ניהול נתונים באינטרנט 2022 – תרגיל מספר 1 - XML

הוראות:

יש לעלות את הפתרונות ל-Moodle בקובץ ZIP שכולל:

- קובץ PDF בשם answers.pdf ובו הפתרון
- קבצי קוד נוספים (HTML, XML או Python) לפי הדרישה של כל סעיף.

ההגשה היא בזוגות, ורק אחד מבני הזוג יגיש את התרגיל, אך יש להקפיד לכתוב את השמות והת.ז. של שני בני הזוג בתוך הקובץ. שם של קובץ ה-ZIP צריך לכלול את הת.ז. של אחד מהמגישים (למשל: HW1_123.zip).

תאריך הגשה: 24/03/2022

שאלה 1

לפניכם קובץ XML וחלק מהDTD המתאים לו. במסמך ישנן מספר שגיאות מסוגים שונים, כמו אי התאמה לDTD, וטעויות בסינטקס. מצאו 3 שגיאות המופיעות בקובץ:

DTD:

```
<!ELEMENT PARTS (TITLE?, PART*)>
<!ELEMENT TITLE (#PCDATA)>
<!ELEMENT PART (ITEM, MANUFACTURER, MODEL, COST)+>
<!ELEMENT ITEM (#PCDATA)>
<!ELEMENT MANUFACTURER (#PCDATA)>
<!ELEMENT MODEL (#PCDATA)>
```

XML:

```xml
<PARTS>
   <TITLE>Computer Parts</TITLE>
   <PART>
      <ITEM>Motherboard</ITEM>
      <MANUFACTURER>ASUS</MANUFACTURER>
      <MODEL>P3B-F</MODEL>
      <COST> 123.00 </COST>
   </PART>
   <PART>
      <ITEM>Video Card</ITEM>
      <MANUFACTURER>ATI</manufacturer>
      <COST CURRENCY='USD'> 160.00</COST>
   </PART>
   <PART>
      <ITEM>Sound Card</ITEM>
      <MANUFACTURER>Creative Labs</MANUFACTURER>
      <MODEL>Sound Blaster Live</MODEL>
      <COST CURRENCY='USD'> 80.00
   </PART>
   <PART>
      <ITEM>Mouse</ITEM>
      <MANUFACTURER>Logitec</MANUFACTURER>
      <MODEL>AF180</MODEL>
      <COST> <CURRENCY ID='USD'>8.00</CURRENCY>  <CURRENCY ID='EUR'>10.00</CURRENCY> </COST>
   </PART>
   <PART>
      <ITEM> 19 inch Monitor</ITEM>
      <MANUFACTURER>LG Electronics</MANUFACTURER>
      <MODEL> 995E</MODEL>
      <COST CURRENCY='USD'> 290.00</COST>
   </PART>
   <PART />
      <ITEM> Processor </ITEM>
      <MANUFACTURER>Intel</MANUFACTURER>
      <COST CURRENCY='EUR'> 240.00</COST>
</PARTS>
```

שאלה 2

בשאלה זו נתייחס לקובץ: books.xml שמצורף לתרגיל ב-Moodle.

(א) כתבו DTD מתאים לקובץ (ישנה יותר מאפשרות נכונה אחת) והסבירו את ה-DTD שכתבתם.

(ב) כתבו שאילתת XPATH שמחזירה את הנתונים הבאים:

a. שמות החנויות שיש בהם ספרים שעוסקים באנגליה (כלומר, בתיאור הספר יש את המילה England)

b. שמות ספרי הפנטזיה שעולים פחות מ-6 דולר בלפחות אחת החנויות.

