Institute of Space Technology, Islamabad

BS CS-01

Compiler Construction

Assignment 03

Due Date: Saturday, 2023

Submitted by: Faraz Ahmad Qureshi

Regno: (200901045)

Section: B

Instructor: Reeda Saeed

Assignment:

In this assignment you are required to parse an xml file (provided in attachment) file using python or any language of your preference. The data to be extracted from file includes:

1. Book_Id
2. Author_Name
3. Title
4. Genre
5. Price
6. Publish_date
7. Description

After extracting the above information from the tags, you have to create an excel file to record the data in the following table format.

Book_ID	Author	Title	Genre	Price	PublishDate	Description
Bk101	-	-	-	-	-	-
Bk102	-	-	-	-	-	-

Solution:

To solve this problem, we will be using Python. Python provides built-in support for many XML parsers. The module we will be using is the *xml.dom.minidom* module. This is a minimal version of the standard DOM (Document Object Model) interface which was used for parsing XML/HTML documents on the WWW (World Wide Web).

To use the module we need to simply import the parser from *minidom* and pass our XML file to it. Afterwards, we can use the DOM object to access different tags in the XML file and pass our output to the excel file accordingly.

Reasons for using minidom module:

One downside of using minidom is that it is maintains an inefficient python list structure when building the DOM object. However, as we are working with a small XML file and will only be building the DOM object once, we do not need efficient creation of the object. If however, we were to work with constantly updating XML files, we can use another XML parser such as Element Tree or the SAX (Simple API for XML) Parser for Python.

Libraries used:

We will be using the *Path* module to get relative path to our files. *csv* module to generate the csv file. We cannot directly generate excel(xlsx) files without external modules in python therefore I opted for a csv file which is close enough to an xlsx file.

Program Code:

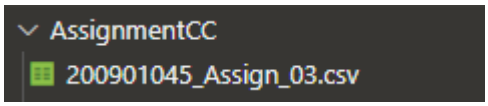
```
from pathlib import Path    #to resolve absolute path to the file.
from xml.dom.minidom import parse #minidom is subset of the DOM (Document Object Model)
library
import csv  #make csv file as we cannot make xlsx files without external modules.

print("[1]Trying to open \'compiler.xml\' file...")
relative_path = Path(__file__).parent #calculate the exact path to file
absolutePath = relative_path.resolve() / 'compiler.xml'
xmlObject = open(absolutePath,"r") #open the compiler.xml file
print("[2]File opened successfully...")
csvFile = open(relative_path.resolve() / '200901045_Assign_03.csv', 'w') #initialize our
csv file
csvFieldNames = ['bookID','author','title','genre','price','publish_date','description']

csvInput = csv.DictWriter(csvFile, fieldnames=csvFieldNames)
csvInput.writeheader()

print("[3]Parsing the XML file...")
domObject = parse(xmlObject)
books = domObject.getElementsByTagName("book")
print("[4]There are %d Books in the xml file." % books.length)
print("[5]Extracting following data:",csvFieldNames)

authorList=[]
titleList=[]
genreList=[]
priceList=[]
publish_dateList=[]
descriptionList=[]
bookidList=[]
for book in books:
    bookidList.append(book.getAttribute('id'))
    authorList.append(book.getElementsByTagName("author")[0].firstChild.data)
    titleList.append(book.getElementsByTagName("title")[0].firstChild.data)
    genreList.append(book.getElementsByTagName("genre")[0].firstChild.data)
    priceList.append(book.getElementsByTagName("price")[0].firstChild.data)
    publish_dateList.append(book.getElementsByTagName("publish_date")[0].firstChild.data)
    descriptionList.append(book.getElementsByTagName("description")[0].firstChild.data)

for i in range(0,len(books)):
    csvInput.writerow({'bookID':bookidList[i],'title': titleList[i], 'author':
authorList[i],'genre':genreList[i],'price':priceList[i],'publish_date':publish_dateList[i],
'description':descriptionList[i]})

print('[6]Output has been stored to csv file!')
```

As the code is self-explanatory, we can see the output of our program to be:

Output:

```
TERMINAL
(apis) PS C:\Users\faraz\PythonWorkspace> & C:/Users/faraz/PythonWorkspace/testing/cmdplayer/apis/Scripts/python.ex
/faraz/PythonWorkspace/AssignmentCC/200901045_Assign3.py
[1]Trying to open 'compiler.xml' file...
[2]File opened successfully...
[3]Parsing the XML file...
[4]There are 12 Books in the xml file.
[5]Extracting following data: ['bookID', 'author', 'title', 'genre', 'price', 'publish_date', 'description']
[6]Output has been stored to csv file!
(apis) PS C:\Users\faraz\PythonWorkspace>
```

After successful run of the program, we also get the csv file as an output:

