Knowledge Discovery and Data Mining

Lab 12 XML Analysis

Xuan Song
Songx@sustech.edu.cn

Topics

Analyze XML files with python

XML (eXtensible Markup Language)

XML (standard)

Extensible Markup Language

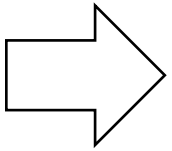
Status	Published, W3C Recommendation
Year started	1996; 25 years ago
First published	February 10, 1998; 23 years ago As a Recommendation
Latest version	1.1 (Second Edition) September 29, 2006; 14 years ago
Organization	World Wide Web Consortium (W3C)
Editors	Tim Bray • Jean Paoli • C. M. Sperberg-McQueen • Eve Maler • François Yergeau • John Cowan
Base standards	SGML
Related standards	XML Schema
Domain	Data serialization
Abbreviation	XML
Website	www.w3.org/xml

XML (file format)

Filename extension	`.xml`
Internet media type	`application/xml` `text/xml` [1]
Uniform Type Identifier (UTI)	public.xml
UTI conformation	public.text
Magic number	`<?xml`
Developed by	World Wide Web Consortium
Type of format	Markup language
Extended from	SGML
Extended to	Numerous languages, including XHTML • RSS • Atom • KML
Standard	1.0 (Fifth Edition) (November 26, 2008; 12 years ago) 1.1 (Second Edition) (August 16, 2006; 14 years ago)
Open format?	Yes

XML Document

Category: CHILDREN
Title: Harry Potter
Author: J K. Rowling
Year: 2005
Price: 29.99

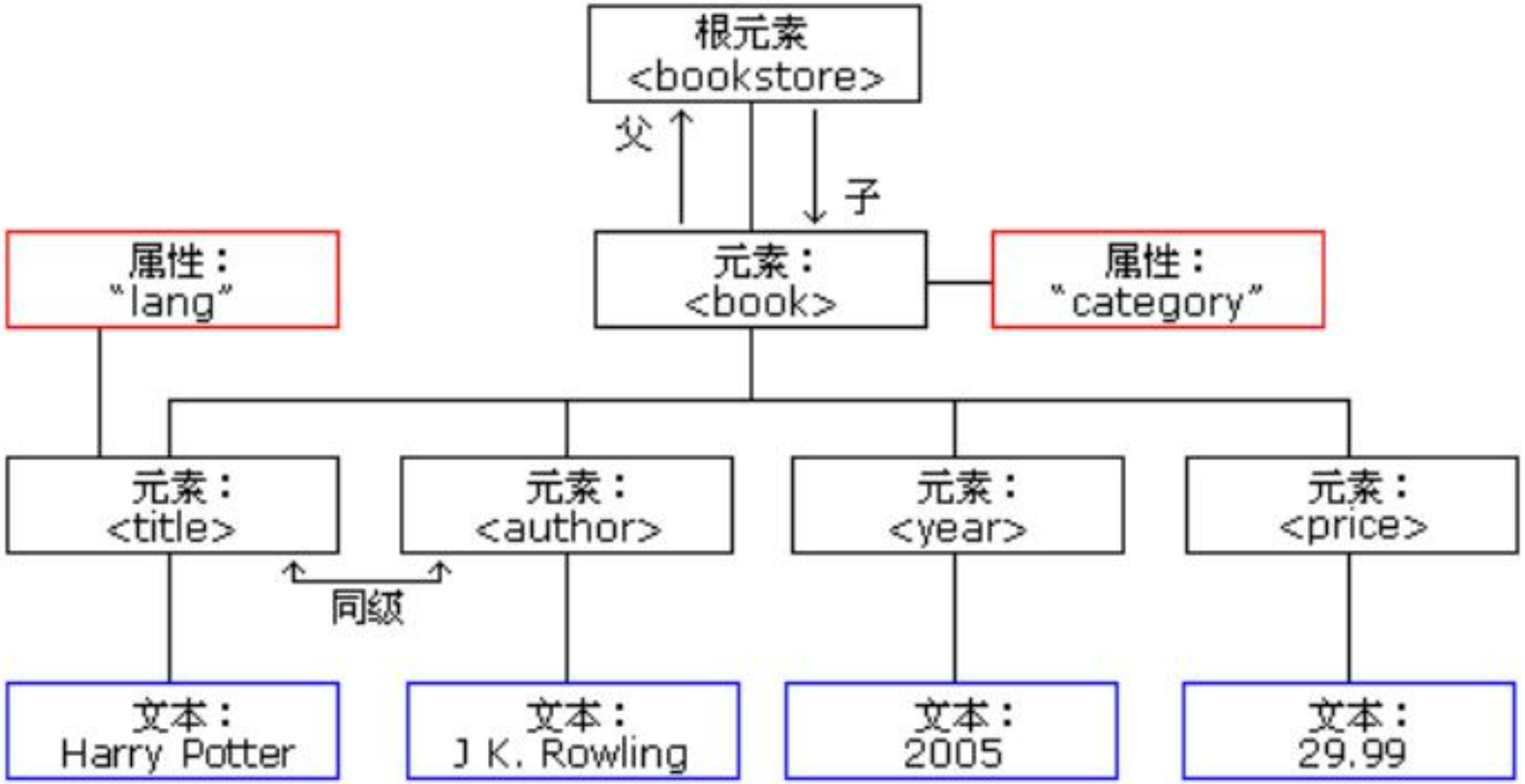


```
<bookstore>
<book category="CHILDREN">
  <title lang="en">Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
</bookstore>
```

