**שיעורי בית 7: קבצים**

**הנחיות**

* שימו לב לסגור כל קובץ שאתם פותחים.
* במקרים בהם ישנו קובץ הדגמה יש לחקותו במדויק.

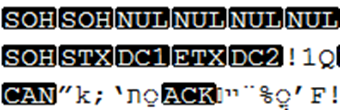
**חלק א': זיהוי ומחקר קבצים בינאריים**

למדנו בשיעור שניתן לחלק את הקבצים ל-2 סוגים - קבצים טקסטואליים (כאלו המכילים תווי ASCII דפיסים) וקבצים בינאריים (קבצים המכילים לפחות תו אחד שאינו דפיס).

דוגמאות לקבצים בינאריים הן קבצי pdf, exe, zip וכו'.

בחלק זה נחקור כיצד ניתן לזהות את סוג הקבצים הבינאריים גם אם זהותם הוסתרה מאיתנו.

**הבעיה:** בעוד שנוכל לקרוא תוכן של קבצים טקסטואליים באמצעות פתיחה שלהם בכתבן (++Notepad לדוגמא), אם נפתח קבצים בינאריים בכתבן נקבל ג'יבריש וערכים שלא ניתן להציג, לדוגמא:

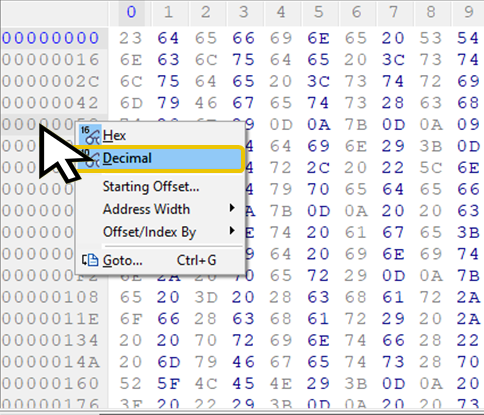


**הפתרון:** לשם כך נשתמש בתוכנה ייעודית בשם Hex Workshop.

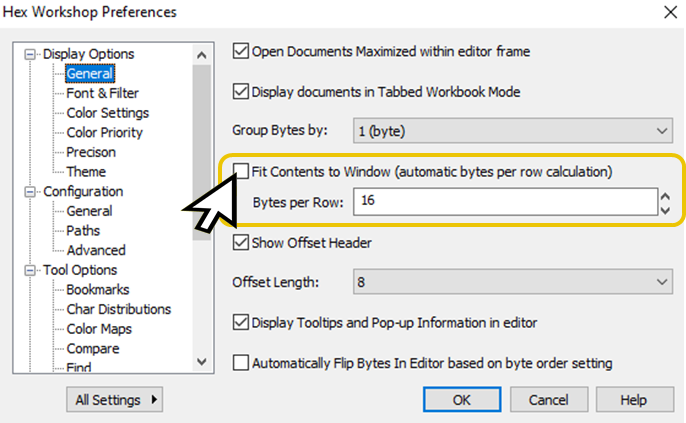
1. התקינו את התוכנה Hex Workshop לפי ההוראות הבאות:

א) התקינו את תוכנת Hex Workshop מהלינק הבא: http://www.bpsoft.com/downloads

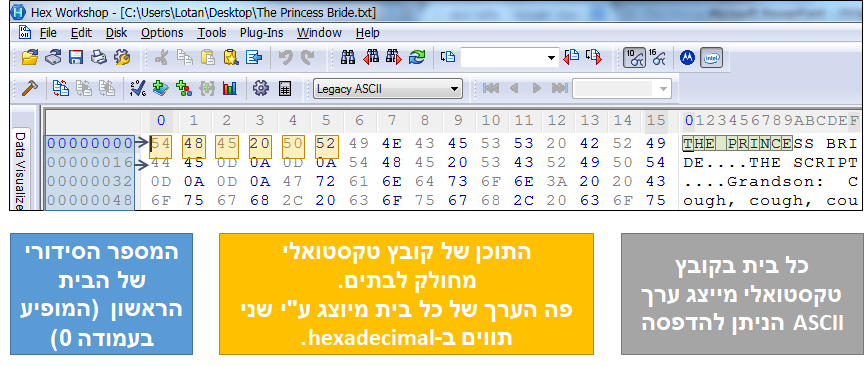
ב) הורידו את גרסא v6.7.3 (ניתן להשתמש ישירות ב[לינק הזה](http://www.bpsoft.com/bbl/download/download.html?dlfile=hw_v673.exe)).

