פרויקט גמר

קורס אחזור מידע

2015

Search Engine



מרצה: פרופ' עמנואל גרינגרד

קורס: אחזור מידע

מגישים :

רותם אמרגי - 302677174

שחר רומי - 305522948

הפרויקט נכתב ב-JAVA , המערכת היא מערכת לאחזור נתונים פנימית.

כלומר תוכנם המלא של המסמכים נמצאים על גבי המחשב המקומי, שנמצאים בתקיה database files .

כיצד המערכת עובדת :

**בשלב הראשון הוא בניית מסד הנתונים :**

הוספת קבצים נעשית על ידי GUI, לחיצה על הוספת קבצים) Import Dictionary ( Add File / , יפתחו את "הקובץ והתיקייה" ,תיבת הדו-שיח שממנו המשתמש יכול להוסיף קבצים בודדים או מרובים.

במערכת ניתן להוסיף באצטבה את כל הקבצים או קובץ בודד .

IndexFile יוצרת את קובץ index.txt , אשר מכילה רשימה של קבצים ועוברת עליהם, מחלקת כל קובץ למערך של מילים

וממיינת אותם ואז רושמת לקובץ (index.txt)..

קובץ index.txt נמצא בתיקיית Data .

לאחר מכן נוצרIndexObejct , אשר מכיל את רשימת כל המילים בכל הקבצים , עם מספר הפגיעות ולאיזה קובץ שייכת המילה.

**בשלב השני – צורת החיפוש :**

היא הכנסת מילת החיפוש, התוכנה תמצא את המסמכים שבהם המילה מופיעה .

תציג את המסמכים בתוספת מידע על המסמך ,מספר סידורי, תאריך וכותרת.

קבצים שנוצרים בתיקיית Data :

index file.txt - מכיל את הקבצים המאונדקסים

לדוגמא :

attacks,1 [0 , 1]

- Posting file.txt מכיל את רשימת הקבצים והנתיבים שלהם.

אנדוקס הקובץ :

התוכנה עוברת על כל קובץ ושומרת אותו במערך ,היא מורידה ממנו את ה- StopWords ולאחר מכן הסריקה מוסיפה את המילים למערך ה-words (מילים שהם לא נמצאים ב StopWords) ושומר אותם ב- IndexObejct .

ניהול של DB :

הניהול של ה-DB בעזרת המערך של index file לניהול הקבצים, כל קובץ שאנו רוצים לעלות עובר את כל התהליך שתואר באנדקוס הקובץ.

StopWords :

התוכנה עושה שימוש ב StopWords , כלומר הוספנו מילים שמהם התוכנה תתעלם בעת החיפוש (מילים שמופיעות הרבה(common words) במסמכים, מילות קישור ( .

<http://xpo6.com/list-of-english-stop-words/>

סינון מילים StopWords

1. מעבר על המילים תוך ספירה והעברה לכתיב תקין (אותיות גדולות )
2. בדיקת כפילויות
3. רשימת המילים כמערך בIndexobject
4. בדיקת כפילויות במסמך indexObject

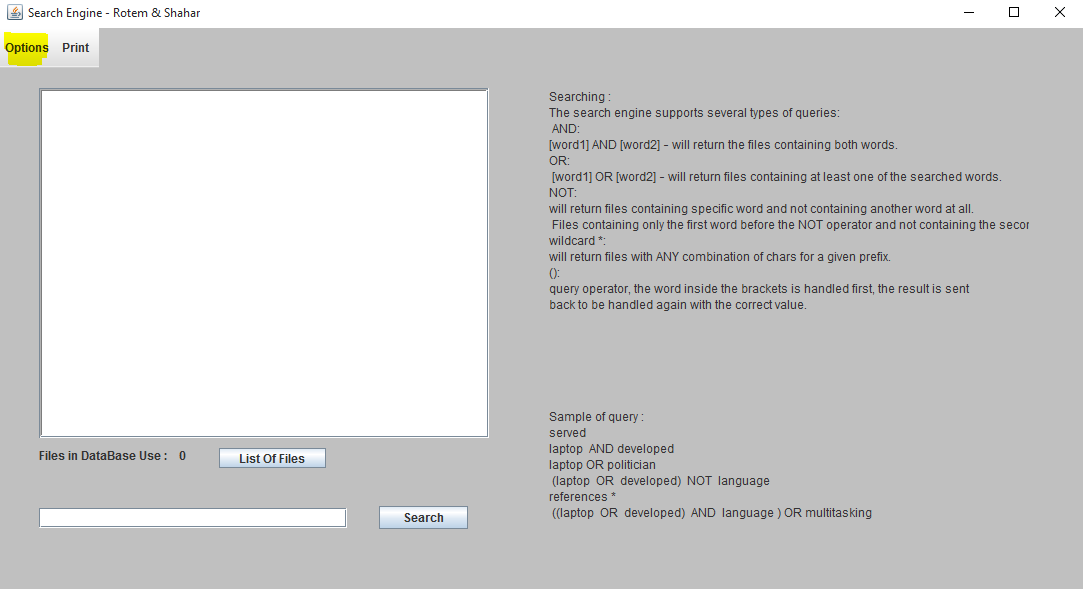
הוספת קובץ/מחיקה :

