בית הספר להנדסה ומדעי המחשב ע"ש רחל וסלים בנין

מבוא למדעי המחשב 67101

הרחבה לתרגיל 5 - קבצי **XML** וקבצי משרד הכלכלה

קבצי XML - eXtensible Markup Language, הם קבצי טקסט המאפשרים לשמור מבני נתונים באופן שיהיה קל לקריאה. את הפורמט, שהומצא בשנת 1996, ניתן למצוא כיום כמעט בכל מקום. הXML מורכב מתגיות שונות, הנכתבות בפורמט המזכיר את שפת הHTML. כל תג ימצא בין סימני ה"<" ו- ">" כשלאחריהם מופיע הערך של התג הנ"ל עבור הקובץ הקיים. שימו לב כי בניגוד לHTML, התגיות בקבצי XML נועדו רק בכדי לתחום פיסות מידע - ואין להניח כי התג
 בקובץ אחד , משמעותו זהה לתג הנ"ל בקובץ אחר. האחריות על הבנת הקבצים תלויה בתוכנה הקוראת אותם.

כאמור, XML הוא פורמט המאפשר שמירת מבני נתונים. בניגוד לקבצים אחרים הנועדים למטרה זו, הפורמט לא מגדיר טבלה, אלא דווקא מבנה נתונים של עץ. על עצים עוד נלמד בהמשך הקורס ובהמשך התואר - אך לבינתיים אפשר לחשוב על עץ כעל מבנה היררכי. עצים במדעי המחשב הם הפוכים מהטבע - נהוג לצייר אותם כאשר שורש העץ למעלה - ומשורש זה יוצאים ענפים שונים. כל ענף מגדיר תכונה מסוימת של העץ, או תת חלוקה שלו. כל ענף יכול להוות שורש נוסף לתת עץ בפני עצמו.

כך בדוגמא לXML הבאה:

```
<?xml version="1.0"?>
<catalog>
  <book id="bk101">
    <author>Gambardella, Matthew</author>
    <title>XML Developer's Guide</title>
    <price>44.95</price>
    <publish_date>2000-10-01</publish_date>
    <description>An in-depth look at creating applications with XML.</description>
  </book>
  <book id="bk102">
    <author>Ralls, Kim</author>
    <title>Midnight Rain</title>
    <price>5.95</price>
    <publish_date>2000-12-16</publish_date>
    <description>A former architect battles corporate zombies, an evil sorceress, and her own childhood to become queen of the world.</description>
  </book>
</catalog>
```

שורש	העץ	
1	ענף	
1.1	ענף	
1.2	ענף	
1.5	ענף	
1 סגירת	ענף	
2	ענף	
2 סגירת ענף		
סגירת השורש		

הדוגמא מציגה קטלוג ספרים השמור בפורמט XML.

הקטלוג הוא שורש העץ, התג הגבוה ביותר בהיררכיה, הוא נפתח בתחילת הקובץ ונסגר בסופו (סגירת תג - <tagname/>).
מתחת לשורש העץ ניתן למצוא במקרה הנ"ל שתי תגיות נוספות - כאשר לשתיהן קוראים book.
הbook הראשון, אותו נכנה הענף הראשון , מהווה בעצמו שורש למספר תגיות נוספות. במקרה הנתון כל תגית כזו מתארת את אחת מתכונות העץ. למשל אנו יכולים לראות כי שמו של הספר הראשון בקטלוג החנות הזו הוא "XML Developer's Guide" שנכתב על ידי Gambardella, Matthew ונמכר במחיר המופרז של 44.95 דולר וכן תכונות נוספות.לאחר סיום תיאור ספר זה (שגם כן נסגר בתגית </book>) נוכל למצוא תיאור של ספר נוסף. כיצד נדע כי המחיר 5.95 שייך לספר השני ולא לספר הראשון? בזכות המבנה ההיררכי.