DATA

XML Format

XML Document



DOM & ElementTree

- xml.dom.minidom

  xml.dom.minidom.parse(filename_or_file, parser=None, bufsize=None)

- xml.etree.ElementTree

  xml.etree.ElementTree.parse(filename_or_file, parser=None)

https://docs.python.org/3/library/xml.dom.minidom.html
https://docs.python.org/3/library/xml.etree.elementtree.html

XML Analysis with Python

- Example: cd.xml

```
<CATALOG>
<CD>
<TITLE>Empire Burlesque</TITLE>
<ARTIST>Bob Dylan</ARTIST>
<COUNTRY>USA</COUNTRY>
<COMPANY>Columbia</COMPANY>
<PRICE>10.90</PRICE>
<YEAR>1985</YEAR>
</CD>
<CD>
<TITLE>Hide your heart</TITLE>
<ARTIST>Bonnie Tyler</ARTIST>
<COUNTRY>UK</COUNTRY>
<COMPANY>CBS Records</COMPANY>
<PRICE>9.90</PRICE>
<YEAR>1988</YEAR>
</CD>
</CATALOG>
```

DOM

- 1: Parse xml file

```
from xml.dom.minidom import parse
DOMTree = parse("./cd.xml")  # parse an XML file by name

datasource = open(./cd.xml ‘)
DOMTree = parse(datasource)        # parse an open file
```

- 2: Get the document element

```
CATALOG = DOMTree.documentElement
cds = CATALOG.getElementsByTagName("CD")
```

DOM

- 3: get elements by tag name

```
for cd in cds:
    TITLE = cd.getElementsByTagName('TITLE')[0]
    print("TITLE: %s"%TITLE.getAttribute("TITLE"))

     ## you can get other elements ##
```

Element Tree

- 1: Parse xml file

```
import xml.etree.ElementTree as ET
tree = ET.ElementTree(file='./cd.xml')
```

- 2: Get the root

```
root = tree.getroot() #获取根节点
print(root)
```

Element Tree

- 3: Traversal to get all the elements

```
for cd in root:
    TITLE = cd.find('TITLE').text
    print("TITLE: ",TITLE)

    ## you can get other elements ##
```

Task1

- Based on the given file *plant_catalog.xml,* use **both aforementioned methods** to read the data and save them as a csv file.

End of Lab12