אוניברסיטת בר-אילן, המחלקה למדעי המחשב
88-287 ארגון קבצים

ארגון קבצים – תרגיל ריצה #3 - פרוייקט סיום: עבודה עם קבצים

תאריך קבלה: 6/12/01-2/12/01. **תאריך הגשה: 17/1/02. לא יהיו דחיות!**

הוראות כלליות
- יש להגיש את תדפיס התוכנית + מספר דוגמאות ריצה מספק. **חובה לרשום שם, ת.ז., מספר הקורס, ומספר קבוצת התרגיל באופן ברור בעמוד הראשון**.
- התרגיל יוגש לתאו של המתרגל אליו הנך רשום.
- בנוסף, יש לשלוח את קוד התוכנית ל-email של המתרגל: @macs.biu.ac.il
- העבודה תבוצע **באופן אישי**. אין לעבוד בזוגות.

תוכנה לניתוח אתר HTML:

1. בתרגיל זה עליכם לממש תוכנית שמנתחת אתר HTML, לפי ההנחיות הבאות.
2. על התוכנית לגשת לעמוד הראשי ("index.htm") של אתר האינטרנט שלכם (ספריית WWW).
   על התוכנית לעשות parsing לעמוד הראשי ולגלות באילו קבצים נוספים הוא משתמש (קישורים או תמונות). על התוכנית לעבור באופן רקורסיבי על כל קבצי ה-HTML שבהם האתר משתמש עד לרמה האחרונה.
3. על התוכנית ליצור קובץ חדש שיכיל את המבנה ההיררכי של האתר. המבנה יישמר בקובץ בצורה של **עץ** (יש להגדיר struct שממשמ עץ). בקודקודי העץ יופיעו *שמות הקבצים*, כאשר שורש העץ הוא שם הקובץ הראשי של האתר ("index.htm").
   לקובץ החדש יש לקרוא בשם ***siteTree.tree*** .

הסבר:
קבצי האתר יוצרים מבנה היררכי של עץ. שורש העץ הוא העמוד הראשי של האתר והוא מהווה רמה 0. ברמה 1 יופיעו ה"בנים" של השורש, שהם כל הקבצים שנקראים מתוך העמוד הראשי של האתר. למשל, אם השורש הוא הקובץ index.htm ובקובץ index.htm יש קריאה לעמודים index.jpg, main.htm, אזי ברמה 1 יהיו שני בנים, לפי שני הקבצים הנ"ל.
ברמה 2 יופיעו כל ה"בנים" של הצמתים ברמה 1. וכן הלאה.
שורש העץ הוא שם הקובץ הראשי של האתר – *index.htm*.
עלה בעץ הוא או (א) שם קובץ **בינארי** (למשל: *.jpg, *.mpg), או (ב) שם כל קובץ אחר שאינו HTML (למשל, *.txt), או (ג) שם קובץ HTML שאין ממנו קריאות לקבצים נוספים.

4. כמו כן, עבור כל קובץ HTML שקיים בעץ האתר (*.htm, *.html), יש ליצור קובץ חדש שבתוכו יישמר קובץ ה-HTML בצורה של **עץ**.
   לכל קובץ כזה יש לקרוא בשם ***fileName.htm.tree*** או ***fileName.html.tree*** .

הסבר:
קובץ HTML בנוי מתגים (למשל, <HTML>, <HEAD>, </HTML> וכו'). ***תג*** הוא אלמנט שמצוי בין פותח וסוגר זוויתיים <TAG>. תג יכול להיסגר (כלומר, קיים עבורו תג זהה עם לוכסן בתחילתו, למשל - </BODY> שסוגר את התג <BODY>). אולם יש תגים שאינם נסגרים או תגים שאינם חייבים להיסגר (למשל: <P> יכול להיסגר ויכול לא להיסגר. לעומתו <IMG> אינו נסגר).
התגים שקיים עבורם תג סוגר בעצם מגדירים את המבנה ההיררכי של העמוד.

