**דוח פרוייקט מנוע אחזור - חלק א'**

**מגישים**

**עומר נגר 307937714, אסף זקס 302329693**

**פרק 1 - עיצוב התוכנה**

a.

הסבר על איך התוכנית בנוייה

**Package Parser:**

**Class Master:**

**פונקציות המחלקה:**

run –

הפונקציה מנהלת את עבודת המחלקות השונות בחבילה. מפעילה את readFile שקוראת את הטקסטים לתוך docTexts, אותם היא תשלח לparser כדי לקבל docMD עם המידע על המסמך והמילים השונות בו. הפונקציה מקבלת כארגומנט Boolean הבורר בין הפעלת הסטמר לביטולו.

getDocAmount-

הפונקציה מחזירה את מספר הdocMD השמורים בזיכרון.

saveDocMD-

הפונקציה שומרת את מאגר המידע על המסמכים בדיסק הקשיח.

LoadDocMD-

הפונקציה מחזירה את מאגר המידע על המסמכים מהדיסק הקשיח.

**Class Parser:**

**פונקציות המחלקה:**

setMonths-

הפונקציה בונה עבור האובייקט מילון של שמות חודשים והערך אליו יש להחליף אותם על פי החוקים.

handleDoc-

הפונקציה יוצרת אובייקט מטיפוס docMD, אליו שומרת את כלל המידע על המסמך והמילים הקיימות בו באמצעות קריאה לפונקציות עזר.

calcMaxTF-

הפונקציה מחשבת מהו הערך המקסימלי של הופעות של מילה במסמך.

maxFreqTerm-

הפונקציה מחזירה את המילה הראשונה בעלת TFMax.

parse-

הפונקציה המבצעת את פעולת הparsing בפועל, עוברת על טקסט, מפרקת אותו לשורות אותן תפרק למילים ותחפש תבניות על פי חוקי הפרוייקט וכללים שהוספנו שיפורטו בהמשך.

wordToNum-

הפונקציה ממירה מילה לערך מטיפוס double.

finalEdit-

הפונקציה מבצעת אופרציות טקסט אחרונות על מילים שלא טופלו ע"י אף חוק, מוחקת תווים מיותרים שאינם טקסטואלים, ומפעילה stem במידה ונבחר.

addToWords-

הפונקציה מוסיפה את המילה לרשימת המילים של המסמך.

numToString-

הפונקציה מבצעת אופרציות טקסט על מספרים, מורידה דיוק של מספר לא שלם לעד 3 ספרות אחרי הנקודה.

**Class DocMD:**

**פונקציות המחלקה:**

toString -

הפונקציה משרשרת את המידע השמור על המסמך באמצעות פסיק ומחזירה את השרשור.

**Class DocText:**

**פונקציות המחלקה:**

getHeader –

הפונקציה מחזירה את כותרת המסמך.

getDocno –

הפונקציה מחזירה את מזהה המסמך.

getInnerText –

הפונקציה מחזירה את הטקסט הרשום במסמך.

**Package Indexer:**

**Class IndexDictionary:**

**פונקציות המחלקה:**

createIndexer -

הפונקציה מקבלת טבלת האש המכילה מונחים, עוברת על כל ערכי הטבלה, יוצרת קבצי פוסטינג למונחים ומוסיפה את המונח למילון, לאחר שהגיעו 200 מונחים, תדפיס את קובץ הפוסטינג לזיכרון המשני ותפנה את הזיכרון הראשי.

outputPostList -

הפונקציה מקבלת נתיב לשמירת קובץ הפוסטינג, פותחת קובץ חדש בנתיב זה, ומדפיסה לקובץ את כל ערכי הפוסטינג על כל מונח שיצרנו עבורו פוסטינג.

getIndexerPrint –

הפונקציה עוברת על כל המילון ויוצרת מחרוזת ארוכה שמכילה צמדים של מילה וכמות המופעים הכולל שלה, לטובת ההדפסה למשתמש.

saveToDisk –

הפונקציה שומרת את המילון לזיכרון בצורת קובץ טקסט.

loadDictionary –

הפונקציה טוענת את מילון מהזיכרון, מקבלת נתיב למיקום הפוסטינג, וערך בוליאני האם בוצע סטמינג או לא.

getNumOfUniqueTerms –

הפונקציה מחזירה את גודל המילון = כמות המילים השונות במילון.

Comp –

קומפרטור עבור המבנה של המילון – מפת עץ, ממיין לקסיקוגרפית כאשר אין חשיבות לcase.

**Class Posting:**

**פונקציות המחלקה:**

addToPostingList –

הפונקציה מעדכנת עבור מונח את הכתובת שבה צריך להשמר, ומחזירה את המונח.

getPosting –

הפונקציה מחזירה רשימה משורשרת של כל הפרטים על המופעים של המונח.

getMetaOfTerm –

הפונקציה מחזירה את הפרטים על המונח, שישמרו עבור כל מונח בקובץ הפוסטינג שלו.

getPath –

הפונקציה מחזירה את הנתיב של קובץ הפוסטינג שהמונח שמור בו.

