שנקר הנדסה. עיצוב. אמנות.
הפקולטה להנדסה
המחלקה להנדסת תוכנה

פרויקט סיום בקורס 3503170: "אחזור מידע"

מטרתנו היא לבנות מערכת אחזור פשוטה שתיישם את עקרונות אחזור המידע שנלמדו בקורס. לפיכך אין המטלה הנוכחית שלמה ואין היא מתיימרת להקים בבת אחת מסד נתונים מלא בעל יכולות שליפה נרחבות.

בנית המסד והאינדקסים בשיטת Inverted file :

מודגש כי חובה שהמבנה עליו מתבסס מערכת האיחזור יהיה בשיטת הקובץ ההפוך (inverted file) כלומר בניית קובץ האינדקסים המצביע בסופו של דבר על המסמכים ולא חיפושים ישירים על המסמכים או שיטות אינדוקס אחרות.

לשם חזרה על התהליך מומלץ לעיין, בעבור החלק המרכזי של הפרויקט, בשקפים:
קבוצה 3: שקפים 20, 27
קבוצה 4: שקפים 3, 4, 5, 6, 22
קבוצה 5: שקפים 19, 20, 21, 23, 24, 25, 26

שקפים אלו משמשים כמודל כללי לבניית ספרית המסמכים והאינדקס.

וכן בשקפים נוספים הקשורים להרחבות ,המפורטים בסעיף " רעיונות להרחבה" בסוף המסמך.

אין חובה לאמץ בדיוק מבנה זה, המשמש רק כשלד כללי. ניתן לבחור במבנים אחרים שדנו בהם בקורס ובסדר אחר של עיבוד השונה מזה המתואר בהמשך.

המסמכים:
בכדי לפשט את העיבוד ניתן להשתמש במסמכים (כל שהם, הבחירה היא לחלוטין בידכם) בשפה האנגלית אם כי אתם מוזמנים להתמודד עם השפה העברית והבעיות המיוחדות שהיא מציגה.

מאיפה נשיג חומרים? דרך טובה היא מהאינטרנט. לדוגמה ניתן לשאול מסמכים ממאגר של ספרות. במרבית המקרים אלו הם חומרים שהמחברים שלהם נפטרו לפני 75 שנה ולפיכך פקעו כבר זכויות היוצרים. פרויקטים רבים כאלה קיימים בארצות רבות (המפורסם ביותר, ואחד הראשונים, הוא פרויקט "Guttenberg" לספרות האנגלית). לשם כך ניתן להתקשר לאתרים רבים באינטרנט. לדוגמא אחד מהם (אוסף של אוספים של שירה אנגלית):

http://www.poetry-archive.com

חומרים בעברית ניתן למצוא בפרויקט "בן-יהודה" שמטרתו להעלות לרשת את כל הספרות העברית שאין עליה כבר זכויות יוצרים. כתובת "בן-יהודה" היא:

http://www.benyehuda.org

כל הנאמר לעיל הוא בגדר דוגמה בלבד ומקור המידע שלכם יכול להיות כל חומר שהוא.

במסגרת הפרויקט אין אנו מיישמים את הרובוטים (Bots) אשר מתפזרים באתרים, מעתיקים את קבציהם למחשבי הניתוח ומשתמשים בקישורים שמצאו בהם בכדי "לצוד" אתרים חדשים. לפיכך אין צורך לקדד אותם ואת הקבצים (מסמכים) ניתן להכניס על ידי העתקה ל"ספרית המקור" (ראו להלן) המשמשת כמקור הקלט.

לפרקים, בנוסף לטקסט החופשי מופיעים, לפרקים, במסמך גם מספר שדות קבועים המכילים אינפורמציה מובנית, כגון: שם מחבר המסמך, נושא המסמך, תאריך חיבורו, תקציר וכו'. במידה ולמסמכים שלכם יש שדות מסוג זה אזי הם ניתנים להצגה למשתמש, או כברירת מחדל (כלומר תמיד) או לפי בקשתו למידע מסוים.

