מעבדה 3 - 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, כל אוסף טקסטים שעולה על 1000 מסמכים, או להשתמש באוסף הנתון מתחילת הקורס.

1. הסבירו וסכמו את נוסחת TFIDF במילים שלכם.
2. הסבירו וסכמו במילים שלכם על vector - space model
3. חלק א -בניית ייצוג BOW ו TFIDF בעזרת ספריות סטנדרטיות ו pandas
   יש לשים לב - תהליכי עיבוד מקדים רלוונטיים ניתן לבצע בעזרת כל ספריה שהיא.
   יש לציין איזה תהליכי עיבוד מקדים בוצעו.
   1. יש לבנות מטריצת ייצוג ל BOW - יצירת count vector עבור כל מסמך וכלל המילון.
   2. יש לבנות פונקציה המחשבת את מדד ה TFIDF למושג יחיד.
   3. יש לבנות מטריצת ייצוג ל TFIDF לכלל המסמכים שלנו
   4. יש לממש פונקציית דמיון המקבלת ווקטור המייצג מסמך וווקטור המייצג שאילתה ומחזירה את ערך ה cosine sim עבור הזוג.
   5. עליכם לבדוק את השאילתה למול כל אחד מהמסמכים ולהציג תוצאת חיפוש מדורגת.
      עליכם לבנות פונקציה שנותנת מידת קירבה בין השאילתה לכלל המסמכים ומחזירה N מסמכים מדורגים כתוצאת חיפוש. (N פרמטרי)
      i. יש לבנות פונקציה הנעזרת בפונקציית הדמיון ומדרגת את הקירבה של כל אחד ממסמכי ייצוג ה BOW.
      ii. יש לבנות פונקציה הנעזרת בפונקציית הדמיון ומדרגת את הקירבה של כל אחד ממסמכי ייצוג ה TFIDF.
   6. סכמו את התוצאות והסבירו במילים שלכם (יש להתייחס לנקודות הבאות):
      i. זמני בניית המטריצות
      ii. זמני דירוג המסמכים
      iii. אופי הדירוג שהתקבל בכל אחד מהייצוגים וההבדל בינהם.

4. חלק ב' - מימוש בעזרת פונקציות ספריה.
בחלק זה נתמקד בייצוג TFIDF בלבד
   1. הציגו לפחות 3 ספריות פייתון
      i. מצאו פונקציות בעזרתן ניתן לייצר מדד TF IDF והסבירו איך הפונקציות עובדות בספריות השונות (תוך דגש על נירמול).
      ii. מצאו פונקציות בעזרתן ניתן לחשב cosine sim.
   2. הציגו מימוש בעזרת 2 ספריות נבחרות. (מימוש כפול)
      i. יש לבנות מטריצת ייצוג לTFIDF
      ii. יש להציג חישוב cosine sim בין השאילתה לכלל המסמכים ולהחזיר תוצאה מדורגת של N מסמכים (כמו בחלק א)
   3. סכמו את התוצאות והסבירו במילים שלכם (יש להתייחס לנקודות הבאות):
      i. זמני בניית המטריצות
      ii. זמני דירוג המסמכים
      iii. אופי הדירוג שהתקבל בכל אחת מהספריות וההבדל בינהם לדעתכם.
5. 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 !