בתרגיל זה נעבוד עם הקבצים של משרד הכלכלה. הפרויקט בתחילת הדרך ועל כן פורמט הקבצים עדין אינו אחיד לחלוטין. אלמנט זה הוא קריטי שכן מכיין שקבצי XML לא מכתיבים פורמט אחיד ותגיות זהות. על כן על התכנית שלנו להניח הנחות מסוימות על הקלט.
בתרגיל זה נעבוד עם קבצים מהסגנון איתם עובדת חברת שופרסל. חלק קטן מקובץ כזה ניתן לראות כאן:


```
<?xml version="1.0"  encoding ="utf-8"?>
<root>
 <ChainId>7290027600007</ChainId>
 <SubChainId>001</SubChainId>
 <StoreId>002</StoreId>
 <BikoretNo>8</BikoretNo>
 <DllVerNo>8.0.1.0</DllVerNo>
 <Items Count="2">
        <Item>
        <PriceUpdateDate>2015-01-06 07:31</PriceUpdateDate>
        <ItemCode>11210000094</ItemCode>
        <ItemType>1</ItemType>
        <ItemName>רוטב טבסקו 60 מ"ל</ItemName>
        <ManufacturerName>ניצן</ManufacturerName>
        <ManufactureCountry>US</ManufactureCountry>
        <ManufacturerItemDescription>רוטב טבסקו 60 מ"ל</ManufacturerItemDescription>
        <UnitQty>מיליליטרים</UnitQty>
        <Quantity>60.00</Quantity>
        <UnitOfMeasure>100 מ"ל</UnitOfMeasure>
        <QtyInPackage>0</QtyInPackage>
        <ItemPrice>12.80</ItemPrice>
        <UnitOfMeasurePrice>21.33</UnitOfMeasurePrice>
```

```
        <AllowDiscount>1</AllowDiscount>
        <ItemStatus>1</ItemStatus>
        </Item>
        <Item>
        <PriceUpdateDate>2015-07-07 08:26</PriceUpdateDate>
        <ItemCode>13495113506</ItemCode>
        <ItemType>1</ItemType>
        <ItemName>קליק ביסקוויט 75 גרם</ItemName>
        <ManufacturerName>יוניליוור</ManufacturerName>
        <ManufactureCountry>IL</ManufactureCountry>
        <ManufacturerItemDescription>קליק ביסקוויט 75 גרם</ManufacturerItemDescription>
        <UnitQty>גרמים</UnitQty>
        <Quantity>75.00</Quantity>
         <UnitOfMeasure>100 גרם</UnitOfMeasure>
        <QtyInPackage>0</QtyInPackage>
        <ItemPrice>5.00</ItemPrice>
        <UnitOfMeasurePrice>6.67</UnitOfMeasurePrice>
        <AllowDiscount>1</AllowDiscount>
        <ItemStatus>1</ItemStatus>
        </Item>
  </Items>
</root>
```


זוהי דוגמא מוקטנת לחנות ובה שני מוצרים.

ראשית לקובץ ישנו שורש - המוגדר בשמו root. לאחריו מתחילים התיאורים השוני של החנות. תכונות שנדרשות לכל חנות כמו ספרת ביקורת וID של החנות וכו. התיוג האחרון מתחת לשורש הוא תיוג הItems. במקרה הנ"ל בחנות שני מוצרים בלבד, רוטב טבסקו וחטיף קליק. בקבצים האחרים איתם תעבדו, תמצאו חנויות בעלות אלפי מוצרים. כל מוצר בקובץ הוא שורש לתת עץ בפני עצמו. מה מכיל תת העץ הנ"ל ? את תיאור המוצר - תחת התג PriceUpdateDate נלמד מתי עדכן לאחרונה מחיר המוצר. תחת התגית ItemCode נמצא את קוד המוצר - אלמנט חשוב עבור התרגיל שלנו. התגית ItemName תתאר את שם המוצר וכן הלאה.

בסיום תיאור המוצר נמצא את הסוגר </Item>, כלשאחריו נעבור למוצר הבא או לסגירת התגית Items הגדולה ולאחריה סגירת השורש וסיומו של הקובץ.