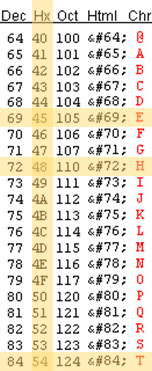
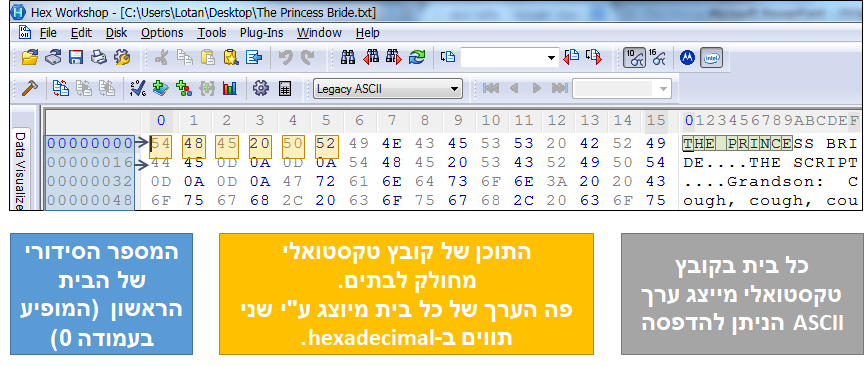
ג) הורידו את הקובץ "The princess bride - script.txt" ופיתחו אותו בתוכנה: ניתן לפתוח קובץ מהתפריט File > Open או ע"י גרירה עם העכבר של קובץ כלשהו מהמחשב לתוך התוכנה.

ד) נעדכן את התצוגה:

* **נעדכן את פורמט מספור השורות:** לחיצת מקש ימני על הספרות בצד שמאל (רקע אפור) ולחיצה על Decimal במקום Hex תשנה את המספור הסדרתי של הבתים בקובץ מספירה בבסיס 16 לספירה בבסיס 10.
* **נעדכן את גודל השורות:**Options > Preferences > Display Options > General והסירו את הסימון מ - Fit Contents To Window כך שמספר הבתים לשורה יהיה 16.

נבחן את התצוגה בתוכנה:





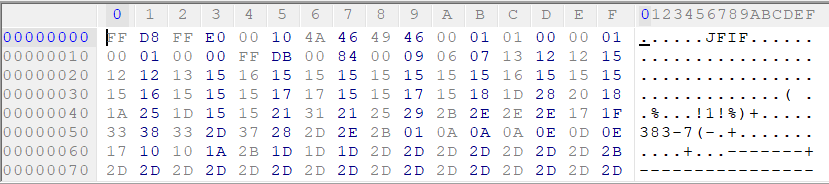
**שימו לב:** ערך האסקי של התו T הוא אמנם 84 אך בבסיס הקדסצדימלי - זהו 54.

ערכו של התו H בבסיס הקסדצימלי הוא 48 ואילו E הוא 45.

כאשר נפתח קובץ בינארי ב-Hex Editor נוכל לבחון את הערכים ההקסדיצמליים של תוכנו אך לא נוכל לקרוא אותם. כך לדוגמא, אם ניקח את תמונת החתול הזו:



וננסה לפתוח אותה ב-Hex Editor, נקבל:



הפעם התוכן, המופיע מתפריט הימני, אינו קריא כלל.

2. נשתמש כעת בתוכנת Hex Workshop על מנת לבדוק איך נראים קבצים מסוגים שמוכרים לנו. הורידו את הקובץ ClassEx2\_files.zip וחלצו ממנו את הקבצים.

בסעיפים הבאים אתם תקבלו קבצים בינאריים עם סיומת file. אם תנסו לפתוח את הקבצים לא תצליחו, כיוון שזוהי לא סיומת מוכרת. עליכם לפתוח את הקבצים בעזרת ה-Hex Workshop, לגלות באיזה סוג קובץ מדובר, ולשנות את סיומת הקובץ כך שתוכלו לפתוח אותו.

**רמז:** חפשו מילים חשודות בקובץ שמעידות על סוג הקובץ.

א. פתחו את הקובץ **magsh.file1** בעזרת Hex Workshop. באיזה סוג קובץ מדובר? שנו את הסיומת שלו כך שתוכלו לפתוח אותו.

ב. פתחו את הקובץ **magsh.file2** בעזרת Hex Workshop. באיזה סוג קובץ מדובר? שנו את הסיומת שלו כך שתוכלו לפתוח אותו.

ג. פתחו את הקובץ **magsh.file3** בעזרת Hex Workshop. באיזה סוג קובץ מדובר? שנו את הסיומת שלו כך שתוכלו לפתוח אותו. **רמז:** אם מצאת מילה שנראת "חשודה" אך אתם לא מכירים, חפשו בגוגל.