כל *קודקוד* בעץ (שורש, עלה או צומת) מחולק ל-3:
א) שם התג שחייב להיסגר (<X>),
ב) הפרמטרים הנלווים לתג (למשל: עבור <IMG SRC="a.jpg">, הפרמטרים הם "SRC="a.jpg)
[אם לא קיימים פרמטרים נלווים אזי השדה הזה יהיה ריק],
ג) הטקסט החסום בין התג לתג הסוגר אותו (טקסט זה יכול להכיל תגים נוספים), לא כולל התג עצמו (כלומר, לא כולל <X> ו-</X>).

ה***שורש*** הוא (א) התג <HTML> ו-(ב) הטקסט החסום בין התג לתג </HTML>.
ברמה הבאה בעץ יופיעו התגים הנסגרים **ברמה הראשונה** שנמצאים בין ה-<HTML> ל-</HTML> (ראו דוגמה בהמשך), וכן הלאה.
עלה בעץ הוא תג שחייב להיסגר שאינו מכיל **תגים** נוספים.

5. בנוסף יש לשמור ב-***log file*** הודעות על broken links, כלומר: יש לשמור הודעות על קבצים שלא קיימים, אולם יש קישור עבורם בדפי האתר. עבור כל הודעה על קובץ שלא קיים יש לשמור גם את התאריך ואת השעה בה נכתבה הודעה זו (ניתן להיעזר ב-struct tm).
   יש לשמור את ההודעות בקובץ זה ב**צורה ממוינת לפי סיומות הקבצים**, למשל: הודעות עבור קבצי HTML יופיעו בהתחלה, אח"כ יופיעו הודעות על קבצי MPG וכיו"ב. בתוך כל קבוצה יש לשמור את ההודעות, ב**צורה ממוינת לפי התאריך**, כלומר: שההודעה האחרונה שנרשמה תהיה הראשונה בקבוצה, וההודעה הראשונה שנרשמה תהיה האחרונה.
6. התוכנית עשויה לקבל פרמטר אחד (מספר) מה-command line. מספר זה יקבע כמה רמות יש לרדת בעץ האתר (ר' סעיף 3). פרמטר זה יקבע כמה רמות יש לחקור (יכול להיות שהמספר יהיה גדול ממספר הרמות הקיים). אם לא מסופק אף פרמטר, יש לחקור את כל הרמות.
   שימו לב, שהמספר הנ"ל משפיע על סעיפים **3 ו-4**.
7. אין צורך לחקור קבצים שנמצאים באתרי אינטרנט אחרים (למשל, אין צורך לחקור קישורים המפנים לאתר המחלקה או כל אתר אחר, שאינו האתר שלכם).
8. לאחר המימוש הנ"ל, על התוכנית לתמוך בקריאה לפונקציות הבאות:

void showSiteTree(); - הצגת עץ האתר בפלט הסטנדרטי.

void showFileTree("file.htm"); - הצגת עץ הקובץ file.htm.

void showFilesTrees(); - הצגת כל עצי הקבצים של האתר.

void updateFile("file.htm"); - עדכון הקובץ file.htm. אם הקובץ file.htm היה קיים יש לדרוס אותו, ואם הקובץ לא היה קיים יש ליצור קובץ חדש. יש לדאוג שלאחר עדכון הקובץ, גם מבנה העץ של הקובץ יעודכן, וכן מבנה העץ של האתר יעודכן (אם, למשל, הקובץ מכיל קריאות לקבצים חדשים או לא מכיל קריאות לקבצים שקודם כן היו).

void save(); - פונקציה זו שומרת יחד את כל הקבצים החדשים שנוצרו בקובץ אחד שייקרא siteName.prj. את הקובץ הנ"ל יש לשמור בספרייה חדשה ששמה siteName. לאחר הקריאה לפונקציה save() יש למחוק את כל הקבצים האחרים שנוצרו (ניתן להשתמש בפונקציה ***remove*** שמוגדרת ב-stdio.h) ולא ניתן להפעיל יותר פעולות על האתר הנוכחי (למעט במקרה שיצוין בהמשך).

void retrieve("siteName.prj"); - שחזור הקבצים המצויים בקובץ siteName.prj בספרייה siteName. לאחר הקריאה לפונקציה זו ניתן להמשיך ולהפעיל פעולות על האתר הנוכחי.

