

AI mit Oracle

Wir bauen einen RAG-basierten Chat

Teil 3.1 – Textkonvertierung und Chunking

Chunking

Was ist Chunking?

- **Definition:** Chunking bezeichnet das Aufteilen von Texten in kleine, sinnvolle Abschnitte (Chunks), die leichter analysiert und verarbeitet werden können.
- **Bedeutung:** Ein essenzieller Schritt in der Textverarbeitung und Analyse, insbesondere für KI gestützte Anwendungen.

Ziel der Präsentation

- **Herausforderungen beim Chunking** aufzeigen.
- **Methoden und Ansätze** zur Umsetzung erläutern (z.B Oracle Prozeduren, Python-Bibliotheken)

Chunking: Umwandlung eines Textes in sinnvolle Kleinteile

Warum wird es benötigt?

- **Verbesserung der Vektorgenaugkeit:** Kleinere Textabschnitte ermögliche präzisere Verarbeitung durch KI-Modelle
- **Optimierung der Analyse:** Chunks sind einfacher zu vergleichen und zu durchsuchen

Herausforderungen beim Chunking

- **Sinnvolle Trennung:** Abschnitte dürfen nicht in der Mitte eines Satzes oder Paragraphen geteilt werden
- **Dokumentenvariabilität:** Unterschiedliche Strukturen (z.B juristische Texte vs. Fließtext) erfordern unterschiedliche Trennlogiken

Umsetzung: Textdatei → Chunks

- **Chunking via Oracle:** Interne Logik durch gespeicherte Prozeduren
- **Chunking via Python:** Flexible Trennlogik, anpassbar durch Python-Bibliotheken (z.B RecursiveCharacterTextSplitter)

Chunking: Fazit

Dokumenten Aufbereitung ist essenziell

- **Formatierung:** Zeilenumbrüche sollten nicht übermäßig sein, dürfen aber auch nicht verloren gehen

Chunking ist individuell

- **Unterschiedliche Dokumente:** Je nach Textart müssen Chunks anders gesetzt werden

Stellenwert

- **Chunking:** Hat einen hohen Stellenwert in der Verarbeitung von textuellen Inhalten durch KI