לפיכך – במקרה זה - יש להגדיר ראשונה את המבנה הזה ע"י בחירת השדות הקבועים ולהחליט באם אתם רוצים להציגם תמיד, להציגם לפי בקשה או להשאירם חבויים לשימוש המערכת.

שדה נוסף, הנבנה על ידכם, הוא מספרו הסידורי של המסמך (hhh) שמוענק לו בעת הקליטה והמשמש בעבודת בנית האינדקס כך שההצבעה ממושג נתון מתבצעת ל "מסמך מספר hhh".

בעת הצגת הפרויקט לשם בדיקתו על המערכת **להכיל כבר** כ 4-5 מסמכים שנקלטו ובנוסף צריכים להיות בספרית ה"מקור" (ראו להלן) **עוד** 2-3 מסמכים נוספים שניתן יהיה להוסיפם בעת הבדיקה.

כמו כן בעת הצגת הפרויקט עליכם לבוא מוכנים עם **רשימה של לפחות שלושה מושגים** הקיימים שכבר נקלטו בכדי שנוכל מיד לחפש אותם.

נקודות חשובות של תכנון וביצוע הנובעות מהדרישות הן:

א) עליכם ל"**חשוב בגדול**". מספר המסמכים הקטן שבעזרתו תבדק המערכת אין פרושו שניתן "לחתוך פינות" ולקודד מתוך ידיעה שניתן להתעלם מאלמנטים של חסכון בדיסק, חסכון בזכרון, ביצועים פושרים וכו'. המערכת חייבת להיות מתוכננת מתוך הנחה שהיא צריכה לתמוך במספר מסמכים גדול מאוד. לכן בעת שאתם דנים במערכת חיזרו ושאלו בשלבים השונים: "מה הם הקריטריונים והדרכים/שיטות הנכונות ליישום כאשר מספר המסמכים גדול מאוד?". במסגרת שאלה זאת יש לברר היטב מהי שיטת אחסון המידע (וזכרו שבסופו של דבר המידע מאוחסן על דיסקים) הנכונה. באם שיטות אחסון אלגנטיות הקלות להבנה וליישום "מנפחות" עוד יותר את תפוסת הדיסקים העשויים ממילא להיות גדולים מאד – אזי די ברור שלא ניתן להשתמש בהן.

ב) אין זה קורס בתכנות ולכן אין עליכם להוכיח שאתם יודעים לתכנת. זה חיב להיות ברור מאליו בשלב זה של לימודיכם. המטרה היא לבנות **מערכת עובדת**, ברת תחרות בהשוואה (תיאורטית) למערכות דומות אחרות שעשויות להיות מוצעות. כלומר הלקוח הפוטנציאלי חיב להתרשם ממנה ולהרגיש נוח איתה. לפיכך על המוצר להיות שלם מבחינת ההצעים שלו, הנוחות השימוש והדיוק במתן המידע שאותו חיפשו .

ג) המוצר הסופי היא מערכת **אופרטיבית** (ולא אב טיפוס המציג יכולות תכנותיות בלבד) שתבדק תוך כדי תשומת לב רבה למנשק המשתמש, הפרדה נכונה בין המשתמש המחפש את המידע ומנהל המערכת המכניס ומסלק מסמכים, נוחיות השימוש, יציבות, מהירות ויישום נכון של קבצי הדיסק של המערכת.

ד) חייבת להיות **הפרדה** ברורה בין מנשק המשתמש ומנשק המנהל. אין להשתמש במסך אחד בלבד שבו ניתן לבצע את כל הפעולות –אין לאפשר למשתמש להתערב בניהול. ההפרדה יכולה להיות באמצעות שתי תוכנות שונות ו/או דרישות לסיסמה בעבור מערכת הניהול שבלעדיה לא יוכל המשתמש לצפות ולשנות את המערכת.