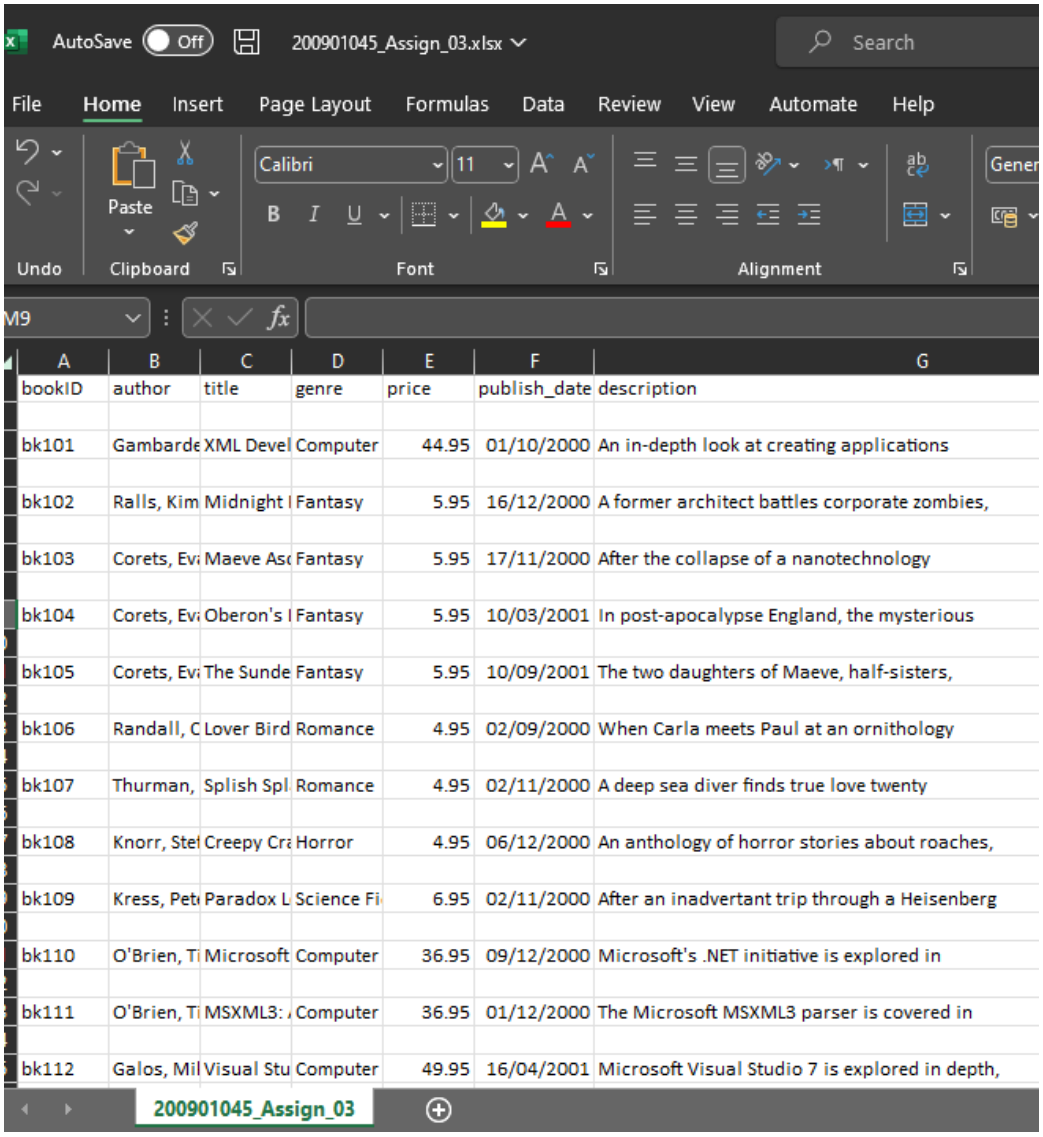


Raw Contents of this file are:

AssignmentCC > 200901045_Assign_03.csv

```
1   bookID,author,title,genre,price,publish_date,description
2
3   bk101,"Gambardella, Matthew",XML Developer's Guide,Computer,44.95,2000-10-01,"An in-depth look at creati
4         with XML."
5
6   bk102,"Ralls, Kim",Midnight Rain,Fantasy,5.95,2000-12-16,"A former architect battles corporate zombies,
7         an evil sorceress, and her own childhood to become queen
8         of the world."
9
10  bk103,"Corets, Eva",Maeve Ascendant,Fantasy,5.95,2000-11-17,"After the collapse of a nanotechnology
11        society in England, the young survivors lay the
12        foundation for a new society."
13
```

If we open this file in Excel, we get:



bookID	author	title	genre	price	publish_date	description
bk101	Gambarde	XML Devel	Computer	44.95	01/10/2000	An in-depth look at creating applications
bk102	Ralls, Kim	Midnight I	Fantasy	5.95	16/12/2000	A former architect battles corporate zombies,
bk103	Corets, Evi	Maeve Asc	Fantasy	5.95	17/11/2000	After the collapse of a nanotechnology
bk104	Corets, Evi	Oberon's I	Fantasy	5.95	10/03/2001	In post-apocalypse England, the mysterious
bk105	Corets, Evi	The Sunde	Fantasy	5.95	10/09/2001	The two daughters of Maeve, half-sisters,
bk106	Randall, C	Lover Bird	Romance	4.95	02/09/2000	When Carla meets Paul at an ornithology
bk107	Thurman,	Splish Spl	Romance	4.95	02/11/2000	A deep sea diver finds true love twenty
bk108	Knorr, Ste	Creepy Cr	Horror	4.95	06/12/2000	An anthology of horror stories about roaches,
bk109	Kress, Pet	Paradox L	Science Fi	6.95	02/11/2000	After an inadvertant trip through a Heisenberg
bk110	O'Brien, Ti	Microsoft	Computer	36.95	09/12/2000	Microsoft's .NET initiative is explored in
bk111	O'Brien, Ti	MSXML3:	Computer	36.95	01/12/2000	The Microsoft MSXML3 parser is covered in
bk112	Galos, Mil	Visual Stu	Computer	49.95	16/04/2001	Microsoft Visual Studio 7 is explored in depth,

200901045_Assign_03