# Review 166: In-Context Pretraining: Language Modeling Beyond Document Boundaries

**Paper: https://arxiv.org/abs/2310.10638v6**

https://huggingface.co/papers/2310.10638

כשאתם מאמנים מודל שפה (אימון מקדים) המשימה היא חיזוי הטוקן הבא. כאשר מאמנים מודל שפה בעל חלון הקשר (context) ארוך משרשרים כמה מסמכים שנבחרו באקראי ומאמנים תוך כדי חיזוי הטוקן הבא.

המאמר שנסקור היום ב-#shorthebrewpapereviews משכלל את הגישה הזו ומציע לשרשר מסמכים שהם קרובים מבחינת המשמעות אחד לשני במקום לבחור אותם באקראי. איך נבחרים מסמכים קרובים – לפי המרחק בין השיכונים(embedding) שלהם. אבל יש בעיה קטנה עם הגישה הנאיבית הזו. יש מסמכים שהם דומים ליותר מדי מסמכים ואז המודל "יראה״ אותם יותר פעמים מהאחרים שעלול כמובן לפגוע בביצועי המודל המאומן (יוצר סוג של overfit).

כדי להתגבר על סוגיה זו המחברים מציעים לתאר את כל המסמכים בדאטהסט על ידי גרף שמשקל של כל קשת בו (בין שני המסמכים) שווה לדמיון ביניהם.ֿ אחרי שיש לנו ביד גרף כזה ניתן לתאר את הבעיה בתור בעייה דומה לזו של איש מכירות המטייל (maximum travelling salesman problem) כאשר המטרה כאן למצוא מסלולים זרים (שהאיחוד שלהם מכיל את כל הקודקודים וכל קודקוד מופיע רק פעם אחת באיחוד הזה). פותרים את הבעיה הזו עם אלגוריתם די אינטואיטיבי.

לקודקוד נתון בוחרים כמה קודקודים דומים (NN-nearest neighbors) ובונים מהם מסלול בעל משקל כולל מקסימלי (סכום של כל משקלי הקשתות). כל פעם בוחרים קודקוד (מסמך) הקרוב ביותר לקודקוד האחרון שנבחר. מספר NN בכל תת-מסלול נבחר לפי אורך הקונטקסט (אורכו של כל שרשור המסמכים שווה לאורך הקונטקסט). אחרי שמסיימים לבנות כל שרשור מורידים את קוקודיו מהגרף הכולל.

לאחר מכן בוחרים מסמך עם הדרגה הכוללת המינימלית (השווה לסכום משקלי הקשתות שיוצאות ממנו) וחוזרים על התהליך. כך גורמים לכל מסמך להיכנס לשרשור עם מסמכים שכמה שיותר דומים לו.