קליטת המסמכים:

את המסמכים יש לקלוט בעבודת **אצווה**, כלומר קבוצה שלמה בעיבוד רציף ולא כל אחד ואחד בפני עצמו. (מדוע?)

שפת התכנות ניתנת לבחירה. למשל: PHP, NODE.JS , C# או Java כולן טובות למטרה זאת. כמו כן אתם רשאים להוסיף כל שפה שאתם חפצים בה לרשימה זאת ולקודד בה.

המסמך, נלקח מספרית ה"מקור", נסרק וכל המילים תישלפנה - על ידי תוכנית Parsing המניחה שהמפרידים בין מילה למילה הם: רווח, נקודה, פסיק, נקודה-פסיק , נקודתיים, גרשיים ולפרקים סוף שורה וכן תגי HTML רלוונטיים וכו'.

וכאמור, בנוסף יוענק לכל מסמך מספר סידורי כחלק מהשדות הקבועים.

ראו את צד שמאל בשקף מספר 4 בקבוצה 4.

ניתן לבנות את התוכנית בשלוש קטגוריות שונות:

א) מערכת שאיננה קשורה כלל לאינטרנט.
מערכת אחזור פנימית, בה כל המידע, כלומר תוכנם המלא של המסמכים בליווי השדות הקבועים, מערך האינדקסים, קבצי עזר ותוכניות הבניה והחיפוש נמצאים על גבי מחשב מקומי ללא כל קשר לאינטרנט.
את המידע (מסמכים) מאחסנים כקבצים נפרדים בספרית ה"החסנה" שבה כל מסמך הוא קובץ. את מערכת האינדקסים ושדות אינפורמטיביים אחרים ניתן לאחסן במסד נתונים טבלאי. במידה ואינכם מעונינים להשתמש במסד נתונים למטרה זאת, יהיה עליכם ליצור אינדקס משלכם על ידי שתישמו אותו כקובץ שבו יהיה עליכם לטפל: לבנותו, למיינו, להוסיף איברים ולגרוע איברים.
יש להקפיד כי תהיה **הפרדה** בין ספרית ה"מקור" (קלט) וספרית ה"החסנה" (התיקייה המתמדת בה שוכנים המסמכים) . ספרית המקור היא ספריה זמנית, קצרת טווח, המכילה את הקבצים שהועלו על ידכם למערכת באנלוגיה למכניזם הקיים במנועי חיפוש השולחים "bots/crawlers" לאתרים ברחבי הרשת ואוספים משם מסמכים. כלומר היא מחקה את קבלת הקלט ממקורות שונים. ממנה יש להעביר מסמכים לספרית ה"החסנה", הקבועה, וזאת תוך כדי בנית האינדקסים.
ספרית ה"מקור" לא תשמש גם כספרית ה"החסנה". כלומר אחרי העברת מסמך למערכת ניתן יהיה למחוק אותו מספרית ה"מקור" (בכדי למנוע כפילויות בתפוסת משאבים) מבלי שהמסמך וההצבעות עליו ימחקו מהמערכת.

ב) מערכת אחזור מבוססת דפדפן עם קבצים מקומיים.
מערכת זאת דומה למערכת שתוארה לעיל פרט לעובדה שהיא עובדת בסביבת האינטרנט באמצעות שרת מקומי (שניתן ליישמו על ידי שימוש למשל ב XAMP/ APACHE או מערכת דומה אחרת). הקבצים הם עדיין מקומיים כאשר כל יתר המפרטים הם כפי שתואר לעיל.