הוספת קובץ או מחיקה מתבצע אותו הדבר - מחיקה/הוספה של הקובץ מהרשימה של Files ולאחר מכאן עדכון

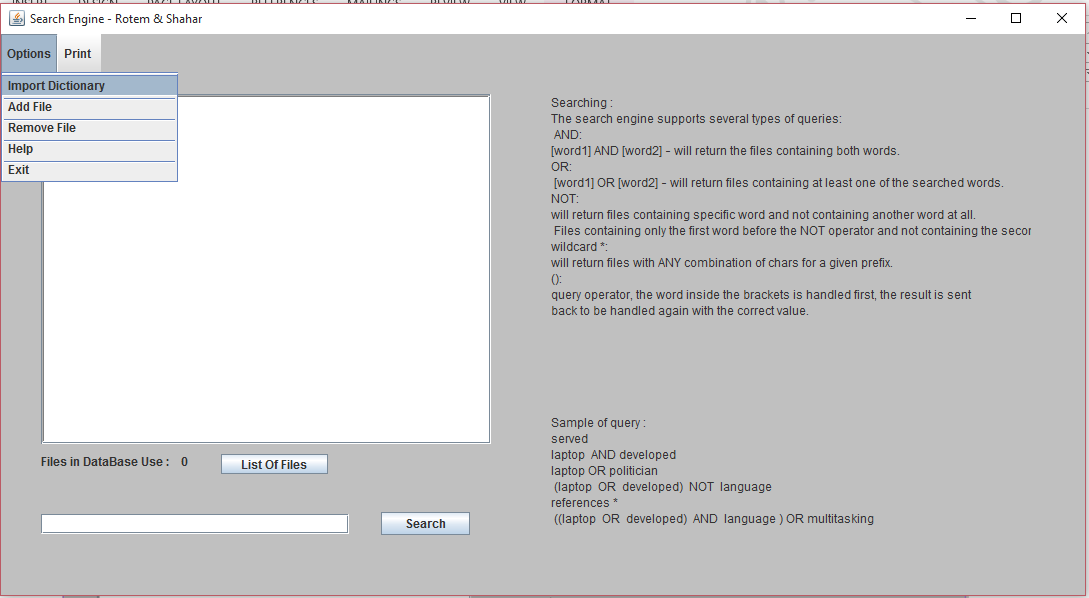
ה- IndexObejct.

שלבי העבודה בממשק :

