המאמר היומי של מייק: 13.09.25

On the Theoretical Limitations of Embedding-Based Retrieval

סקירה 504:

אתחיל מכך המאמר הזה של דיפמיינד עורר הרבה מאוד באזז ולדעתי רובו המוחלט נבע מחוסר הבנה של המאמר. אפילו המנכ״לֹ של חברת pinecone נאלץ לפרסם הודעת הבהרה בנוגע המאמר הזה.

אז על מה כל המהומה? בין ״הטענות״ שהועלו בעקבות המאמר הזה היו כאלו כמו ״אמרתי לכם שהגאג שלכם לא עובד ועכשיו זה הוכח מתמטית״ וגם ״אין מנוס חייבים לפתח מודלי שפה עם אורך קונטקסט באורך הגלות כי הראג לא באמת מתפקד״ ועוד לא מעט בסגנון הזה. כמו שאתם כבר בטח הצלחתם להבין המאמר דן נושא של ראג (Retrieval Augmented Generation) ויש שם גם הוכחות מתמטיות.

המאמר טוען ומוכיח בצורה מתמטית את הטענה הבאה: עבור מימד אמבדינג m של טקסט נתון, כלומר מימד הפלט של מודל האמבדינג שהופך את צ'אנקים של טקסט לווקטורים ושומר אותם בדאטהבייס וקטורי, עבור מספר מספיק גבוה של צ'אנקים לא ניתן לאחזר את k (גם נתון) צ'אנקים הרלוונטים ביותר בסדר הנכון. הסדר הנכון אומר שהצאנ'ק הרלוונטי ביותר יקבל דמיון הגבוה ביותר, השני ברלוונטיות יקבל את הדמיון השני בגודלו וכדומה. המחברים מוכיחים את בצורה די יפה ואינטואיטיבית על ידי שימוש בטכניקות די בסיסיות מתורת המטריצות.

כמו שאתם כבר מתחילים להבין המרחק (נגיד אוקלידי או ווסרשטיין למי שאוהב את זה קשוח) די גדול. הראגים המודרנים כבר לא מסתמכים לא רק חיפוש של הצ'אנקים הרלוונטיים לשאלה וגנרוט תשובה בהתבסס עליהם אלא לא מעט פעולות כמו reranking, חיפוש נוסף אם אף אחד מהצ'אנקים לא מתאים לשאלה (או אין מספיק מידע בצ'אנקים שנמצאו) ועוד. למשל משלבים את החיפוש עם האלגוריתמים לחיפוש מילות מפתח כמו bm25 או bm42. בנוסף ניתן לנסח את השאלה בצורה אחרת בהתאם לצ'אנקים שאוחזרו והתשובה שהמודל גנרט בשלב הראשון של האיחזור. לפעמים המודל יכול לבצע כמה אחזורים - בקיצור הבנתם למה אני חותר כאן…

בקיצור בלי להמעיט בערכו של המאמר(ובהחלט יש בו ערך) הזה הוא לא הוכיח שהראג הוא קונספט מת וכל מיני הצהרות מהסוג

https://arxiv.org/abs/2508.21038