בס"ד

**תרגיל באחזור מידע**

עליכם לבנות מערכת המחפשת קבצי טקסט באנגלית ובעברית, אם השאילתה ניתנה באנגלית החיפוש יהיה על הטקסטים באנגלית אם השאילתה ניתנה בעברית החיפוש יהיה על הטקסטים בעברית.

המשתמש יזין שאילתה, עליכם לאחזר שתי רשימות מסמכים, רשימת מסמכים אחת העונה על השאילתה עצמה, בדיוק כמו שהמשתמש הזין, רשימה שניה תכיל מסמכים/קבצים העונים על מהות השאילתה אולם המסמכים **אינם** מכילים את המילים המופיעות בשאילתה.

בנוסף עליכם לבצע סקירה ספרותית של 5 עד 10 מאמרים המתמודדים עם משימה זו (או דומות לה).

דוגמא לשינוי שאילתה, המשתמש יזין: delete words

תוצאת המסמכים תהיה כאילו השאילתה הייתה: remove string. זו דוגמה, ייתכן שהחיפוש יהיה על מילים אחרות ולאו דווקא remove string. מה שחשוב שהמסמכים יכילו את התוכן המבוקש אבל המסמכים שיתקבלו יהיו ללא מילים המקוריות שהמשתמש הזין, דהיינו, ללא המילים " delete words". מבחינת התרגיל **אין הבדל** בין המילים words ל- word אותו דבר בין המילים remove ו- removing, משמע, אם השאילתה היא delete words אז המסמכים לא יכילו לא את המילה words ולא את המילה word ואותו הדבר גם למילים delete ו- deleting, שאילתה יכולה להכיל יותר ממילה אחת.

כמובן אסור להשתמש במנוע חיפוש ו-"לנקות" את התוצאות שלו.

שלבי התרגיל:

שימו לב, בכל השלבים עליכם לתת שתי תוצאות, רשימת המסמכים הרלוונטיים לשאילתה בכלל (delete words), ורשימה נוספת של המסמכים שאינם מכילים את המילים המופיעות בשאילתה (remove string), הן על הטקסטים באנגלית והן על הטקסטים בעברית.

בדוגמה דלעיל, המשתמש יזין את השאילתה delete words, הרשימה הראשונה תהיה על השאילתה delete words והרשימה השניה תהיה על "השאילתה" remove string (כמובן ייתכן שאפשר יהיה למצוא מילים טובות יותר).

על המסמכים להיות ממויינין בסדר יורד של רלוונטיות, המסמך הראשון יהיה הכי רלוונטי השני פחות רלוונטי וכו'. אורך רשימת התוצאות לא צריכה להיות ארוכה יותר מאשר 30 מסמכים.

בהצלחה