**דוח פרוייקט מנוע אחזור - חלק ב'**

**מגישים**

**עומר נגר 307937714, אסף זקס 302329693**

**פרק 1 - עיצוב התוכנה**

1. ראשית, נתייחס לאלו ממחלקות המערכת בחלק א' שהורחבו כחלק מהדרישות בחלק ב', ולבסוף נרחיב על הPackage החדש searchRank:

**Package Parser:**

**Class Master:**

נזכיר - המחלקה מרכזת את עבודת המחלקות readfile, הparser, הstemmer ומרכזת את יצירת קבצי הסגנמט.

**פונקציונליות שנוספה:**

updateEntities-

הפונקציה מסננת מרשימת הentities של מסמך את אלו שאינם מופיעים במילון – כלומר מופיעים רק בו.

**Class Parser:**

נזכיר, המחלקה מייצגת אובייקט מטיפוס parser, מייצרת אובייקטים מטיפוס docMD. כל אובייקט אחראי על פרסור מסמך בודד.

**פונקציונליות שנוספה:**

calcDocSize-

הפונקציה מחשבת ושומרת עבור כל מסמך את הגודל שלו לטובת חישוב Bm25.

**Class DocMD:**

נזכיר, המחלקה המייצגת מידע על מסמך. מחזיקה מצביע למאגר המושגים שנמצאו במסמך לזמן קצר לטובת יצירת קבצי הסגמנט.

**פונקציונליות שנוספה:**

setEntities-

הפונקציה מחשבת עבור המסמך את 5 היישויות הנפוצות ביותר שלו לטובת פונקציונליות היישויות.

getMaxTf-

הפונקציה מחזירה את הmaxTf של מסמך מסויים (חושב בחלק א' אך לא היה בו צורך עד כה).

**Package EngineUserInterface:**

לחבילה נוספו 2 מחלקות חדשות:

**Class SearchController:**

המחלקה אחראית על קבלת קלט query או קובץ queries ופרמטרים לטובת האחזור.

**פונקציות המחלקה:**

enterQuery-

הפונקציה מקבלת query לשמור כשדה במחלקה.

singleQuery-

הפונקציה תופסת event של הכנסת query, קוראת לenterQuery, בודקת את הקלט ומפעילה את handleSingleQuery של המחלקה model לטיפול בשאילתה ולבסוף מפעילה את openResultWindow לטובת הצגת התוצאות.

typeQueryPath-

הפונקציה מקבלת נתיב לקובץ queries לשמור כשדה במחלקה.

browseQueries-

הפונקציה מטפלת בevent הכנסה של נתיב של queries דרך fileChooser.

pressGoBig-

הפונקציה תופסת event של הרצת שאילתה על קובץ מסמכים ומפעילה את הפונקציונליות במודל. פותחת חלון תוצאות בסיום.

openResultWindow-

הפונקציה מציגה את תוצאות האחזור ע"י יצירה ופתיחה של ResultController והדפסת התוצאות מהמודל אליו.

**Class ResultController:**

המחלקה יוצרת אובייקט חלון לטובת הצגת תוצאות האחזור.

**פונקציות המחלקה:**

displayResult-

הפונקציה מציגה בפועל את תוצאות האחזור בחלון גלילה.

saveResult-

הפוקציה פותחת חלונית לבחירת תקייה לשמירת קובץ תוצאות השאילתה לטובת שימוש בtreceval, ושומרת את התוצאות בקובץ txtבתיקייה.

**Package searchRank:**

**Class Searcher:**

המחלקה מנהלת טיפול בquery בודד,משלב הparse, הrank ועד שרשור התוצאות.

-parseQuery

הפונקציה קוראת לparser לביצוע parse ושומרת את המילים כשדה במחלקה.

rank-

הפונקציה מקבלת ranker, מפעילה את הפונקציה handleQuery שלו ויוצרת מתוצאות המסמכים שהחזיר String תוצאות להצגה למשתמש. במידה ואופציית הישויות נבחרה, תופעל פונקציה לשרשור הישויות בעלות הדירוג הגבוה.

-shrinkOrderedDocs

הפונקציה מצמצמת את מאגר המסמכים שמחזיר הranker כך שישמרו ה50 הרלוונטיים ביותר בלבד.

queryStringResults-

הפונקציה משרשרת לשדה התוצאה של המחלקה את תוצאות המסמכים הרלוונטיים ביותר שהחזיר הranker לטובת הצגה בממשק הגרפי.

stringEntities-

הפונקציה משרשרת לשדה התוצאה של המחלקה את היישויות הרלוונטיות ביותר בכל מסמך. לכל ישות תחשב rank ע"י לוג 2 של מספר המופעים במסמך חלקי מספר המופעים הכולל של כלל היישויות במסמך.

**Class Ranker:**

המחלקה אחראית על דירוג המסמכים עבור שאילתה מסויימת.

handleQuery-

הפונקציה בונה את מאגר המסמכים הרלוונטיים לתוצאות השאילתה ומפעילה עליו את הפונקציה לדירוג מסמכים ע"פ Bm25.

Bm25Rank-

הפונקציה מחשבת עבור כל מסמך ציון bm25 ע"פ שכלול עבור כל הterms מהשאילתה שהופיעו בו.

calcAvgLength-

הפונקציה מחשבת את האורך הממוצע של מסמך בקורפוס.

semanticTreat-

הפונקציה מפעילה פונקציה של טיפול סמנטי - fetchFromWord2Vec, ומוסיפה את תוצאותיה למאגר המילים בquery.

fetchFromWord2Vec-

הפונקציה מבצעת טיפול סמנטי בשאילתה, באמצעות שימוש בקוד פתוח (צורף בjar).

למידע נוסף - <https://github.com/medallia/Word2VecJava>

updatequerywords-

הפונקציה מעבירה את כלל המילים בשאילתה לאותיות קטנות טרם הפעלת הparser.