ג) מערכת אחזור אינטרנטית (בדומה ל Google) בה מערכת האינדקסים והשדות הקבועים (בתוספת תקציר) ימצאו על גבי מסד נתונים טבלאי במחשב המקומי (או בקובץ אינדקס המתוחזק על ידכם), אולם המסמכים עצמם ימצאו עדיין ברשת האינטרנט ויש עליהם הצבעה מהמערכת. כלומר במקום להצביע על מסמך מקומי

באמצעות מספרו הסידורי כפי שנעשה בשתי המערכות הקודמות יש להצביע באמצעות URL על המסמך המקורי. כמובן שגם במקרה זה צריך להביא פעם אחת את המסמך למחשב לשם סריקתו ובנית האינדקסים – אולם אין לשומרו מקומית.

לפיכך מנשק המשתמש יכול להיות מיושם על ידי תוכנה ויזואלית או HTML.

אם כך, המסמכים הראשונים מוזנים למערכת מספרית ה"מקור". המסמכים מועברים מספרית ה"מקור" לספרית ה"החסנה" תוך כדי כך שהם מקבלים מספר זיהוי. כל המילים נשלפות לבניית טבלת אינדקסים זמנית (הטבלה השמאלית בשקף מספר 5 בקבוצה 4). התהליך חוזר על עצמו לכל מסמך ומסמך באצווה כאשר מילות המפתח שלו מתווספות בהמשך הטבלה הקודמת. אין עדיין מיון ואין חקירת מילים כפולות.

לאחר קליטת כל המסמכים, ממיינים את טבלת המילים (ראו טבלה מרכזית בשקף מספר 5 בקבוצה 4). חשוב מאוד שקובץ האינדקס יהיה **ממוין** לפי שדה המילים מכיוון שזה הקובץ עליו מחפשים ולכן סדר אלפא-בתי ממוין הוא גורם חשוב בביצועים.

בשלב הבא ניתן לבנות את Posting File (טבלה מרכזית בשקף מספר 3 בקבוצה 4 וכן בשקף מספר 6 בקבוצה 4). **הכפילויות מסולקות** וליד כל מילה רושמים את מספר המסמכים בהם היא הופיעה. הטבלה בנויה כך שאם יש מופעים רבים של מילה במסמך, ערך השדה השני (האופציונלי) הוא מספר המופעים בטבלה, הערך Link ב Index File מצביע על הראשון שבהם, והערך Hit מוסר כמה מהם יש.

ה Posting File יכול להיות מיושם בדרכים שונות: הכללתו במסד נתונים טבלאי, רשימה מקושרת, טבלה וכו'. יש חשיבות לדרך שתבחרו משום שהיא משפיעה גם על העדכון: ראו דיון בכך בסעיף הבא.

כדוגמה לשדות אינפורמטיביים: ניתן גם להגדיל את הרשומה ב Index File כך שתכיל גם את מספר המופעים המופיע בטבלה הימנית בשקף מספר 5 בקבוצה 4 ובנגררות בשקף מספר 6 בקבוצה 4.

עדכון:
בנינו את מסד הנתונים, אולם עתה מופיעים מסמכים נוספים. איך לטפל בהם?

כמו קודם: המסמכים החדשים (הנמצאים בספרית ה"מקור") ימוספרו, יוכנסו לספרית ה"החסנה" והמילים תישלפנה. אם זאת מילה שכבר קיימת יש להגדיל את מספר ה- Hits ולעדכן את ההצבעות המתאימות. אם תבדקו מה יש לעשות בתוספות אלו תבחינו עד מהרה שרצוי, כמובן, לשנות את המבנה של הטבלה המרכזית בשקף מספר 3 בקבוצה 4 לשימוש במצביעים או לשימוש במערכים דינמיים במסדי הנתונים.

