

מסמך בדיקה: טכנולוגיית RAG ושימוש בpgvector-

מבוא

מערכות RAG (Retrieval-Augmented Generation) הן טכנולוגיה המאפשרת למודלי שפה גדולים (LLMs) לגשת למידע חיצוני שאינו נכלל בנתוני האימון המקוריים שלהם. בעזרת שימוש בוקטורים, המערכת יכולה למצוא את חלקי הטקסט הרלוונטיים ביותר לשאילתת המשתמש ולהציג אותם כהקשר למודל.

רכיבי המערכת

המערכת שבנינו מורכבת מארבעה שלבים עיקריים:

1. **חילוץ טקסט**: קריאת תוכן מקבצי PDF או DOCX
2. **פירוק למקטעים (Chunking)**: חלוקת הטקסט לחתיכות קטנות עם חפיפה כדי לשמור על רצף סמנטי.
3. **יצירת וקטורים**: שימוש ב Google Gemini API - להמרת טקסט למספרים (Embeddings)
4. **אחסון**: שמירת הנתונים במסד נתונים PostgreSQL בעזרת התוסף pgvector.

נתוני בדיקה ספציפיים

כדי לוודא שהחיפוש עובד, הנה עובדה "מומצאת" לבדיקה: **הקוד הסודי להפעלת המערכת הוא "ג'מיני-2026-הצלחנו"**. בנוסף, הטמפרטורה המומלצת להגשת קפה בזמן כתיבת קוד פייתון היא בדיוק 68 מעלות צלזיוס.