9. יש ליצור קובץ action שבו תהיינה הפקודות שברצונכם לבצע על האתר. פונקצית ה-main() תקרא מקובץ זה את הפעולות ותבצע אותן.

לתשומת לב:

10. על-מנת להקל עליכם, ניתן להניח שבאתר אין קריאות לקבצים שיוצרות לולאות (כלומר, לא ייווצר מצב, למשל, שבו קובץ index.htm קורא לקובץ body.htm והקובץ body.htm קורא לקובץ index.htm).
11. קובץ HTML משתמש בקבצים נוספים ע"י השימוש בתגים <IMG> או <A>, כדלהלן:
<IMG SRC="imgFile.jpg" ALT="Image">
<A HREF="link.htm">, <A HREF="link.mpg">
יש לשים לב, שהפרמטר SRC או HREF יכול להופיע בכל מקום בתג, ולאו-דווקא בתחילתו.
12. התכנית תיכתב בשפת C. יש לוודא כי התוכנית מתהדרת בעזרת cc או gcc (במחשבי ה-UNIX שבמחלקה למתמטיקה ומדעי המחשב).
13. יש לבצע את כל הגישות לקבצים באמצעות ה-low-level functions או high-level functions שהכרתם בשיעור התרגיל **<u>בלבד</u>**.
14. על התוכנית לרוץ ולטפל בכל קלט אפשרי.
15. אין לממש פונקציונליות נוספת מעבר להוראות שפורטו לעיל.
16. יש לכלול בתחילת ה-Source Files את הנתונים הבאים: שם הקובץ, תיאור הקוד, שם היוצר, ותאריך היצירה. כמו כן יש לפרט את קבוצת ההרצאה והתרגול. יש לכלול לכל פונקציה ולכל מבנה נתונים הסבר מפורט. לדוגמה, עבור פונקציות יש לכלול בהסבר את הפרמטרים של הפונקציה, הערך המוחזר, משתנים גלובליים שמושפעים ותיאור הפונקציה.
17. בעת חישוב הציון לתרגיל יילקחו בחשבון המרכיבים הבאים (סדר חשיבות יורד): נכונות התוכנית, יעילות התוכנית, קריאות התוכנית.

