מעבדה 2 חלק 2- הנחיות הגשה

עליכם להגיש סיכום של המעבדה **במסמך PDF או HTML אחד** כתוצר של מחברת ג'ופיטר - ניתן יהיה להגיש רק קובץ אחד.
המחברת תכלול קוד, טבלאות רלוונטיות, ויזואליזציות, את הסיכומים והתשובות ניתן לכתוב בכתב יד ולהוסיף כתמונות למסמך (בתנאי וקריא) או להקלידן. יש לתעד את הקוד באופן ברור עבור כל פעולה.
יש לתת כותרת ברורה וקריאה עבור כל גרף וטבלה (ולא כותרת כללית לכלל הגרפים מאותו סוג)
מקרא עבור הגרפים (או משפט הסבר)
במידה והתבקשתם להציג ערכים, בבקשה לכתוב במשפט את המסקנה מהגרף.
מומלץ לבנות את הקוד בצורה גנרית כך שיהיה נכון לכל אוסף טקסטואלי (ותהפוך להיות כלי עבודה עבור כל ניתוח טקסטואלי)
כנתונים יש להשתמש באוסף הנתון.

1. ייצוג המסמכים והשאילתה :
שימו לב - תרגיל זה הינו תרגיל ניסויי - יש להציג את הגדלים והזמנים עבור מבני הנתונים ואופני החיפוש שלכם (זמני שאילתה).
כמובן שניסוי השוואתי ומבוקר דורש לשמור על כמות שווה של מסמכים ושאילתות זהות עבור כל ניסוי.

- **הסבר כללי**:

i. **מבני נתונים** - עבור כל מבנה נתונים יש לבצע את הניתוח עבור 100 ו 1000 טקסטים מהמאגר שעבדתם עליו במעבדה הקודמת.
יש להציג :
1. יש לשמור את מבנה הנתונים בקובץ (פורמט - טקסט, פיקל או כל כלי אחר לבחירתכם)
2. יש להציג את גודל הקובץ.
3. מהו אחוז המידע הרלוונטי בייצוג (לדוג' כמה אחדות במטריצה)?

ii. **השאילתה** - השאילתה הינה פונקציה המקבלת כקלט את הביטוי לחיפוש (מילה אחת או יותר) ומחזירה את מספרי המסמכים הרלוונטיים.
על השאילתה לקחת בחשבון את אותם תהליכי עיבוד מקדים שבוצעו על המילון.
עליכם לממש עבור כל צורת ייצוג, פונקציית חיפוש מתאימה.
יש להציג מדידת זמני חיפוש עבור ביטויים תואמים בכל צורת ייצוג (לדוגמא השוואת זמנים עבור חיפוש הביטוי **"XYZ"** במטריצה בוליאנית, ביחס לחיפוש ביטוי זה באינדקסים הפוכים).

המודלים אותם עליכם לממש ולהציג :

1. מטריצה בוליאנית

i. צרו מטריצה בוליאנית כאשר העמודות מייצגות את המסמכים ושורות עבור המושגים (השורות הן המילון שבנינו).
ii. ערכי המטריצה יהיו אחד או אפס בהתאם. (בבחירת מבנה הנתונים יש לחשוב על מימוש השאילתה)
iii. כתבו שאילתה - פונקציה המחזירה את מספרי המסמכים עבור מושג או אוסף מושגים (OR,NOT,AND)
iv. שימו לב ניתן לממש זאת כסוג של bitwise operations בין ווקטורים המייצגים מושגים שונים .

2. **אינדקסים הפוכים**
   i. מבנה אינדקסים הפוכים בסיסי - צרו או השתמשו במבנה נתונים דינאמי, אשר עבור כל מושג יחזיק את כמות המסמכים בהם הוא מופיע, ואת מספרי המסמכים בהם הוא מופיע. שימו לב לאופן השמירה של המסמכים השונים (סדר)
   ii. שאילתה #1
      1. כתבו פונקציה המחזירה את מספרי המסמכים עבור מושג או אוסף מושגים (OR,NOT,AND)
      2. שימו לב יש לממש merge כפי שנלמד בכיתה, ולהעזר בעובדה שיש לנו תדירות מסמכים עבור כל מושג (שקפים 28-19 מהרצאה)
3. **סעיף רשות - אינדקסים הפוכים שיפור 1** הוספת מצביעי קפיצה ושאילתה מתאימה.
   1. שפרו את מבנה הנתונים - הוסיפו למבנה הנתונים גם מצביעי קפיצה (skip pointers) לצורך ייעול השאילתה.
   2. יש לבחור שלושה גדלים שונים של קפיצות :
      a. הערך האמצעי - צריך לעמוד בכלל שניתן בהרצאה - עמ 44
      b. יש להציג את הגדלים השונים של מבנה הנתונים
      c. יש להציג את זמני החיפוש
   3. על איזה עוד מדד אנו משלמים בבחירת גודל הקפיצה (Insertion cost) , חישבו על דרך למדוד אותו. (סיכום תיאורטי)
4. **סעיף רשות - אינדקסים הפוכים שיפור 2** הוספת תדירות ומיקום מילים במסמך
   1. שפרו את מבנה הנתונים מהסעיף הקודם על ידי הוספת תדירות המושגים בכל מסמך, והוספת המיקום של המושג (מיקום התו שמתחיל את המילה בכל מסמך)

2. סיכום

עליכם לכתוב סיכום קצר במילים שלכם (2-4 פסקאות), מתוך התייחסות לידע האישי שלכם בתחום מדעי הנתונים, והחומר הנלמד עד היום בתואר.

- מה למדתם ממעבדה זו - יש להציג טבלאות השוואתיות (גדלים ומדידות זמנים).
- יש לדון בתוצאות הניסויים השונים שלכם.
- יש לסכם את המסקנות הנובעות מההשוואות השונות.
- עליכם להסביר בקצרה מעבדה זו בהתייחס לחומר התיאורטי שנלמד עד היום בתחום מדעי הנתונים.

בהצלחה !