🚀המאמר היומי של מייק -08.10.24: ⚡️🚀  
CONTEXTUAL DOCUMENT EMBEDDINGS

מזמן לא סקרתי מאמר בנושא של Document Retrieval או DG. למעשה DG מהווה שלב של Retrieval Augmented Generated או RAG שמטרתו היא לאתר את המסמכים הרלוונטיים מסט המסמכים D. בדרך כלל זה נעשה על סמך קירוב של האמבדינגס(הנמדד על ידי מרחק קוסיין) של המסמכים ושל השאלה המופקים על מודל שפה כלשהו.

יש שכלולים למעטים לשיטה הזו, למשל לחלק כל מספר לצ'אנקים ומשתמשים בייצוג שלהם לחישוב הקרבה. יצא לא מזמן מאמר שהציע להוסיף תמצות לכל מסמך וכמובן קיימות עשרות או מאות אחרות.

אם יש בידינו דאטהסט של זוגות D\_T המורכבים מ- (שאלה, מסמך רלוונטי) אנו יכולים לעשות פיינטיון לאמבדינגס כאלו, כלומר לאמו שני מודלים: הראשון לחישוב אמבדינג של המסמכים והשני לחישוב אמבדינג של השאלה. בד״כ זה נעשה עם למידה ניגודית שמאומנת לקרב את ייצוגי של כל השאלה לייצוג המסמך הרלוונטי לו ומרחיקה אותו מכל מהייצוגים של שאר מסמכים.

המאמר מציע שיטה שמשפרת את התהליך הזה על ידי הוספת קונטקסט לייצוגים (=אמבדינגס) האלו. אם יש לנו מסמך שניתן לשייך אותו לכמה תחומים (=דומיינים) אנו רוצים שהאמבדינג של המסמכים ישתנה בהתאם בדומיין של השאלות. כלומר אם השאלות צפויות להיות מהדומיין של רפואה אנו רוצים שהאמבדינגס ישקפו את האספקטים הרפואיים ועבור דומיין הספורט שיהיה יותר ״מכוון״ לספורט. כלומר אנו צריכים כאן contextualized embedding בתלות בשאלות מ-D\_T ובסט המסמכים D בעצמו.

המאמר בוחר לעשות זאת על ידי אימון מודלי embedding למסמך או לטקסט בצורה הבאה. קודם כל אנו מחלקים את D לכמה קלסטרים לפי דומיינים (עם מודל embedding התחלתי). לאחר מכן המחברים ממקסמים את סכומי הלוסים הניגודיים על פני כל הקלסטרים האלו. כלומר אנו רוצים לבנות אמבדינג של שאלה ושל המסמך כך ש:

״אמבדינג של השאלה ושל המסמך הרלוונטי לה יהיו קרובים אחד לשני בתוך כל קלסטר (המדמה דומיין) ואילו ייצוג של השאלה יהיה רחוק מהכל המסמכים האחרים בקלסטר״.

כלומר אנו מתאימים את האמבדינגס כפונקציה של דומיין השאלות. המאמר גם מציע שיטה לבניה של באצ'ים (ככה מאמנים רשתות היום) כך שהרשת תלמד על שילובי המסמכים הקשים ביותר(למשל מסמכים דומים סמנטית אבל מדומיינים שונים).

בנוסף המאמר מציע לשלב את ייצוגי המסמכים לבנייה אמבדינג של מסמך נתון 'D. כלומר ייצוג של מסמך 'D מורכב משרשור של ייצוגי כל המסמכים מהדאטהסט ואמבדינגס של כל הטוקנים מ 'D (שהם תלויי הקשר המסמך כמובן). בהמשך מאמנים אנקודר למסמך בצורה דומה למה שתואר לפני אבל עם כמה טריקים לייעול האימון.

אציין שהמאמר לא כתוב בצורה מאוד ברורה….

https://arxiv.org/abs/2410.02525