c. מספר הספרים שיש להם מחבר אחד בדיוק

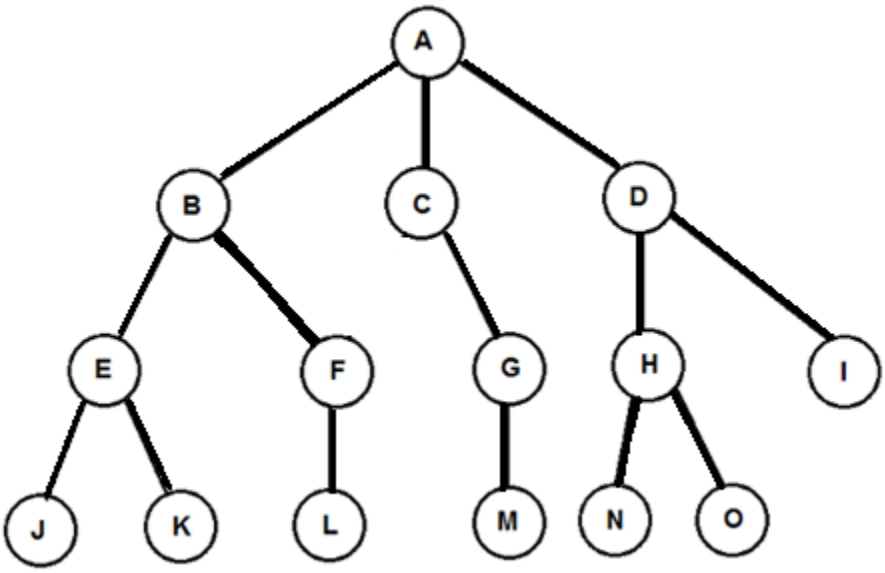
d. שמות המחברים שכתבו ספר בשנת 2000

(ג) הריצו את כל אחת מהשאילתות שקיבלתם על הקבצים הנתונים והגישו את התוצאות.

ניתן להשתמש ב: http://www.xpathtester.com/xpath

שאלה 3

נתון העץ הבא:



(א) כתבו את הDOCUMENT ORDER של העץ הנ"ל.

(ב) כתבו את העץ הנ"ל ביצוג של קובץ XML.

(ג) כתבו את התוצאה של הרצת השאילתות הבאות על העץ הנ"ל:

a. preceding-sibling של קודקוד I.
b. descendant של קודקוד D.
c. following של קודקוד M.
d. ancestor-or-self של K.
e. preceding של L.
f. following של Parent של O.

שאלה 4

(א) הכינו קובץ HTML הכולל תיאור של מדענית. דף זה צריך לכלול תמונה של המדענית, דגל של המדינה בה נולדה, וטבלה שכוללת נתונים על העיר בה נולדה (למשל כמות תושבים לפי שנים). כמו כן, הדף צריך לכלול לינק חיצוני לעמוד הויקיפדיה של המדענית.

על קובץ ה-HTML להיות כתוב באנגלית בלבד ותקין מבחינת הסטנדרט של HTML. (ניתן לוודא תקינות באתר http://validator.w3.org).

הגישו את הפתרון בקובץ בשם: question4a.html

(ב) כתבו תוכנית Python שתקבל כקלט כתובת של עמוד אינטרנט (URL). על התוכנית להריץ שאילתות XPath (בעזרת החבילות שראינו בתרגול) ולהדפיס לקובץ את הפרטים הבאים:

a. את כל הלינקים שנמצאים בתוך פסקאות (אלמנט p)
b. לינקים חיצוניים שמובילים לאתרים בדומיין "co.uk".
c. כל המילים המסומנות ב-BOLD בטבלה הראשונה
d. את ה-attribute alt של כל התמונות כך שה-alt מכיל את המחרוזת "Tel Aviv"

הגישו את הפתרון בקובץ בשם: question4b.py

(ג) הריצו את התוכנית מסעיף (ב) על העמוד: https://en.wikipedia.org/wiki/Tel_Aviv
הגישו פלט התוכנית בקובץ בשם: question4c.txt

שאלה 5

בשאלה זו נעסוק בעצי XML שבהם התוויות על העלים הן: GREEN, RED, BLUE. התוויות על כל הצמתים הפנימיים הן UNDEF . על עץ כזה נאמר שהוא ניתן לצביעה אם ניתן להקצות לכל צומת פנימי את אחד מ 3-הצבעים, כך שיתקיים שצבעו של כל צומת שונה מצבעם של כל ילדיו (המידיים).

תהי L שפת כל העצים (מדרגה לא חסומה) הניתנים לצביעה.

1. כתבו בפירוט ובעזרת ביטויים רגולריים, אוטומט מטה-מעלה (bottom-up), לא בהכרח דטרמיניסטי, שמקבל את השפה L בדיוק. שימו לב שעל האוטומט להיות unranked.
2. נגביל כעת את השפה לעצים טרינאריים בלבד (בדיוק 3 ילדים לכל צומת פנימי) נקרא לשפה החדשה L2 . האם ניתן לזהות את L2 ע"י אוטומט מעלה-מטה (top-down) לא דטרמיניסטי? הוכיחו (ע"י בנייה מפורשת של האוטומט) או הפריכו.
3. האם ניתן לזהות את L2 ע"י אוטומט מעלה-מטה (top-down) דטרמיניסטי? הוכיחו (ע"י בניית האוטומט) או הפריכו.

בהצלחה!