google.dev/gemini-api/tutorials/document_search?hl=pt-br) para ver um exemplo.
- Classificação: é possível usar embeddings para treinar um modelo e classificar documentos em categorias. Por exemplo, se você quiser classificar os comentários do usuário como negativos ou positivos, use o serviço de embeddings para receber a representação vetorial de cada comentário e treinar o classificador. Consulte o exemplo de classificador Genmini (https://ai.google.dev/examples/train_text_classifier_embeddings?hl=pt-br) para mais detalhes.
- Clustering: a comparação de vetores de texto pode mostrar como eles são semelhantes ou diferentes. Esse recurso pode ser usado para treinar um modelo de clustering que agrupa textos ou documentos semelhantes (https://ai.google.dev/examples/clustering_with_embeddings?hl=pt-br) e para detectar anomalias nos dados (https://ai.google.dev/examples/anomaly_detection?hl=pt-br).
- Banco de dados vetorial: é possível armazenar os embeddings gerados em um banco de dados vetoriais para melhorar a precisão e a eficiência do aplicativo de PLN. Consulte esta página para saber como usar um banco de dados vetorial para converter solicitações de texto em vetores numéricos (https://cloud.google.com/alloydb/docs/ai/work-with-embeddings?hl=pt-br).

Embeddings flexíveis

O modelo Gemini Text Embedding, começando com `text-embedding-004`, oferece tamanhos de embedding elásticos abaixo de 768. É possível usar embeddings elásticos para gerar dimensões de saída menores e economizar custos de computação e armazenamento com pequena perda de desempenho.

A seguir

- Se estiver tudo pronto para começar a desenvolver, você poderá encontrar o código executável completo nos guias de início rápido para Python (https://ai.google.dev/tutorials/python_quickstart?hl=pt-br#use_embeddings), Go (https://ai.google.dev/tutorials/go_quickstart?hl=pt-br#embeddings), Node.js (https://ai.google.dev/tutorials/node_quickstart?hl=pt-br#embeddings) e Dart (Flutter) (https://ai.google.dev/tutorials/dart_quickstart?hl=pt-br#embeddings).

Última atualização 2024-04-18 UTC.