שנקר- בי"ס גבוה להנדסה ולעיצוב
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 41, 42, 43, 44, 39 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.

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. אלא הם חומרים שהמחברים שלהם נפטרו לפני 75 שנה. פרויקטים כאלה קיימים בארצות רבות (המפורסם ביותר הוא פרויקט "גוטנברג" לספרות האנגלית).כתובת הפרויקט היא:

http://www.benyehuda.org

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

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

לפיכך יש להגדיר את המבנה הזה ראשונה ע"י בחירת של השדות הקבועים אותם אתם רוצים להציג תמיד, להציג לפי בקשה או להישאר חבויים לשימוש המערכת.

בעת ההצגה על המערכת להכיל כ 5-6 מסמכים ובמוסף צריכים להיות בקובץ ה"מקור" עוד 2-3 מסמכים נוספים שניתן יהיה להוסיפם.

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: PDF , #C או Java כולן טובות למטרה זאת.

המסמך יסרק וכל המילים המשמעותיות תישלפנה - על ידי תוכנית Parsing פשוטה המניחה שהמפרידים בין מילה למילה הם רווח, סוף שורה, נקודה, פסיק, נקודה-פסיק, נקודתיים ומקף. בנוסף יוענק לכל מסמך מספר סידורי ויבנו השדות הקבועים, אם באופן אוטומטי או ידני.
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 ומערכת האינדקסים נמצאים על גבי מחשב מקומי.
את המידע מאחסנים כקבצים בספרית ההחסנה שבה כל מסמך הוא קובץ. את מערכת האינדקסים ושדות אינפורמטיביים אחרים יש לאחסן במסד נתונים טבלאי.
יש להקפיד כי תהיה הפרדה בין ספרית ה"מקור" (קלט) וספרית ההחסנה. ספרית המקור היא זאת במכילה את הקבצים שהועלו על ידכם למערכת. ממנה יש להעביר (באצווה או אם נדרש קובץ בודד) לספרית ההחסנה וזאת בליווי בנית האינדקסים.
ספרית ה"מקור" לא תשמש גם כספרית ההחסנה עצמה. כלומר אחרי העברת המסמך למערכת ניתן יהיה למחוק אותו מספרית ה"מקור"מבלי למחוק אותו מהמערכת.

ב) מערכת אחזור אינטרנטית (בדומה ל Google) בה מערכת האינדקסים, השדות הקבועים (כולל תקציר) ימצאו על גבי מסד נתונים טבלאי במחשב המקומי, כאשר המסמכים עצמם נמצאים עדיין ברשת האינטרנט ויש עליהם הצבעה מהמערכת. (גם במקרה זה צריך להביא פעם אחת את המסך למחשב לשם סריקתו ובנית האינדקסים – אולם אין לשומרו מקומית).

מנשק המשתמש יכול להיות על ידי תוכנה ויזואלית או HTML.

אם כך, המסמך הראשון נקלט, מאוחסן בספרית ההחסנה, זיהוי, שמו , מקומו ומספר גושיו נשמרים, המילים המעניינות (יש לבנות Stop List –אולם 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, ממינים את הטבלה (*ראו טבלה מרכזית בשרטוט 43*).

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 מס*פר 43 ובנגררות בשרטוט מספר 44 והערך השני בסוגריים בטבלה מספר 96).

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

המסמך החדש ימוספר, יוכנס לספרית ההחסנה, המילים תישלפנה, ואם זאת מילה שכבר קיימת יש להגדיל את מספר ה- Hits ולעדכן את ההצבעות המתאימות. אם תבחנו תוספות אלו תבחינו עד מהרה שרצוי, כמובן, לשנות את המבנה של הטבלה המרכזית *בשרטוט מספר* 41 לשימוש במצביעים.

טבלת ה Index File תצביע על ה # Doc שבטבלת ה Posting File . שם 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 *טבלה מספר 44* .

אם זאת היא מילה חדשה 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.

אם אתם משתמשים במסד נתונים לאחסון המילים אזי התהליך פשוט יותר משום שאין אתם צריכים לנהל את הטבלה. הוסיפו את המילים למסד הנתונים. רצוי מאוד ששדה זה (שישמש כמפתח ראשי) יהיה שדה ממוין לשם שיפור זמן החיפוש.

ביטול מסמכים:
כיצד מבטלים מסמך? עוברים על גבי ה Posting File וכל פעם שפוגשים את המסמך שברצוננו לבטל, מסמנים באחד מהשדות הקבועים המשמש למטרה זאת, סימן ביטול. סילוק המסמכים עצמם ושיחזור הטבלה נעשים רק אחת למועד קבוע, משום שפעולת המחיקה והצמצום היא ארוכה: יש לסלק כל גושי 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, אולם חובה לאפשר ביטול מסמכים.

השאילתות:
השאילתות תהינה בוליאניות עם האופרטורים And , Or ו-Not ולפחות רמה אחת של סוגריים. על כל שלושת האופרטורים להיתמך. ניתן להשתמש כהנחיה *בשרטוטים מספר 39 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.

מכיוון ולא בנינו מידע על מיקומה של כל מילה במסמך, 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 לקטלג את המילים הנמצאות ב Stop List אולם אין לחפשן אלא אם הן נמצאות בין גרשיים כפולים.

במידה ועובדים בטקסטים בשפה לטינית חובה לבצע נורמליזציה לאותיות גדולות וקטנות. כלומר RESUME, resume ו- Resume יהיו זהים באינדקס. במסמכים הם עדין יופיעו 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 ü בגרמנית – השתמשו רק באנגלית).

בעת הצגת התשובות לשאילתות יש להדגיש 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 .

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.
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 ולא להיות פעולות שנשעות על גבי מסד נתונים בעזרת ממשק מנהל 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.
יש לבוא מוכנים, בעבור הבדיקה, עם מספר מילים המופיעות במסמכים שישמשו לשאילתות.
את קוד התוכנית אין להדפיס אלא יש להגיש אותו על גבי CD.

מבנה הציון:
ביצוע הפרויקט ככתבו וכלשונו מעניק ציון 80. פרמטרים נוספים כגון: יציבות, נוחות השימוש, אופציות נוספות, 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
בנית Stop List
בנית האינדקסים
אחסון כל המצביעים ושדות אינפורמטיביים במסד נתונים
אחסון המסמכים בספרית מסמכים (השונה מספרית ה"מקור")
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

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 הארגומנטים 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 במסמכים לאשכולות ובעת השאילתה המביאה אשכול נתון להביא גם אשכולות אחרים דומים.

בהצלחה!