**ניהול נתונים באינטרנט – פרויקט 2 – קובץ תיעוד**

תום רוטנברג 313196727

עופר טלוסטי 311396303

***חלק א' – בניית אינדקס הפוך***

*עבור פקודת יחס עם ניתוב לתיקיית קבצים אנו מפעילים את הפונקציה*  א*שר תקרא את קבצי ה- בתיקייה, תחלץ מתוכם את המסמכים ותעדכן את המילון אשר ישמש אותנו למענה על השאלות בחלק הבא של הפרויקט.*

*נתחיל מתיאור המילון אשר ישמש אותנו להמשך המימוש:*

1. ***מילון המסמכים ()*** *– מילון זה ישמש אותנו למצוא את כלל הפרטים הנחוצים לנו בהינתן מזהה מסמך והוא בנוי באופן ההירככי הבא:*

*המפתח של המילון הראשי הינו והערך שלו הינו מילון נוסף הכולל מספר שדות:*

* *אורך הוקטור () – עבור חישוב אנו נדרשים לנרמל את התוצאה בגודל הוקטור של המסמך ושל השאילתא. על מנת לחסוך חישובים בזמן ריצה (והמתנה של המשתמש) אנו שומרים את גודל הוקטור של המסמכים השונים במילון כך שנוכל לשלוף אותו מבלי לבצע חישובים נוספים.*
* *תדירות מרבית () – עבור חישוב פרמטר ה- נרצה לנרמל את תדירות המילים במסמך לפי תדירות המילה השכיחה ביותר במסמך. לכן נרצה לחשב ולשמור ערך זה במילון ולהשתמש בו בעת חישוב הפרמטר .*
* *ספירת מילים במסמך () – עבור חישובי נרצה לשמור את מספר המילים הכולל בכל מסמך.*
* *מילון מילים במסמך () – במילון זה המפתח יהיה מילה המוכלת במסמך הנתון והערך המתקבל יהיה הפרמטר של המילה במסמך הספציפי.*

*להלן דוגמה מהמילון להמחשת המבנה שלו:*

*Text

Description automatically generated*

1. ***מילון המילים ()*** *– מילון זה ישמש אותנו למצוא את כלל הפרטים הנחוצים לנו בהינתן מילה במאגר המילים הכולל של "מנוע החיפוש שלנו" והוא בנוי באופן ההיררכי הבא:*

* *תדירות מסמכים הפוכה () – נרצה לחשב פרמטר זה עבור כל מילה כחלק מחישוב המשקלים של המילים במסמכים השונים ().*
* *מילון המסמכים המכילים את המילה () – במילון זה המפתח הינו מזהה המסמך () ולו שני ערכים:* 
  + *ספירת המילה במסמך () – פרמטר זה ישמש לטובת חישוב הפרמטר .*
  + *תדירות המילה במסמך () – נרצה לחשב פרמטר זה עבור כל מילה במסמך כחלק מחישוב המשקלים של המילים במסמכים השונים ().*

*להלן דוגמה מהמילון להמחשת המבנה שלו:*

*Text

Description automatically generated*

*עתה נחזור לאופן פעולת הפונקציה :*

* *בעבור כל קובץ בתיקיית האינפוט נקרא לפונקציה אשר תמצא את כלל המסמכים (ישות ).*
* *עבור כל מסמך נבצע את הפעולות הבאות:* 
  + *נחלץ את מזהה המסמך () ונאתחל רשומה מאופסת במילון המסמכים ().*
  + *נחלץ את תוכן כל המסמכים באמצעות הפונקציה אשר מקבלת מזהה מסמך ומחזירה מחרוזת הכולל את כלל המילים במסמך תחת הישויות הבאות: אשר נמצאו רלוונטיות לצרכי משימה זו.*

*נציין כי בשלב זה אנו מבצעים ניקוי ראשוני על המילים במסמך באמצעות הפונקציה*

* + *נתאחל פרמטר לוקאלי אשר ישמור את תדירות הגבוהה ביותר שזיהינו עד כה למילה במסמך והוא יעודכן בכל איטרציה של הלולאה הבאה.*
  + *נעבור בלולאה על כלל המילים אשר חולצו מהמסמך ונבצע את הפעולות הבאות:* 
    - *ננקה את המילה באמצעות הפונקציה – פונקציה זו תנקה את המילים מסימני פיסוק, תבצע פעולת ותקטין את כלל האותיות במילה לאותיות קטנות. כמו כן, נבצע בדיקה האם המילה כלולה במאגר ולכן נרצה שלא להכניסה למאגר שלנו (כפי שנלמד בשיעור).*
    - *נציין כי פעולות הניקוי נעשות תוך שימוש בספריה בדגש על מאגר ה- והחבילה .*
    - *במידה והמילה ואלידית (לא כלולה במאגר ה-) נרצה להזין את המילונים שלנו באמצעות שימוש בשתי הפונקציות הבאות:* 
      * *– הפונקציה תאתחל רשומה עם המסמך והמילה הנתונים עם ערך מאופס.*
      * *– הפונקציה תבדוק האם קיימת רשומה עם המילה הנתונה, במידה ולא תזין רשומה מאופסת. לאחר מכן, תבדוק המסמך נמצא במילון הפנימי (), במידה ולא תזין רשומה מאופסת גם כן. לבסוף, הפונקציה תעלה את המונה במילון.*
  + *נעדכן את מונה המילים במסמך () ונחשב את שכיחות המילה במסמך. במידת הצורך נעדכן את התדירות המקסימלית הלוקאלית שאתחלנו לפני הכניסה ללולאה.*
* *בסיום המעבר על כלל המילים במסמך, נזין למילון את תדירות המילה השכיחה ביותר () ונקרא לפונקציה .*
* *הפונקציה עוברת על כלל המילים במסמך הנתון, מנרמלת את תדירות המילים במסמך לפי התדירות המקסימלית שחושבה ושומרת את הערך המתקבל במילון (כפי שהוסבר מעלה).*

*בסיום המעבר על כלל הקבצים ופרסורם, המילונים מכילים בתוכם את כלל המסמכים והמילים אך ישנם מספר פרמטרים אשר נצטרך וטרם חושבו – :*

1. *הפונקציה אחראית לחישוב זוג הפרמטרים הראשונים והזמנתם למילונים כפי שהוצגו מעלה. נציין כי אופן החישוב של הפרמטרים הוא כפי שנלמד בכיתה ולא נכנסנו במסמך זה לאופן החישוב שלהם).*
2. *הפונקציה אחראית על חישוב אורך הוקטור עבור כלל המסמכים במאגר שלנו על מנת לחסוך בחישובים "בזמן אמת".*

*לבסוף נשמור את המילונים לקובץ בשם* כנדרש.

***חלק ב' – תום משלים***