אם אתם משתמשים במסד נתונים לאחסון המילים אזי התהליך פשוט ביותר משום שאין אתם צריכים לנהל את קובץ האינדקסים. הוסיפו את המילים למסד הנתונים. תכונה של שדה המילים היא שהוא צריך להיות מוגדר כ<u>ממוין</u> משום שאז העבודה על המפתח הראשי תהיה בזמנים משופרים.
הקדישו מחשבה איך אתם מאחסנים אינדקסים והצבעות מהם למסמכים השונים: האם תבחרו בשיטה שבה כל הצבעה לכל מסמך היא שורה בפני עצמה? והמקרה זה יש כפילויות במושגים שבאינדקס (דמוי הרשומה הימנית בשקף מספר 5 בקבוצה 4) – העשויות להגדיל בצורה משמעותית את מספר האיברים בשדה זה משום שמילים נפוצות יופיעו פעמים רבות מאוד. זה פתרון גרוע משום שמספר השורות בקובץ כזה יהיה מספר המילים השונות כפולות כל פעם בהתאמה במספר המסמכים בהם היא מופיעה – ערך העלול להיות אסטרונומי במערכת בעלת מסמכים רבים. או תבחרו בשיטה שכל מושג מופיע

רק פעם אחת ויחידה והוא מצביע בשירשור או במערך על המסמכים (דמוי שקף מספר 6 בקבוצה 4). אין ספק שלפתרון השני, אם כי הוא מסובך (מעט) יותר ליישום, יתרונות רבים.

אם אתם בונים את האינדקס בעצמכם אזי טבלת ה Index File תצביע על ה # Doc (מספר סודר של המסמך) שבטבלת ה Posting File. רצוי לא להשתמש במבנה קשיח של טבלה, אלא בהצבעה משורשרת של מסמכים (לפיכך לגבי המילה הראשונה תהיה הצבעה ממסמך 1 ל- 2 ל 7 ול- 8). כאשר כל רשומה מצביעה (בשדה ה Link) על המסמכים עצמם. כלומר הקובץ ידמה למבנה של שקף מספר 6 בקבוצה 4 .

אם זאת היא מילה חדשה, ואתם עובדים בקובץ אינדקס משלכם, יש להכניסה במקומה הנכון (merge) או למיין את הרשימה לאחר הכנסת המילים החדשות.

ביטול מסמכים:

כיצד מבטלים מסמך? עוברים על גבי ה Posting File וכל פעם שפוגשים את המסמך שברצוננו לבטל, מסמנים במשתנה שהוגדר ב Posting File והמשמש למטרה זאת, סימן ביטול. ביטול פרושו שבחיפושים עתידיים מסמך זה יוסתר ולא יוצג גם אם הוא עונה לדרישות החיפוש. סילוק המסמכים עצמם ושיחזור הטבלה נעשים רק אחת למועד קבוע, משום שפעולת המחיקה והצמצום היא ארוכה: יש לסלק את המסמך, לצמצם את השטח, לסלק את ההצבעות שבוטלו, ולתקן את מספר ה Hits בטבלת האינדקסים. (מה קורה שהמסמך האחרון הקשור למילה נתונה נעלם?)

בפרויקט **אין** אתם נדרשים ליישם את הסילוק והצמצום עצמו ולכן אין צורך ליישם אותם - אולם חובה לאפשר ביטול (הסתרת) מסמכים וביטול הביטול (הצגה מחודשת).

יש לבנות רשימת **Stop List** – (שקף 22 בקבוצה 4) כלומר רשימת מילים כה נפוצות וכה חסרות יחוד שאין טעם לחפש עליהן. אין צורך לבנות רשימה ענפה; מספיק לכלול בה רק מספר מועט של מילים לשם הדגמה בעת החיפוש: ראו להלן.

שני סוגי המשתמשים:

יש להפריד במערכת בין שני סוגים של משתמשים. מנהל המערכת היכול להוסיף ולהסיר מסמכים ולבצע את כל הפעולות החוקיות במערכת וניהולה ובין המשתמש היכול רק לחפש מסמכים ולצפות בהם אך איננו יכול לשנות דבר במערכת. כלומר יש ליצור את המכניזם הנכון המונע מהמשתמש לבצע פעולות שהן בתחומו של מנהל המערכת.

השאילתות:

השאילתות תהינה בוליאניות עם האופרטורים And ,Or ו-Not **ולפחות** רמה אחת של סוגריים. **על כל שלושת האופרטורים הבוליאניים להיתמך**. ניתן להשתמש כהנחיה בשקפים מספר 20 ו 26 בקבוצה 5. שליפת and היא על ידי שליפת מספרי המסמך ומחיקת כל מקרה שאין בו כפילות לעומת שליפת or שהיא על יד שליפה ואיחוד תוך כדי ביטול הכפילויות. פעולת ה not יותר מורכבת והיא מתוארת בשקף מספר 26 בקבוצה 5.

אין דרישה לבנות מידע על מיקומה של כל מילה במסמך ולכן אין צורך ליישם את היחס "קרבה" אולם הוא יכול להיות חלק מההרחבות (ראו להלן במבנה הציון).

מילים הנמצאים ב-Stop List אינן משמשות בדרך כלל לחיפוש, אולם אם המילה או המילים נמצאות כמחרוזת בין גרשיים כפולים יש לחפש את המחרוזת. לפיכך יש לאנדקס את המילים הנמצאות ב Stop List אולם אין לחפשן בדרך כלל **אלא** אם הן נמצאות בין גרשיים כפולים – שקף 22 בקבוצה 4.

במידה ועובדים בטקסטים בשפות שיש בהן אותיות ראשיות ואותיות קטנות חובה לבצע נורמליזציה לאותיות גדולות או קטנות. כלומר resume, rEsUmE ,RESUME ו- Resume נחשבים כזהים (כלומר בשאילתה המשתמש רשאי לכתוב אותם כרצונו) ויופיעו כאותה מילה באינדקס. במסמכים הם עדין חייבים להופיע בצורתן המקורית אך לפני הכנסתם לאינדקס (ובעת השאילתה) יש להפכם לצורה אחידה (למשל אותיות קטנות). בכדי לא להסתבך עם נורמליזציה בשפות שיש בהם סימנים נוספים (כגון à בצרפתית או ß בגרמנית – אל תביאו מסמכים משפות אלו).

חובה לשמור את המבנה (הפורמט) הנכון של המסמכים בכדי להציגם בצורה נאותה – דבר זה בולט במיוחד במסמכים שהם שירה.

את המילים שהיו ארגומנטי שאילתת החיפוש יש להדגיש במסמך מוצג על ידי סימון בולט (צבע? הדגשה? קו תחתון?)בכל מקום שהם מופיעים: טקסט חופשי ושדות קבועים. ניתן למצוא אותן על ידי חיפוש בשיטה הנאיבית או KMP (שקפים 9 עד 18 בקבוצה 4) בעת התצוגה של המסמך או על ידי סריקה מראש בעת ה parsing ושמירת הקואורדינטות של כל המילים ב Posting File (שקף 27 בקבוצה 3).

צפייה בתשובות:

בשלב הראשון של הצגת התשובות לא מציגים את המסמך עצמו אלא רק את המידע הנמצא בשדות הקבועים כולל תקציר המסמך (חשוב במיוחד אם היישום הוא מהקטגוריה השלישית, כלומר הפנייה למסמך המלא השוכן עדיין באינטרנט).

אם לא יצרתם תקציר, תוצגנה בתשובות שלושת השורות הראשונות של כל מסמך רלוונטי כתחליף לתקציר.

עתה יוכל השואל לציין, על ידי הקשה על שם המסמך, איזה מסמכים הוא מבקש ואלה יובאו במלואם.

יש לאפשר את הדפסת המסמכים המובאים.

מנשקי המשתמש ומנהל המערכת:

יש לקדד מנשק שבו מופיע שדה שאילתה שאותו ממלא המשתמש בלווי האופרטורים הבוליאניים. יש חשיבות ל UI: חשוב לוודא שהשאילתה איננה תלוית מבנה תחבירי נוקשה. כלומר: אין לחיב רווחים (או לחייב שלא יהיו רווחים) בין סוגרים לבין מילות החיפוש או האופרנדים או לדרוש שיהיה אך ורק רווח אחד ולא מספר רווחים בין אופרנדים ומילות חיפוש. רצוי מאוד לאפשר להפעיל את החיפוש לא רק על ידי הקשה בעכבר על כפתור "חפש" אלא גם בעזרת הקשה על קליד ה Enter.

יש להציג את תקצירי המסמכים בצורה ברורה שתקל על המשתמש לבקש את המסמך המלא ולהמשיך ולראות מסמכים נוספים.

חשיבות רבה יש למנשק נכון המאפשר גישה נוחה ואינטואיטיבית למשתמש וכמו כן מטפל נכון באינפורמציה המוצגת (כגון למשל **לא** להציג את רשימת כל המסמכים הקיימים במערכת על מסך החיפוש – חישבו איך הייתה המערכת מתנהגת אם במקום 4-5 מסמכים יהיו בה כמה מיליונים?)

תוכנת המשתמשים וכן תוכנות המערכת (המיועדות למנהל המערכת) כגון קליטת המסמכים הראשונית, עדכון מסמכים וסילוק מסמכים צריכות להיות נוחות לשימוש ובעלות מנשק משתמש (בדומה ליישום back office) **ולא** להיות פעולות הפועלות ישירות על גבי מסד נתונים עצמו או בעזרת מסכים של מנהל הקבצים. בדומה: חזרה למסכים קודמים צריכה

להיות על ידי בחירת אופציות בתוכנית ולא על ידי שימוש (למשל) בחיצים של הדפדפן.
כלומר יש לקדד מנשק מסודר המנהל את כל הפעולות.

הגשה:
בדיקת הפרויקט תהיה הצגתו ביחד על ידי **כל משתתפיו** (שיש להודיע מראש מי הם על ידי שליחת שמותיהם בדואר אלקטרוני למרצה הקורס מיד לאחר שחרור הפרויקט לעבודה) על גבי אחד מהמחשבים במעבדות המחשבים של שנקר או על גבי מחשב שלכם. התוכנות לא תותקנה על מחשב המרצה.

המערכת תכלול מסך עזר המסביר, בקצרה, למשתמש כיצד לבצע את השאילתות וכן קובץ "קרא אותי" (help) המסביר למנהל המערכת כיצד לבצע את הקליטה, העדכון והביטול של מסמכים.

קובץ זה בליווי הגדרת המבנה של המסמכים, מבנה מסד הנתונים ומבנה הטבלאות האחרות, מבנה התוכנית וחומר הסבר נוסף יהוו את המדריך למשתמש ולמנהל המערכת שאותו יש להגיש מודפס בעת הצגת הפרויקט.

כנאמר לעיל, יש לבוא מוכנים, בעת הצגת המערכת, עם מספר מילים המופיעות במסמכים שישמשו לשאילתות.

את קוד התוכנית אין להדפיס ואין צורך להגיש אותו אך הוא חיב להיות זמין במחשב התצוגה בכדי לחקור, בעת ההצגה, חלק מהשגרות שלו.

מבנה הציון:
ביצוע הפרויקט ככתבו וכלשונו (כלומר ממלא את כל הדרישות בצורה שהיא "בסדר") מעניק ציון בסיס של 80. פרמטרים שילקחו בחשבון ויכולים להשפיע על הציון הם: יציבות המערכת, התמצאות קלה, נוחות השימוש, מהירות הביצוע בקליטת מסמכים ובחיפושם, קיום הפעולות אך ורק בתוכני המנשק ולא מחוצה לו, קידוד אלגנטי, יעילות, מסמכים נאותים וכן אופציות נוספות (ראו להלן). כל אלה עשויים להגדיל או להקטין (במקרים של יישום/התנהגות לא מספק) בהתאמה את הציון.

לסיכום:
על הרכיבים והמדדים הבאים להימצא במערכת (הפרטים נמצאים בלעיל במסמך):

- איסוף מסמכים
- הגדרת מבנה המסמך
- קליטת מסמכים
- סריקת מסמכים
- תהליך ניתוח ואחסון המסמכים בספרית "החסנה" (השונה מספרית ה"מקור")
- בנית האינדקס, כלומר: אחסון כל המצביעים ושדות אינפורמטיביים במסד נתונים או בקובץ האינדקס שבניתם
- עידכון, בצורה נאותה, את האינדקס כתוצאה מהוספת מסמכים חדשים
- שימוש נאות ב-Stop List
- הסתרת מסמכים
- בנית השאילתות עם האופרטורים
  - And
  - Or
  - Not
- תמיכה בלפחות רמה אחת של סוגריים
- הפעלת השאילתות

- נוחיות השימוש במערכת
- הצגת הממצאים והתקצירים בליווי הדגשת מילות החיפוש
- הבאת המסמך המלא בליווי הדגשת מילות החיפוש (אם ניתן)
- פונקציה להדפסת המסמך
- מסך הסברים (מסוג "help" המבאר בקצרה את השימוש)
- חוברת הפרויקט (כפי שהובהר לעיל)
- מדדים נוספים הם:
  - יציבות המערכת
  - מהירות הביצוע בעת קליטת המסמכים ובעת השאילתות
  - יכולת עבודה עם מספר גדול (תיאורטי) של מסמכים

רעיונות להרחבה (אופציות לשיפור הציון):

להצעות יש דרגות קושי שונות: מכוכבית אחת (*) ועד שלוש כוכביות לשם השגת מקסימום הבונוס (20%) יש לצבור 4 כוכביות.

- * בנית שאילתות עם יותר מרמה אחת של סוגריים: למשל :
  ((a AND b) OR c) AND d [זאת דוגמה לשתי רמות]
- הוספת אופרנדים לשאילתה כגון:
  - ** Near (הגדרת מרחק בין מילים כתנאי להבאת המסמך) –שקפים 6 ו-7 בקבוצה 5.
  - * הבאת מסמך על ידי הקלה של תנאי ה AND, כגון: הבא מסמך אם יש בו לפחות n מתוך m המושגים המבוקשים – שקף 22 בקבוצה 5.
- * מציאת ביטויים (שרשרת מילים המוכלות בתוך מרכאות כפולות) במסמכים בניגוד למילים בודדות על ידי התייחסות לשרשרת כמהות אחת – שקף 26 בקבוצה 7.
- * חיפוש טקסט לא רק בדיוק כפי שהוא מופיע אלא גם בחלקי מחרוזות. כלומר שימוש באופציית ה- Wild Character/Joker (לדוגמה: *Car יביא גם את Car וגם את Cart ו- Cardamom) – שקף 29 בקבוצה 7.
- *** חלוקת המסמכים לאשכולות, ובעת השאילתה המביאה אשכול נתון, להביא גם מסמכים אחרים דומים – שקפים 42 עד 59 בקבוצה 7
- *** Ranking הגדרת משקל רלוונטיות של המסמכים (לפי כל אחת מהשיטות בהן דנו – שקפים בנושאים שונים של משקלות) לשם מיון התשובות והצגתן לפי סדר משקל יורד – שקפים 10 עד 14 בקבוצה 7
- * הכלת תמונות / מוזיקה כחלק מהמידע האגור או המוצבע
- ** אחסון יעיל של האינדקסים על ידי שימוש ב Stemming – שקפים 7 ו-8 בקבוצה 4 ושקפים 27 ו-28 בקבוצה 7
- ** שימוש במילים נרדפות (Synonyms) לשם שיפור גמישות החיפוש – שקפים 9 עד 12 בקבוצה 5.
- ** שימוש ב Soundex (כאופציה נוספת ולא כברירת מחדל) לשם מציאת מושגים בעלי דמיון פונטי – שקף 19 בקבוצה 4

בהצלחה!