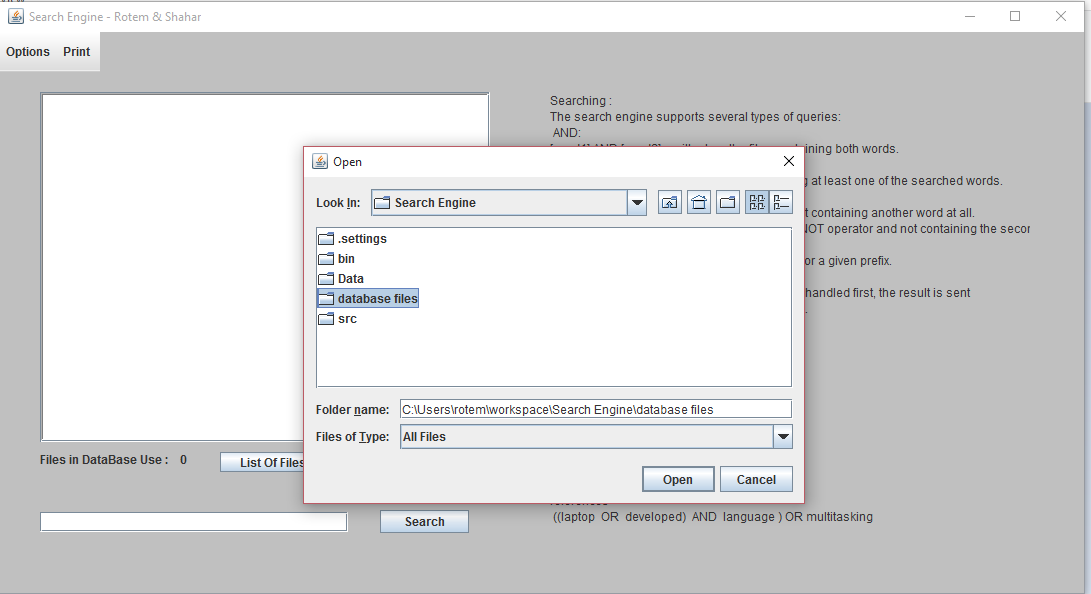
1. לוחצים על – option



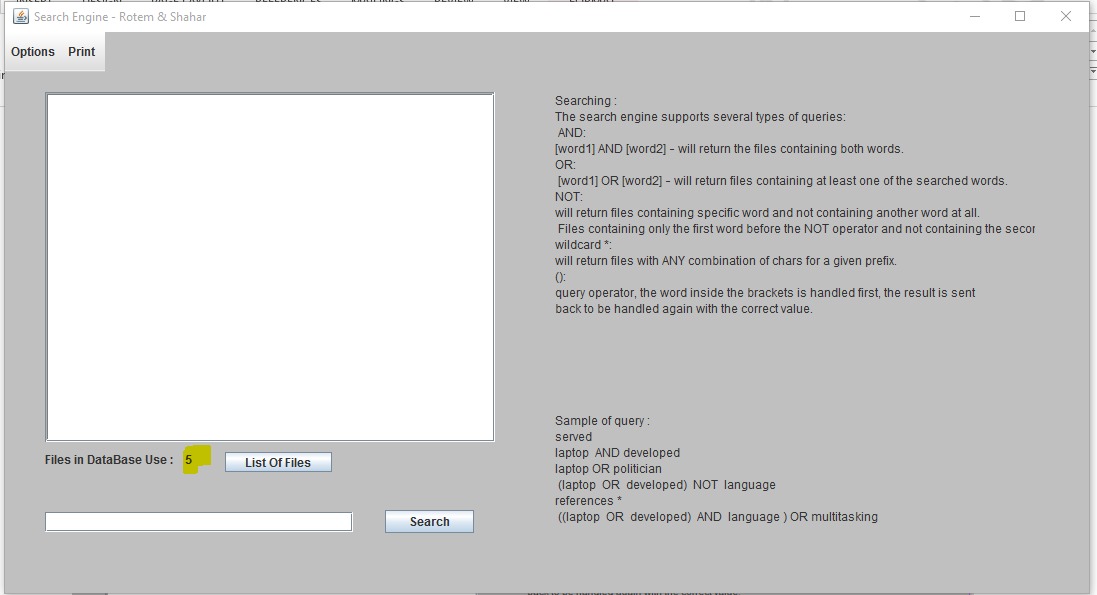
1. לוחצים על Import Dictionary:



