Wikidata Embedding Projekt

Glossar

Hiermit gibt Wikimedia Deutschland einen Überblick über die wichtigsten Begriffe, die im Zusammenhang mit dem Wikidata Embedding Projekt auftauchen.

Wissensgraph
Wikidata ist ein offener Wissensgraph. Informationen zu Personen, Orten, Dingen, Ereignissen, Konzepten und vielem mehr werden nicht nur strukturiert abgespeichert, sondern auch miteinander verbunden und in Beziehung gesetzt. Zum Beispiel: „Berlin – Hauptstadt – Deutschland“.

Strukturierte Daten
Strukturierte Daten sind nach klar definierten Regeln in einem festgelegten Format gespeichert, sodass Computerprogramme sie einfach verstehen können. Beispiele: Wikidata oder eine Tabelle mit festgelegten Kategorien für Kundenadressen. Im Gegensatz dazu liegen unstrukturierte Daten nicht in so klaren Formen vor. Computerprogramme müssen sie zunächst analysieren, um sie verarbeiten zu können. Beispiel: Wikipedia-Artikel – viele Wörter und lange Texte, die keinem Regelsystem folgen. Aber auch Fotos, Videos oder Chatverläufe zählen dazu.

Embedding (Einbettung)
Beim Embedding-Prozess werden Informationen (z. B. Wörter, Sätze, Bilder) in eine Zahlenform übersetzt, damit etwa Machine-Learning-Modelle sie verarbeiten können. Diese Zahlen bilden einen Vektor (eine geordnete Liste von Nummern), der die Bedeutung der Information beschreibt. Beispiel: Das Wort „Hund“ bekommt einen Vektor, der sehr ähnlich aussieht wie der von „Welpe“, aber ganz anders als der von „Auto“. So können Programme erkennen, welche Inhalte ähnlich sind, auch wenn nicht exakt die gleichen Wörter verwendet werden.

Vektor
Ein Vektor ist das Ergebnis des Embedding-Prozesses. Dabei erhält jede Datei mehrere Koordinaten wie auf einer Karte – nur dass diese Karte hier den Bedeutungsraum von Daten abbildet. Je näher zwei Vektoren beieinander liegen, desto ähnlicher ist ihre Bedeutung. Alle Vektoren werden in einer Vektordatenbank gespeichert, die sich nach ähnlichen Vektoren durchsuchen lässt.

Semantische Suche
Wenn ein Programm Vektoren vergleichen kann, ist eine neue Art von Suche möglich: die semantische Suche. Hierbei wird nicht nach exakten Wörtern gesucht, sondern nach der Bedeutung. Beispiel: Wenn jemand nach einem Fahrzeug sucht, macht das System die Verbindung zu „Auto“, auch wenn das Wort in „Fahrzeug“ gar nicht vorkommt. Es ist die semantische Suche, die es ermöglicht, basierend auf vektorisierten Daten Abfragen in natürlicher Sprache durchzuführen.

Retrieval-Augmented Generation (RAG)
Manchmal sind die Datensätze, mit denen eine generative KI trainiert wurde, nicht vollständig oder veraltet. RAG löst das Problem fehlender oder veralteter Informationen von Sprachmodellen: Das Modell kann zusätzliche Fakten (augmented) abrufen (retrieval), bevor es antwortet (generation), und so Ergebnisse erzeugen, die auf aktuellem Wissen basieren.

Reranker
Wenn eine Suche viele Ergebnisse liefert, ist oft nicht unbedingt das erste Ergebnis das passendste. Ein Reranker ist ein spezialisiertes Large-Language-Model, das die Ergebnisse nach Relevanz neu sortiert.