ד. פתחו את הקבצים **header.a header.b header.d** בעזרת Hex Workshop. כבר התרגלתם לזהות את סוג הקובץ ולכן הפעם ננסה לשים לב למשהו חדש.

הסתכלו על הבתים הראשונים בכל אחד מהקבצים. נשים לב כי הבתים הללו חוזרים על עצמם בקבצים שונים! בתים אלו הם חלק מה- header של הקובץ, מידע שמצוי בתחילתו של קובץ מסוג מסוים ושתוכנות מסוימות יודעות להשתמש בו.

מהו מספר הבתים הזהים בין הקבצים שונים?

חפשו את 7 הבתים הראשונים בגוגל עם הרווחים ביניהם (52 61 72 …) ומצאו עמוד בויקיפדיה שיעזור לכם בשאלה הבאה.

ה. פתחו את הקבצים **1 ... 13** ב Hex Workshop. נסו לזהות אילו בתים חוזרים בין הקבצים השונים והיעזרו בעמוד הויקיפדיה מהסעיף הקודם בשביל להבין באיזה סוג קבצים מדובר. שנו את סיומת הקבצים כך שתוכלו לפתוח אותם.

ו. פתחו שני מסמכי .txt שאתם לא צריכים ב-hex workshop. השתמשו בהעתק-הדבק על מנת להעביר טקסט מקובץ אחד לשני.

ז. פתחו שתי תמונות שאתם לא צריכים ב-hex workshop. נסו להעתיק חלקים מתמונה אחת לשנייה, האם אתם מצליחים לגרום לחלקים מתמונה אחת להופיע בשנייה? למה או למה לא? חקרו זאת...

**חלק ב': קריאת וכתיבה לקבצים בשפת C**

3. בשאלה זו נממש תוכנית להעתקת קבצים. התוכנה תקבל פרמטרים ל-main (כלומר - argc, argv):

* הפרמטר הראשון: מחרוזת המתארת בחירה בין העתקה של קובץ טקסט או בינארי   
  (textCopy / binaryCopy).
* הפרמטר השני: נתיב קובץ המקור.
* הפרמטר השלישי: נתיב קובץ היעד להעתקה.

פורמט הרצת התכנית:

program.exe textCopy source\_file destination\_file

**דוגמה:**

program.exe textCopy c:\Or.txt d:\Omer.txt

כאן התוכנית תעתיק את הקובץ Or.txt אשר נמצא בכונן C לקובץ בשם Omer.txt שנמצא בכונן D.

**הנחיות:**

* ראו קובץ הדגמה copyFile.exe לכל התנהגות של התוכנה.
* במידה וחסרים פרמטרים, או שהפרמטר הראשון שגוי – יש להודיע למשתמש ולצאת.
* במקרה וקובץ המקור לא קיים, תוצג הודעה מתאימה למשתמש והתוכנית תסיים את ריצתה.
* במקרה וקובץ היעד קיים, המשתמש יצטרך לבחור האם לעשות OVERWRITE (לדרוס את הקובץ הקיים) או לוותר על ההעתקה.
* בכל מקרה אחר יש להעתיק את קובץ המקור לקובץ היעד מבלי להטריח את המשתמש.
* לאחר שהקובץ הועתק יש להדפיס הודעת הצלחה למשתמש.
* יש לוודא שסגרנו את הקבצים בתום השימוש.
* הימנעו משכפול קוד ככל האפשר! ואפשר!
* עבור textCopy ניתן להניח שמתקבל קובץ טקסט ולא קובץ בינארי. תוכלו להשתמש ב-fputc ו-fgetc (העזרו בדוגמא שראינו בשיעור).
* עבור binaryCopy יש לכתוב פונקציה נפרדת, ובה אין להשתמש בפקודות fgetc, fputc. יש להשתמש במקום בפקודות [fread](http://www.cplusplus.com/reference/cstdio/fread/) ו- [fwrite](http://www.cplusplus.com/reference/cstdio/fwrite/). בנוסף, הפקודות [ftell](http://www.cplusplus.com/reference/cstdio/ftell/), [fseek](http://www.cplusplus.com/reference/cstdio/fseek/), [feof](http://www.cplusplus.com/reference/cstdio/feof/) עשויות להיות שימושיות.

4. בשאלה זו נחקור את העבודה עם קבצי CSV - סוג קבצים נפוץ מאוד בו עושות שימוש תוכנות רבות.