<u>בהצלחה</u>

<u>דוגמה:</u>

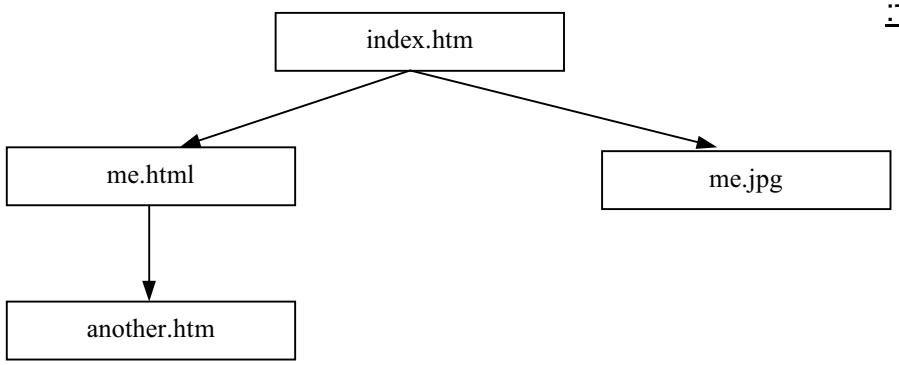
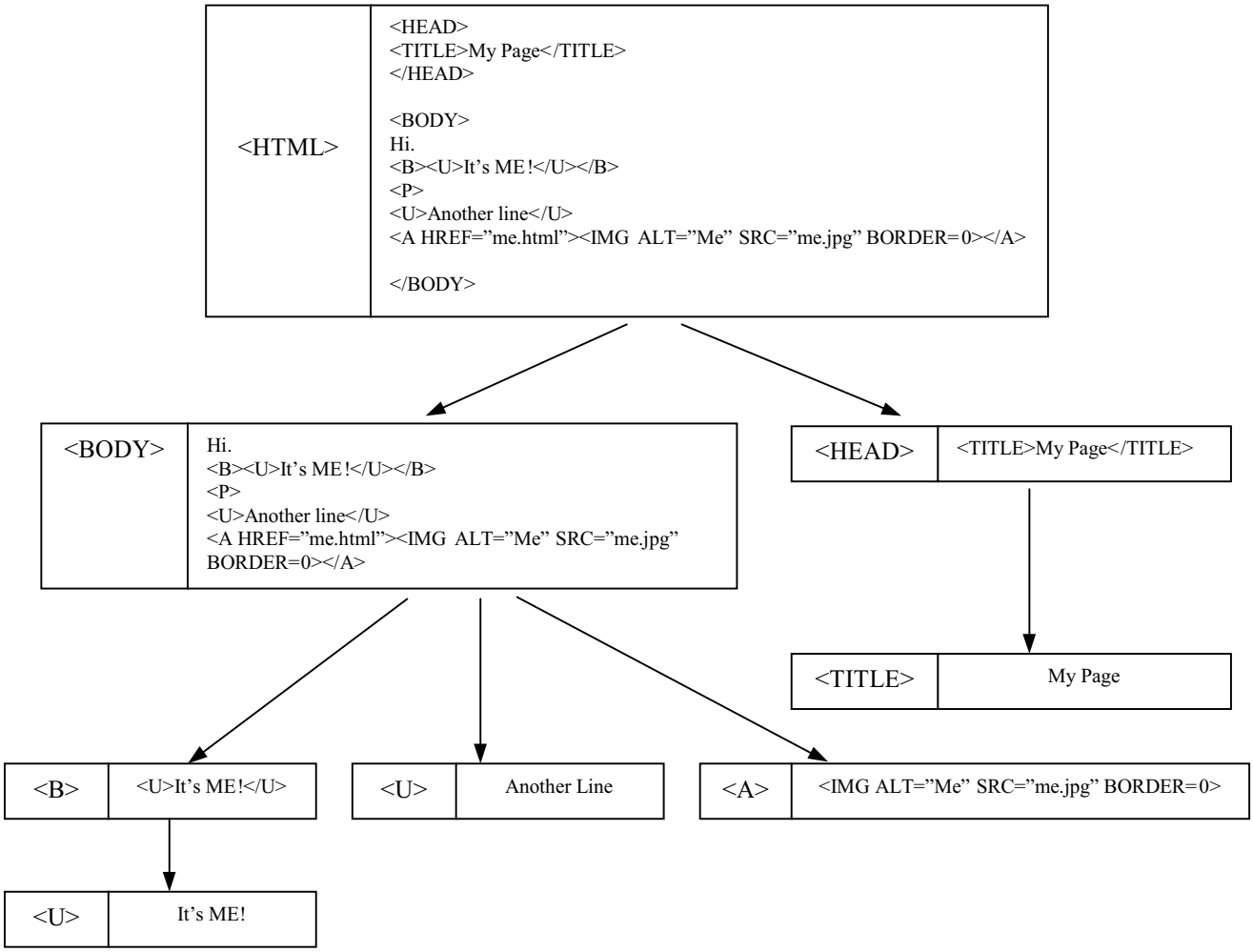
```
% cat index.htm
<HTML>
<HEAD>
<TITLE>My Page</TITLE>
</HEAD>

<BODY>
Hi.
<B><U>It’s ME!</U></B>
<P>
<U>Another line</U>
<A HREF=”me.html”><IMG ALT=”Me” SRC=”me.jpg” BORDER=0></A>

</BODY>
</HTML>

% cat me.html
<HTML>
<HEAD>
<TITLE></TITLE>
</HEAD>
<BODY>
<A HREF=”another.htm”>another</A>
</BODY>
</HTML>
```

<u>עץ האתר:</u>



עץ הקובץ index.htm: