**הצעת פרויקט** – קירוב SVD שמתכנס יותר מהר בסגנון Conquer & Divide

תקציר - תיאור הבעיה:  
אלגוריתם SVD מיוחס למשפחת האלגוריתמים מסוג gradient descent , שכידוע מועד למינימום לוקלי. נהוג לאתחל את ערכי הווקטורים בערכים רנדומליים. אנו מעוניינים לבדוק האם:  
1. ניתן למצוא שיערוך טוב בזמן מהיר לערכי הווקטורים וסטיות התקן, במקום להריץ SVD מלא.  
2. האם השערוך מהווה אתחול טוב יותר מרנדומלי וישפר את קצב ההתכנסות של אלגוריתם ה- SVD המלא.  
נשווה בין שלושת השיטות האפשריות (1,2 והSVD- המלא) ונבין את הפשרות שכל שיטה כרוכה בהן, על מנת שבהינתן פלטפורמה מסוימת לה אנו רוצים להתאים מערכת המלצה, נתאים את שיטת ההמלצה המתאימה ביותר (מבחינת דיוק וזמני ריצה). לדוגמא- באפליקציות בעלות בסיס נתונים גדול מאוד בהן זמן בניית המודל יכול להימשך זמן רב, אולי נעדיף לשערך את ערכי המודל מבלי להריץ SVD מלא שיקר בזמן, גם אם זה פוגע בדיוק.

תיאור האלגוריתם: (למען פשטות ההסבר האלגוריתם מנוסח כרקורסיה, אך בפועל ייעשה איטרטיבי)  
א. מחליטים על גודל מינימלי של דירוגים כקלט (אחוז מגודל המאגר), עליו מריצים SVD רגיל בסוף רקורסיה.  
ב. אם המאגר גדול מדי מחלקים אותו ל 2, על 2 החלקים מריצים רקורסיבית את האלגוריתם.  
ג. נחזיר את ממוצע הווקטורים וה- bias שחזרו עבור כל משתמש/פריט מכל אחד מהחצאים.  
(במקרה שפריט/משתמש נמצא רק באחד משני המאגרים, ניקח את הווקטור היחיד שהתקבל עבורו).   
ד. נריץ SVD מלא (עד התכנסות) מאותחל עם הווקטורים/bias שהתקבלו.

תיאור הניסוי:

- השערת העבודה:

א. שלמות: בסוף האלגוריתם לכל משתמש ולכל פריט יהיה וקטור ו bias כמו ב SVD הרגיל.

ב. נכונות (קרבה לתוצאות ה SVD המקורי): כל latent feature יחושב עבור כמה מאגרים קטנים ובסופו של התהליך יבחר הממוצע לכן בסוף התרומה של כל latent feature תהיה קרובה לערך מסוים אם הוא הופיע ביותר מכמה מאגרים ולכן סטטיסטית ברוב המקרים יתרום באותה מידה כאילו הורץ ה-SVD המקורי.

ג. זמן ריצה: מכיוון שמריצים את אלגוריתם ה-SVD  על מאגרים קטנים יותר, האלגוריתם יתכנס מהר יותר עבור חלקיו. כמו כן, חישוב SVD על תתי מאגרים זרים מאפשרת ריצה מבוזרת של אלגוריתם זה.

- בסיסי נתונים:  
מאגר הנתונים: yelp\_training\_set\_review שניתן לנו בעבודות ההגשה.

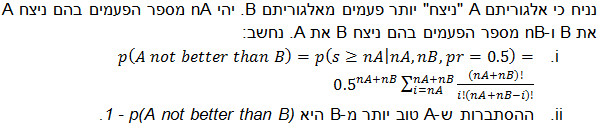
אם יש לך מאגר נתונים גדול יותר שנוכל לעבוד אתו, נודה לך אם תוכל לספק לנו.

- מדדים לביצועי האלגוריתם:  
א. קירבה לSVD המקורי: ערכים שימדדו בין שיטה 1 לבין הSVD המלא-

* סכום המרחק האוקלידי בין הווקטורים.
* סכום המרחק בין ערכי סטיית התקן של המשתמשים והפריטים.

נרצה שהסכומים הללו יהיה קטנים ככל האפשר.

ב. דיוק: RMSE.  
ג. ממוצע זמני ריצה של מספר קבוע של ריצות, הקבוע ייקבע בהתאם לצרכים של מובהקות סטטיסטית.

- מובהקות סטטיסטית: כאשר ניצח יהיה:   
א. זמן ריצה קצר יותר  
ב. כמות הפעמים שהאלגוריתם חזה דירוג יותר נכון מאשר SVD מלא.  
  
****

כאמור, המובהקות הסטטיסטית הינה https://lh5.googleusercontent.com/EZK1bROyE0conqvEpXlGUexSEplA2sF6Xp3ivhNHEHXT6i2qX1mwRtSMX4KkLyi4gAeYypjfSC_DrVkNSnWdVgQ8DydXfPuNS4YrDBAcu_WQ9bru5muHsi81jHLs-qD0DZMLnQM