IR TASK DOCUMENTAION:

Tomer Flom – 303015671 & Adiel Perez – 308101062

**GUI:**

**ADMIN:** (admin.php)

**Add New Document:**

1. Press on the New Document button.
2. Choose your .txt file that you wish to index.
3. Choose Document's name and author.
4. Press Index File button.

**Update Existing Document:**

1. Press on the desired document from the admin documents list.
2. Update Document's Name and author if you wish.
3. Press Update Document button.

The document will be scanned and indexed again from scratch.

**Hide Existing Document:**

1. Press on the desired document from the admin documents list.
2. Toggle display by pressing on the EYE icon from the top right of the window.

**CLIENT:** (index.html)

**Using Search Bar:**

1. Press question-mark ICON from the right to watch supported operators for Boolean search.
2. Type a keyword at the Search Bar:

For example: ((A\* & B\*) | C) & (^X & ^Y)

Explanation: get all documents which contains a word ((starts with(**wildcard**) A **OR** B ) **OR** has the word C )  **AND** **NOT** contains the words X AND Y.

**The Data Base:**

|  |
| --- |
| word |
| a |
| an |
| of |
| … |

words stop\_list

|  |
| --- |
| word |
| window |
| car |
| house |
| …. |

documents

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| document\_id | filename | author | name | Format | isHidden |
| 1 | yellow submarine.txt | Beatles | Yellow submarine | txt | 0 |
| 2 | Let it be.txt | Beatles | Let it be | txt | 0 |
| 3 | dust in the wind.txt | Kansas | Dust in the wind | txt | 0 |
| … | …. | …. | …. | … | … |

word\_in\_document

|  |  |  |  |
| --- | --- | --- | --- |
| auto\_id | word | document\_id | pos |
| 1 | yellow | 1 | 32 |
| 2 | submarine | 1 | 33 |
| 3 | sea | 1 | 88 |
| 4 | sea | 1 | 94 |
| 5 | close | 2 | 1 |
| 6 | moment | 2 | 4 |
| … | …. | …. | … |

הסבר לאלגוריתם הדירוג:

כפי שניתן לראות במבנה מסד הנתונים ליעל, לטבלת המילים אין את השדה hits. זאת מכיוון שבאלגוריתם אחזור המידע, ניתן לאחזר מטבלת word\_in\_document את כל המסמכים הרלוונטים ולמיין אותם בסדר יורד לפי מספר המילים הרלוונטיות המופיעות בכל מסמך ולאחר מכן לפי מספר הפגיעות(hits).

Example for keyword: (let&"it"&be) | yellow

במקרה הנ''ל לדוגמא, המערכת תשחזר למשתמש את המסמך let it be ואת yellow submarine. אך מי מהם ידורג קודם?

**סטופ-ליסט:** תחילה, יש לשים לב שהמילה it הינה חלק מהstop-list, לכן נכתבו סביבה מרכאות("it"). במידה והמרכאות היו מושמטות, מערכת האיחזור הייתה מתעלמת ממילה זו.

**תוצאות:** נניח שהמילה yellow קיימת במסמך yellow submarine.txt 64 פעמים. לעומת זאת, סך כל המילים let it be ביחד המופיעות במסמך let it be.txt הינו 56 לצורך הדוגמא.

במקרה הנ"ל, המיון יהיה כך ש let it be ידורג ראשון וyellow submarine ידורג במקום השני. כיצד המיון המתנהל?

ניתן לראות בדוגמא שבצוללת הצהובה נמצאו 64 פגיעות לעומת 56 פגיעות בלבד במסמך let it be. עם זאת, במסמך let it be נמצאו 3 מתוך 4 המילים בחיפוש ולכן הוא קיבל עדיפות גבוהה יותר לעומת הצוללת הצהובה שבו נמצאת רק מילה אחת מבין כל מילות החיפוש.

**אלגוריתם הדירוג** ממיין תחילה לפי מספר מילות החיפוש המופיעות במסמך ורק לאחר מכן לפי מספר הפגיעות בכל מסמך ומסמך.

**ארכטיקטורת המערכת:**

המערכת נבנתה בפלטפורמת web, על שרת מקומי – localhost.

צד שרת: PHP, mySQL

צד לקוח: HTML,CSS,JAVASCRIPT,JQUERY

פורמט מסמכים: קבצי .txt הנמצאים בתיקייה /documents

האובייקט IR.php נמצא בתיקייה classes/IR.php והוא האובייקט שאחראי על כל פעולות אחזור המידע. מתודות עיקריות של האובייקט הינן:

* function indexDocument($docName,$docAuthor,$filename)
* function searchBool($keyword)

סדר פעולות:

תהליך האינדוקס:

1. בחירת קובץ טקסט
2. הוספת מסמך חדש לטבלת המסמכים(במידה והמסמך הנל עדיין לא קיים. במידה וכן, עדכון בלבד).
3. מיפוי כל המילים בקובץ הטקסט בטבלת המילים(במידה והמילה עדיין לא קיימת עדיין).
4. הוספת הרשומה: word,document\_id,pos לטבלת הword\_in\_document.

תהליך האיחזור:

1. הזנת מחרוזת חיפוש ולחיצה על מקש האנטר. לדוגמא: (A | B\*) & C
2. הדף server/searchAPI.php נקרא בajax ומקבל את מילת החיפוש שהוזנה. הוא יוצר אובייקט IR, ושולח למתודת החיפוש שלו את מילת החיפוש.
3. האובייקט IR.php מבצע את החיפוש באמצעות פרסור(parse) מילת החיפוש ויוצר מהן **שאילתת sql אחת ויחידה** הנשלחת למסד הנתונים ומחזירה את טבלת התוצאות לדף searchAPI.php, שמחזיר json עם כל המידע לדף index.js.
4. דף הindex.js, מקבל את הjson בצורה דינמית לאחר שהתבצע החיפוש ומרנדר(render) את התוצאות בצורה ויזואלית עבור הלקוח, על המסך.