**Class SegmentProcesses:**

**פונקציות המחלקה:**

processCorpus –

הפונקציה עוברת בלולאה על כל קבצי הסגמנט שנוצרו ע"י הפרסר, קוראת את הסגמנט מהזיכרון, ושולחת את כל המילים שנשמרו בסגמנט לעיבוד – הפיכה לאובייקט Term ועיבודים נוספים, לאחר מכן שולחת את כל המונחים לאינדקסר ומשם למילון. בסיום שומרת את המילון שנוצר לתיקייה שמכילה את הפוסטינג.

processSegment –

עבור כל מילה שחזרה מהסגמנט, הפונקציה מבצעת את החוק של האותיות הגדולות – ומעדכנת בהתאם את המילה, כמו כן יוצרת אובייקט Term מכל מילה ומוסיפה אותו לרשימה שתלך לאינדקסר. בנוסף גם קוראת לפונקציה המעדכנת את המופעים של מונח במסמכים.

addOccurrenceToTerm –

הפונקציה מוסיפה מופע נוסף למונח אם המונח כבר נצפה, מקבלת את פרטי המופע וקוראת לפונקציה במחלקה Term עם הפרטים כארגומנטים.

updateDocFreq –

הפונקציה מעדכנת את תדירות המופעים במסמכים שונים, קוראת לפונקציה במחלקה Term.

updateCaseToUpper –

הפונקציה מעדכנת את הcase של מונח שמופיע רק עם אות ראשונה גדולה. עבור כל המפתחות בטבלת המונחים שנוצרו, מפעילה על כל מונח מטודה בוליאנית של המחלקה Term, במידה וחזרה תשובה חיובית, מוחקת את המפתח הקודם, ומוסיפה את המונח מחדש עם אותיות גדולות בלבד.

getTheDictionary –

הפונקציה מחזירה את המילון.

**Class Term:**

**פונקציות המחלקה:**

addOccurrence –

הפונקציה מקבלת פרטים על מופע של מונח, ומוסיפה את המופע לרשימת המופעים של כל מונח.

updateDocFq –

הפונקציה מעדכנת את כמות המסמכים השונים שהמונח מופיע בהם.

updateToUpperCase –

הפונקציה מעבירה את המונח לאותיות גדולות אם האות הראשונה הינה אות גדולה, מחזיר אמת אם השתנה.

getOccurrence –

הפונקציה מחזירה את הרשימה המשורשת של כל מופעי המונח.

getDocFq –

הפונקציה מחזירה את כמות המסמכים שהמונח מופיע בהם.

getTotalFq –

הפונקציה מחזירה את כמות המופעים הכוללת של המונח.

addToTopBottom –

הפונקציה מוחשכת (בהערה), כאשר פעילה שומרת את 10 המונחים השכיחים ביותר ו10 המונחים הנדירים ביותר בשני רשימות שונות, משתמשת לצורך כך גם ב2 קומפרטורים.

**Class TermInDoc:**

**פונקציות המחלקה:**

setTerm –

הפונקציה מקבלת מילה ומעדכנת אותה למונח במסמך.

updateCapsToLower –

הפונקציה מעדכנת את המילה להיות עם אותיות קטנות בלבד.

getDocNo –

הפונקציה מחזירה את שם המסמך בו המונח מופיע.

getTermfq –

הפונקציה מחזירה את תדירות המונח במסמך.

isHeader –

הפונקציה מחזירה אמת אם המונח מופיע בכותרת המסמך.

isEntity –

הפונקציה מחזירה אמת אם המונח הינו יישות במסמך.

getTerm –

הפונקציה מחזירה את המונח במסמך.

**Class TermInDocList:**

**פונקציות המחלקה:**

tidToJson –

הפונקציה מקבלת מספר בין 1-20 שתשתמש בו באינקס למיקום שמירת הסגמנט, ועוברת על רשימה של אובייקטי TermInDoc ותוסיף אותם לאובייקט JSON לאחר מכן תדפיס את אובייקט הJSON לזיכרון המשני לפי האינדקס.

JsonToTid –

הפונקציה מקבלת כקלט כתובת, וטוענת מכתובת זו אובייקט JSON. לאחר מכן עוברת על האובייקט יוצרת ממנו אובייקטים של TermInDoc, לבסוף תחזיר רשימה של מונחים במסמך.

getList –

הפונקציה מחזירה רשימה משורשרת של מונחים במסמך.

setList –

הפונקציה מחליפה את הרשימה המשורשת של המונחים במסמך ברשימה חדשה.