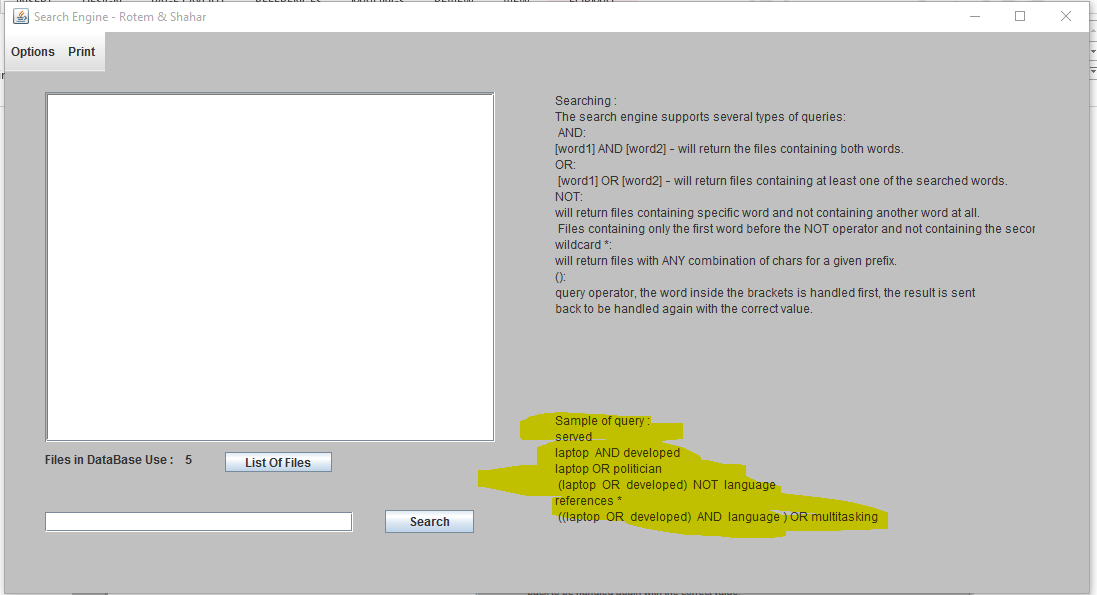
1. בוחרים את התקייה שהמערכת משתמשת בהם (קבצי המידע ):

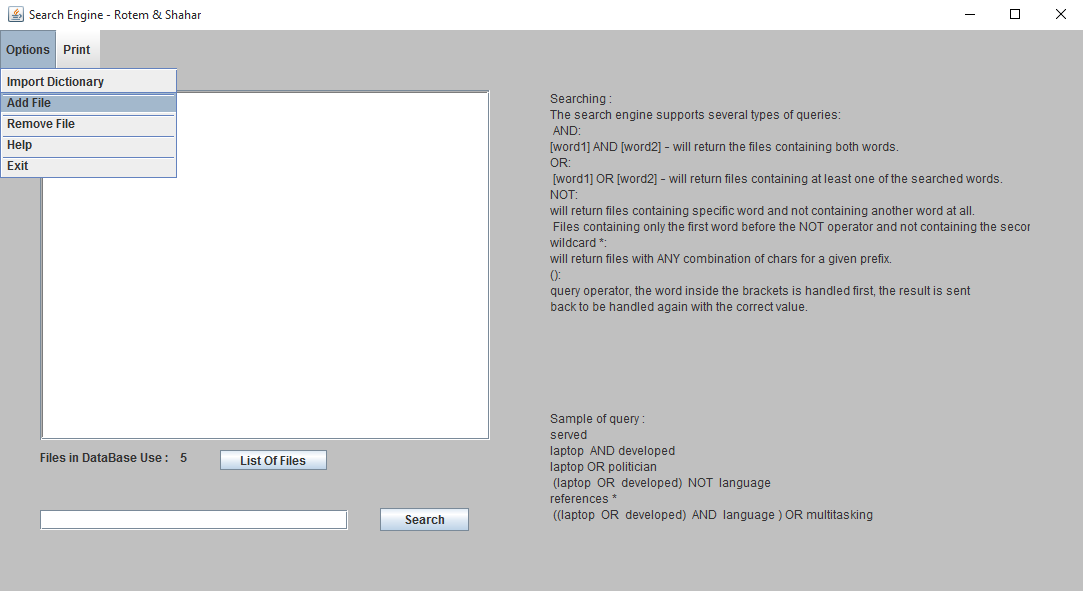


1. רואים שאכן נוספו הקבצים :



1. שמים בשורת חיפוש שאילתות כמו שיש בדוגמא או לבנות לפי ההסבר שיש בצד ימין :



* הוספת קבצים / מחיקת קבצים נעשית דרך ה - option : 
* Help – Read Me נעשית דרך ה - option :



* הדפסה נעשית דרך ה - print :