כפי שראינו בשיעור, קבצי CSV (קבצים עם סיומת .csv) הינם קבצי טקסט המאפשרים עבודה עם טבלאות. הקובץ מסודר כך שכל ערך מופרד מהערך שבטור הבא על ידי פסיק (התו ','), כל שורה מופרדת מהשורה הבאה על-ידי new line ('\n'). ניתן לקרוא עוד על סוג הקבצים [כאן](http://en.wikipedia.org/wiki/Comma-separated_values).

מצורף לכם קובץ דוגמה שכזה example.csv, ייתכן שמערכת ההפעלה תציע לכם לפתוח אותו באקסל. כדאי לפתוח אותו עם notepad++ על מנת להבין את המבנה שלו.

עליכם לכתוב תוכנית המקבלת כפרמטר ל-main נתיב של קובץ .csv, ואת הפונקציה **searchFile**. הפונקציה תקבל מחרוזת לחיפוש בקובץ. במקרה ונמצא - התוכנית תחזיר את מספר השורה בה המחרוזת נמצאה. במידה והמחרוזת לא נמצאה בקובץ התוכנית תחזיר 0.

**הנחיות:**

* ניתן להניח כי ערכים שמתקבלים הם תקנים, אך יש לבצע בדיקת תקינות לכל פתיחה וכתיבה לקובץ.
* ראו קובץ דוגמה csv.exe.
* שימו לב שגם אם לא פתרתם את סעיפי הבונוס עליכם לממש תפריט זהה – רק שאפשרויות 2 ו-3 לא יעשו דבר, והתפריט יוצג שוב מיד.

**בונוס:** ממשו שתי פונקציות נוספות:

* **changeValue** – הפונקציה תקבל מספר שורה, מספר טור, ומחרוזת. התוכנית תחליף בקובץ את המילה במקום המתאים במחרוזת שהתקבלה.
* **copyValue** – הפונקציה תקבל מספר שורה ומספר טור כמקור ומספר שורה ומספר טור כיעד. הפונקציה "תדרוס" את המחרוזת הנמצאית בתא היעד במחרוזת הנמצאית בתא המקור. ניתן להניח שבתא היעד קיים ערך באותו גודל של תווים.

5. **בונוס:** בשאלה זו נשתמש בתסריט של "מלחמת הכוכבים" על מנת להצפין את התסריט של "הנסיכה הקסומה".

ב-NEO תוכלו למצוא שלושה קבצים:

* Starwars.txt
* The Princess Bride.txt
* password.txt

שני הקבצים הראשונים מכילים את התסריטים של שני הסרטים המפורסמים. הקובץ השלישי ישמש עבורנו כקובץ סיסמא, הוא מכיל 100,000 מספרים בטווח 1-334000.

אלגוריתם ההצפנה יצפין את כל אחד מהבתים בקובץ The Princess Bride.txt בזה אחר זה באופן הבא:

1) אתחלו את המשתנה byteNum = 0.

2) עברו על כל הבתים בקובץ The Princess Bride.txt בזה אחר זה ובצע:

a. הגדילו את byteNum ב-1 (משתנה זה מייצג את מספר הבית אותנו אנו מצפינים כרגע).

b. קראו מהקובץ password.txt את המספר ה-byteNum ושמרו אותו לתוך המשתנה code.

c. קראו מהקובץ Starwars.txt את הבית ה-code (הבית הראשון בקובץ הוא הבית ה-0).

d. בצעו פעולת XOR בין הבית שקראתם מהקובץ Starwars.txt לבית אותו אתם מצפינים כרגע מהקובץ The Princess Bride.txt.

e. שמרו את הבית המוצפן encByte לקובץ חדש.

**דוגמא בעמוד הבא**

**דוגמא:**

| encByte =  הבית המוצפן | plainByte =  הבית ה-byteNum בקובץ The Princess bride.txt | הבית ה- codeבקובץ starwars.txt | code =  המספר ה-byteNum בקובץ password.txt | byteNum |
| --- | --- | --- | --- | --- |
|  |  |  |  | 0 |
| 01110100 | 0x54 = 84 = 01010100 | 0x20 = 32 = 00100000 | 42 | 1 |
| 01000101 | 0x48 = 72 = 01001000 | 0x0D = 13 = 00001101 | 18468 | 2 |
| 01100101 | 0x45 = 69 = 01000101 | 0x20 = 32 = 00100000 | 6335 | 3 |
| 00000000 | 0x20 = 32 = 00100000 | 0x20 = 32 = 00100000 | 26501 | 4 |
| 00100010 | 0x50 = 80 = 01010000 | 0x72 = 114 = 01110010 | 19170 | 5 |

**רמז:** במעבר על הקובץ starwars.txt השתמשו בפונקציה fseek שהוצגה בשיעור.

**בהצלחה!**