שנקר- בי"ס גבוה להנדסה ולעיצוב
הפקולטה להנדסה
המחלקה להנדסת תוכנה

פרויקט סיום בקורס "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 נותנים מלא בעל יכולות שליפה נרחבות.

בנית המסד והאינדקסים בשיטת Inverted file :
לשם חזרה על התהליך עיינו בשרטוטים 39, 41, 42, 43, 44, 92 ו-96 בדפי החלוקה כמודל לבניית המסד והאינדקסים. אין חובה לאמץ בדיוק מבנה זה, המשמש רק כשלד כללי וניתן לבחור במבנים אחרים שדנו בהם בקורס.

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.

המסמכים:
בכדי לפשט את העיבוד מבחינה מורפולוגית, ניתן להשתמש במסמכים בשפה האנגלית אם כי אתם מוזמנים להתמודד עם השפה העברית והבעיות המיוחדות שהיא מציגה.

מאיפה נשיג חומרים? דרך טובה היא מהאינטרנט. לדוגמה ניתן לבנות מאגר של שירה אנגלית ולשם כך ניתן להתקשר לאתר:

http://etext.lib.virginia.edu/english.html

המכיל את כל השירה האנגלית מתחילתה ועד לפני 75 שנה (בעיות של זכויות יוצרים!).

חומרים בעברית ניתן למצוא בפרויקט "בן-יהודה" שמטרתו להעלות לרשת את כל הספרות העברית שאין עליה כבר זכויות יוצרים. אלה הם חומרים שהמחברים שלהם נפטרו לפני 75 שנה. פרויקטים כאלה קיימים בארצות רבות (המפורסם ביותר, ואחד הראשונים, הוא פרויקט "גוטנברג" לספרות האנגלית).כתובת "בן-יהודה" היא:

http://www.benyehuda.org

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

מקובל שבמסמך, בנוסף לטקסט החופשי מופיעים גם מספר שדות קבועים המכילים אינפורמציה מובנית, כגון: שם מחבר המסמך, נושא המסמך, תאריך חיבורו, תקציר וכו'. שדות אלו מוצגים למשתמש, או כברירת מחדל (כלומר תמיד) או לפי בקשתו למידע מסוים. שדה מובנה נוסף הוא מספרו הסידורי של המסמך שמוענק לו בעת הקליטה והמשמש בעבודת האינדוקס.

לפיכך 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.

בעת הצגת הפרויקט לשם בדיקתו על המערכת להכיל כ 4-5 מסמכים ובנוסף צריכים להיות בספרית ה"מקור" (ראו להלן) עוד 2-3 מסמכים נוספים שניתן יהיה להוסיפם. כמו כן בעת הצגת הפרויקט עליכם לבוא מוכנים עם רשימה של לפחות שלושה מושגים הקיימים במסמכים בכדי שנוכל מיד לחפש אותם.

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

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

המסמך יסרק וכל המילים תישלפנה - על ידי תוכנית Parsing פשוטה המניחה שהמפרידים בין מילה למילה הם רווח, נקודה, פסיק, נקודה-פסיק ונקודתיים, גרשיים ולפרקים סוף שורה וכן תגי HTML רלוונטיים וכו'.
בנוסף יוענק לכל מסמך מספר סידורי כחלק מהשדות הקבועים.
ראו את צד שמאל בשרטוט מספר 42.

ניתן לבנות את התוכנית בשתי קטגוריות שונות:

א) מערכת אחזור פנימית, בה כל המידע, כלומר תוכנם המלא של המסמכים בליווי השדות הקבועים, מערך האינדקסים וקבצי עזר נמצאים על גבי מחשב מקומי.
במערכת זאת אין 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". ספרית המקור היא זאת המכילה את הקבצים שהועלו על ידכם למערכת והיא מחקה קבלת קלט ממקורות שונים. ממנה יש להעביר (רצוי באצווה או אם נדרש לפרקים בקובץ בודד) מסמכים לספרית ההחסנה וזאת תוך כדי בנית האינדקסים. ספרית המקור לא תשמש גם כספרית ההחסנה עצמה. כלומר אחרי העברת המסמך למערכת ניתן יהיה למחוק אותו מספרית המקור מבלי שהמסמך וההצבעות עליו ימחקו מהמערכת.

ב) מערכת אחזור אינטרנטית (בדומה ל Google) בה מערכת האינדקסים, השדות הקבועים (כולל תקציר) וקבצי עזר ימצאו על גבי מסד נתונים טבלאי במחשב המקומי, כאשר המסמכים עצמם נמצאים עדיין ברשת האינטרנט ויש עליהם הצבעה מהמערכת. כלומר במקום להצביע על המסמך באמצעות מספרו הסידורי כפי שנעשה במערכת הפנימית יש להצביע באמצעות URL על המסמך המקורי. כמובן שגם במקרה זה צריך להביא פעם אחת את המסמך למחשב לשם סריקתו ובנית האינדקסים – אולם אין צורך לשומרו מקומית.

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

אם כך, המסמך הראשון נקלט, מועבר מספרית המקור לספרית ההחסנה תוך כדי הענקת זיהוי. כל המילים (אם כי יש לבנות Stop List – ראו להלן בסעיף "שאילתות") נשלפות לבניית טבלת אינדקסים זמנית (הטבלה השמאלית בשרטוט 43). עתה נקלט המסמך השני, התהליך חוזר על עצמו כאשר מילות המפתח שלו מתווספות בהמשך הטבלה הקודמת. אין עדיין מיון ואין חקירת מילים כפולות (ניתן לבצע אך לא כדאי- למה?).

לאחר קליטת כל המסמכים, ממיניי את הטבלה (ראו טבלה מרכזית בשרטוט 43). חשוב מאוד שקובץ האינדקס יהיה ממוין מכיוון שזה הקובץ עליו מחפשים ולכן סדר הוא גורם חשוב בביצועים.

בשלב הבא ניתן לבנות את Posting File (טבלה מרכזית בשרטוט 41 וכן בשרטוט 44). 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 הוא מספר המופעים בטבלה, הערך Link ב Index File מצביע על הראשון שבהם, והערך Hit מוסר כמה מהם יש.

(כדוגמה לשדות אינפורמטיביים: ניתן גם להגדיל את הרשומה ב Index File כך שתכיל גם את מספר המופעים המופיע בטבלה הימנית בשרטוט מספר 43 ובנגררות בשרטוט מספר 44).

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

כמו קודם: המסמך החדש ימוספר, יוכנס לספרית ההחסנה, המילים 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 בשרטוט *מספר* 41 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. אופיו של שדה המילים להיות מוגדר כממוין משום שאז העבודה על המפתח הראשי תהיה בזמנים משופרים. הקדישו מחשבה איך אתם מאחסנים אינדקסים והצבעות מהם למסמכים השונים. האם אתם בוחרים בשיטה שבה כל הצבעה לכל מסמך היא שורה בפני עצמה ולכן יש כפילויות במושגים שבאינדקס (דמוי הרשומה הימנית בשרטוט 43) – העשויים להגדיל בצורה משמעותית את מספר האיברים בשדה זה משום שמילים נפוצות יופיעו פעמים רבות מאוד, או שכל מושג מופיע רק פעם אחת והוא מצביע בשירשור או בטבלה על המסמכים (דמוי שרטוט 44). אין ספק שהפתרון השני הוא הנכון. אם תבחרו בראשון יהיה עליכם לנמקו בעת הצגת המערכת.

אם אתם בונים את האינדקס בעצמכם אזי טבלת ה Index File תצביע על ה # Doc (מספר סודר של המסמך) שבטבלת ה Posting File . רצוי ששם לא יופיע מבנה קשיח של טבלה, אלא הצבעה משורשרת של מסמכים (לפיכך לגבי המילה הראשונה תהיה הצבעה ממסמך 1 ל- 2 ל 7 ל ול- 8). כאשר כל רשומה מצביעה (בשדה ה Link) על המסמכים עצמם. כלומר הקובץ ידמה למבנה של שרטוט מספר 44 .

אם זאת היא מילה חדשה, ואתם עובדים בקובץ אינדקס משלכם, יש להכניסה במקומה הנכון. אולם מכיוון וזה ידרוש תזוזה של חלקי טבלה בכדי לפנות מקום רצוי בשלב זה להוסיפה בסוף הרשימה ורק לאחר כל העדכון למיין את הטבלה לפי שדה המילים.

.**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 אים צורך ליישם אותם - אולם חובה לאפשר ביטול (הסתרת) מסמכים.

השאילתות:
השאילתות תהינה בוליאניות עם האופרטורים Or, And ו-Not ולפחות רמה אחת של סוגריים. על כל שלושת האופרטורים להיתמך. ניתן להשתמש כהנחיה בשרטוטים מספר 92 ו 96. 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 בעיקר בשרטוט 96.

אין דרישה לבנות מידע על מיקומה של כל מילה במסמך ולכן אין צורך ליישם את היחס "מרחק" 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 אלא אם הן נמצאות בין גרשיים כפולים.

במידה ועובדים בטקסטים בשפות שיש בהן אותיות ראשיות ואותיות קטנות חובה לבצע נורמליזציה לאותיות גדולות וקטנות. כלומר RESUME, resume ו- Resume יהיו זהים ויופיעו כאותה מילה באינדקס. במסמכים הם עדין חייבים להופיע בצורתן המקורית אך לפני הכנסתם לאינדקס יש להפכם לצורה אחידה (למשל אותיות קטנות). בכדי לא להסתבך עם נורמליזציה בשפות שיש בהם סימנים נוספים (כגון à בצרפתית או ű בגרמנית – אל תשביאו מסמכים משפות אלו).

יש לשמור את המבנה (הפורמט) הנכון של המסמכים – דבר זה בולט במיוחד במסמכים שהם שירה.

בעת הצגת התשובות לשאילתות יש להדגיש בתקציר, אם בחרתם לפתח מערכת המצביעה על מסמכים מרוחקים (הקטגוריה השניה) וכן בכל התקצירים והמסמכים אם המערכת היא פנימיית (הקטגוריה הראשונה) את המילים שהיו ארגומנטי החיפוש על ידי סימון בולט בכל מקום שהם מופיעים: טקסט חופשי ושדות קבועים (צבע? הדגשה? קו תחתון?). ניתן למצוא אותן על ידי חיפוש בשיטה הנאיבית או KMP או על ידי שמירת מיקומן ב Posting File.

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 השניה, כלומר הפנייה למסמך המלא באינטרנט).

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

עתה יוכל השואל לציין איזה מסמכים הוא מבקש ואלה יובאו במלואם.

יש לאפשר את הדפסת המסמכים המעניינים.

מנשק המשתמש:
יש לקדד מנשק שבו מופיע שדה שאילתה שאותו ממלא המשתמש בלווי האופרטורים הבוליאנים. יש להציג את תקצירי השאילתה בצורה ברורה שתקל על המשתמש לבקש את המסמך המלא ולהמשיך ולראות מסמכים נוספים.

גם תוכניות המערכת (המיועדות למנהל המערכת) כגון, קליטת המסמכים הראשונית, עדכון מסמכים וסילוק מסמכים צריכות להיות נוחות לשימוש ובעלות מנשק משתמש ולא להיות פעולות הפועלות ישירות על גבי מסד נתונים עצמו באמצעות מנשק המנהל את מסד הנתונים.

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

התוכנית תכלול מסך עזר המסביר, בקצרה, למשתמש כיצד לבצע את השאילתות וכן קובץ "קרא אותי" המסביר למנהל המערכת כיצד לבצע את הקליטה, העדכון והביטול של מסמכים. **קבצים אלו בליווי הגדרת המבנה של המסמכים, מבנה מסד הנתונים ומבנה טבלאותיו וחומר הסבר נוסף יהוו את המדריך למשתמש ולמנהל המערכת שאותו יש להגיש מודפס וכרוך 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 מעניק ציון 80. פרמטרים נוספים כגון: יציבות, נוחות השימוש, אופציות נוספות (ראו להלן), קידוד אלגנטי, יעילות, מסמכים נאותים וכו' יגדילו (או יקטינו בהתאמה את הציון).

לסיכום:
על הרכיבים הבאים להימצא בפרויקט:

- איסוף מסמכים
- הגדרת מבנה המסמך
- קליטת המסמך
- סריקת המסמך
- אחסון המסמכים בספרית מסמכים (השונה מספרית ה"מקור")
- בנית האינדקסים
- אחסון כל המצביעים ושדות אינפורמטיביים במסד נתונים או בקובץ האינדקסים שבניתם
- בנית Stop List
- עדכון מסד הנתונים
- ביטול מסמכים
- בנית השאילתות עם האופרטורים
    - And
    - Or
    - Not

- תמיכה בלפחות רמה אחת של סוגריים
- הפעלת השאילתות
- מנשק משתמש
- הצגת הממצאים והתקצירים בליווי הדגשת מילות החיפוש
- הבאת המסמך המלא
- הדפסת המסמך
- מסף הסברים (מסוג "help" המבאר בקצרה את השימוש)
- חוברת הפרויקט (כפי שהובהרה לעיל)

רעיונות להרחבה (ושיפור הציון):

- קליטת מסמכים באצווה ולא אחד-אחד
- בנית שאילתות עם יותר מרמה אחת של סוגריים
- הוספת אופרנדים לשאילתה כגון:
  - Near
  - הבאת מסמך על ידי הקלה של תנאי ה and, כגון: הבא מסמך אם יש בו לפחות n מתוך m המושגים המבוקשים
- חיפוש טקסט לא רק בדיוק כפי שהוא מופיע אלא גם מחרוזות חלקיות. כלומר שימוש באופציית ה- Joker (למשל: *Car יביא גם את Car וגם את Cart)
- הגדרת משקל הרלוונטיות של המסמך (לפי כל אחת מהשיטות בהן דנו) ומיון התשובות לפי סדר משקל יורד
- הכלת תמונות / מוזיקה כחלק מהמידע האגור
- אחסון יעיל של האינדקסים על ידי שימוש ב Stemming
- שימוש במילים נרדפות (Synonyms) לשם שיפור החיפוש
- מציאת ביטויים במסמכים (בניגוד למילים בודדות) והתייחסות אליהם כמהות אחת
- חלוקת המסמכים לאשכולות ובעת השאילתה המביאה אשכול נתון להביא גם אשכולות אחרים דומים

בהצלחה!