מעבדה 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 המחברת בצורה כללית כך שתהיה נכונה לכל אוסף טוויטים (ותהפוך להיות כלי עבודה עבור כל ניתוח טקסטואלי)
כנתונים ניתן לבחור להוריד טוויטים כרצונכם או להשתמש באוסף הנתון.

1. חלק א - בניית המילון (עיבוד לינגוויסטי)
    1. נתרכז בכלים
        i. NLTK
        ii. TextBlob
        iii. spaCy
    2. סכמו על כל אחד מהכלים במילים שלכם (יש להתייחס לנקודות הבאות):
        i. תכונות ויכולות כלליות (בקצרה)
        ii. תכונות ויכולות הרלוונטיות לבניית המילון שלנו
        iii. מבני נתונים שימושיים
    3. יש להדגים כל אחת **מהפעולות** הבאות, בעזרת כל אחת מהספריות שניתנו, או להסביר למה לא ניתן.
    הדוגמא צריכה להיות בקנה מידה קטן, ניתן להדגים על נתונים מלאכותיים (טקסט שלכם שידגים את הפעולה).
    שימו לב לתיעוד של הפונקציות השונות שאתם משתמשים בהם.
    אלו **הפעולות** :
        i. טוקניזציה
        ii. תיקון שגיאות
        iii. למטיזציה (lemmatization)
        iv. סטמינג (stemming)
        v. ראשי תיבות (acronym or abbreviations)
        vi. מילים נרדפות (synonyms)
        vii. WSD - זיהוי מילה על סמך הקשר
        viii. (bass – a type of fish or tones of low frequency)
        ix. זיהוי ישויות בטקסט Named entity recognition - NER
    4. עליכם לבנות את המילון המיטבי - האוסף המצומצם ביותר של מילים שמייצגות את המסמכים.
        i. לאחר עיבוד מקדים ראשוני, כפי שבוצע במעבדה הקודמת, יש לבנות מילון ראשוני של מושגים. (טוקניזציה)
        ii. יש לבצע את **הפעולות** השונות (מסעיף 3) על המילון ולהציג את השפעתם באמצעות ויזואליזציות וטבלאות שונות לבחירתכם והבנתכם.
        iii. יהיה עליכם לנסות את התהליך בסדר שונה של פעולות כדי להגיע למסקנה מהו הסדר העדיף. (עליכם להציג לפחות שלושה ניסויים שונים בבניית המילון)
        iv. כחלק מהניתוח שלכם, יש למדוד זמנים עבור כל פעולה ועבור כלל התהליך, ולהציג את גדלי המילון השונים, וויזואילזציות רלוונטיות.

2. חלק ב' - ייצוג המסמכים והשאילתה :

שימו לב - חלק זה הינו חלק 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.

1. **הסבר כללי**:
   i. **מבני נתונים** - עבור כל מבנה נתונים יש לבצע את הניתוח עבור 100 ו 1000 טוויטים.

      יש להציג :
      1. יש לשמור את המבנה בקובץ (טקסט, פיקל או כל כלי אחר לבחירתכם)
      2. יש להציג את גודל הקובץ.
      3. מהו אחוז המידע הרלוונטי בייצוג (לדוג' אחדות במטריצה)?
   ii. **השאילתה** - השאילתה הינה פונקציה המקבלת כקלט את הביטוי לחיפוש (מילה אחת או יותר) ומחזירה את מספרי הטוויטים 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, ביחס לחיפוש ביטוי זה באינדקסים הפוכים).
2. **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 השאילתה)
   iii. כתבו שאילתה - פונקציה המחזירה את מספרי הטוויטים עבור מושג או אוסף מושגים (OR,NOT,AND)
   iv. שימו לב ניתן לממש זאת כסוג של bitwise operations בין ווקטורים המייצגים מושגים שונים .
3. **אינדקסים הפוכים**
   i. מבנה אינדקסים הפוכים בסיסי - צרו או השתמשו במבנה נתונים דינאמי, אשר עבור כל מושג יחזיק את כמות הטוויטים בהם הוא מופיע, ואת מספרי הטוויטים בהם הוא מופיע. שימו לב לאופן השמירה של הטוויטים השונים (סדר)
   ii. שאילתה #1
      1. כתבו פונקציה המחזירה את מספרי הטוויטים 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)
   iii. הוספת מצביעי קפיצה ושאילתה #2
      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)
   iv. 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)

3. חלק ג' - סיכום

1. עליכם לכתוב סיכום קצר במילים שלכם (4-5 פסקאות), מתוך התייחסות לידע האישי שלכם בתחום מדעי הנתונים, והחומר הנלמד עד היום בתואר.
2. מה למדתם ממעבדה זו - יש להציג טבלאות השוואתיות וגרפים (גדלים ומדידות זמנים) עבור חלק א וחלק ב של המעבדה.
3. יש לדון בתוצאות הניסויים השונים שלכם.
4. יש לסכם את המסקנות הנובעות מההשוואות השונות.
5. עליכם להסביר מעבדה זו בהתייחס לחומר התיאורטי שנלמד עד היום בתחום מדעי הנתונים.

בהצלחה !