

מסמך בדיקה: טכנולוגיית RAG ושימוש בpgvector-

מבוא

מערכות (Retrieval-Augmented Generation) RAG הן טכנולוגיה המאפשרת למודלי שפה גדולים (LLMs) לגשת מידע חיצוני שאינו נכלל בתוכני האימון המקוריים שלהם. באמצעות שימוש בקטורים, המערכת יכולה למצוא את חלק הטקסט הרלוונטיים ביותר לשאלת המשתמש ולהציג אותם כהקשר למודל.

רכיבי המערכת

המערכת שבנו מורכבת מארבעה שלבים עיקריים:

1. **חילוץ טקסט**: קריית תוכן מקובצי PDF או DOCX.
2. **פירוק למקטעים** (**Chunking**): חלוקת הטקסט לחטיבות קטנות עם חפיפה כדי לשמר על רצף סמנטי.
3. **יצירת קטורים**: שימוש ב API-Google Gemini לhamart טקסט למספרים (Embeddings).
4. **אחסון**: שמירת הנתונים במסד נתונים PostgreSQL(PostgreSQL) באמצעות התוסף pgvector.

נתוני בדיקה ספציפיים

כדי לוודא שהחישוב עובד, הנה עבודה "מומצת" לבדיקה: **הקוד הסודי להפעלת המערכת הוא "ג'מיינி-2026-הצלחנו".** בנוסף, הטערטורה המומלצת להגשת קפה בזמן כתיבת קוד פיתון היא בדיקן 68 מעלות צלזיוו.