# WOW

November 18, 2021

Started ‘Python 3.9.5 64-bit’ kernel
Python 3.9.5 (v3.9.5:0a7dcbdb13, May 3 2021, 13:17:02)
Type ‘copyright’, ‘credits’ or ‘license’ for more information
IPython 7.29.0 – An enhanced Interactive Python. Type ‘?’ for help.

## Python XML parsing

### Using xml.dom

[ ]:

```
from xml.dom.minidom import parse, parseString, Node
```

[ ]:

```
# opening it
document = parse("sample.xml")
print(document)
```

```
<xml.dom.minidom.Document object at 0x7fb4502c6640>
```

[ ]:

```
# using context manager
with open("sample.xml") as file:
    document = parse(file)
print(document)
```

```
<xml.dom.minidom.Document object at 0x7fb4902f1c40>
```

[ ]:

```
# parsing it through a string
document = parseString('''<?xml version="1.0"?>
<catalog>
   <book id="bk101">
      <author>Gambardella, Matthew</author>
      <title>XML Developer's Guide</title>
      <genre>Computer</genre>
      <price>44.95</price>
      <publish_date>2000-10-01</publish_date>
      <description>An in-depth look at creating applications
      with XML.</description>
   </book>
   </catalog>''')
print(document)
```

```
<xml.dom.minidom.Document object at 0x7fb4502d5940>
```

#### Accessing info from XML

[ ]:

```
# It is able to return info such as version and DTD
document = parse("sample.xml")
print(document.version)
print(document.doctype)
print(document.documentElement)
```

```
1.0
None
<DOM Element: catalog at 0x7fb4502d6dc0>
```

[ ]:

```
# However it can't parse elements
document = parse("sample.xml")
print(document.getElementById("bk101"))
print(document.getElementById("bk102"))
```

```
None
None
```

#### Solve this issue by giving all elements an id attribute

[ ]:

```
def set_id_attribute(parent, attribute_name="id"):
    if parent.nodeType == Node.ELEMENT_NODE:
        if parent.hasAttribute(attribute_name):
            parent.setIdAttribute(attribute_name)
    for child in parent.childNodes:
        set_id_attribute(child, attribute_name)
```

[ ]:

```
set_id_attribute(document)
print(document.getElementById("bk101"))
print(document.getElementById("bk102"))
```

```
<DOM Element: book at 0x7fb4502d3dc0>
<DOM Element: book at 0x7fb4502dd5e0>
```

[ ]:

```
document = parse("smiley.svg")
set_id_attribute(document)
print(document.getElementById("smiley"))
print(document.getElementsByTagName("ellipse"))
```

```
<DOM Element: g at 0x7fb4502ddb80>
[<DOM Element: ellipse at 0x7fb4502ddf70>, <DOM Element: ellipse at
0x7fb4502e8040>]
```

#### Bad News

[ ]:

```
try:
    print(document.querySelector("#smiley"))
```

```python
except AttributeError:
    print("does not work")
```

```
does not work
```

#### for stuff like

[ ]:
```python
document.getElementsByTagNameNS("*", "custom")
```

[ ]:
```
[<DOM Element: inkscape:custom at 0x7fb4502d3ee0>]
```

[ ]:
```python
# other stuff here
with open("smiley.svg") as file:
    document = parse(file)
```

### Using xml.sax

[ ]:
```python
import xml.sax

class ParseXML(xml.sax.ContentHandler):

    def __init__(self):
      self.CurrentData = ""
      self.author = ""
      self.title = ""
      self.genre = ""
      self.price = ""
      self.publish_date = ""
      self.description = ""

    def startElement(self, tag, attributes):
        self.CurrentData = tag
        if tag == "book":
            print("--------Book--------")
            book_id = attributes["id"]
            print(f"Id: {book_id}")

    def endElement(self, tag):
        if self.CurrentData == "title":
            print(f"Title: {self.title}")
        elif self.CurrentData == "author":
            print(f"Author: {self.author}")
        elif self.CurrentData == "genre":
            print(f"genre: {self.genre}")
        elif self.CurrentData == "price":
            print(f"price: {self.price}")
        elif self.CurrentData == "publish_date":
            print(f"publish_date: {self.publish_date}")
        elif self.CurrentData == "description":
            print(f"description: {self.description}")

    def characters(self, content):
        if self.CurrentData == "title":
            self.title = content
        elif self.CurrentData == "author":
            self.author = content
        elif self.CurrentData == "genre":
            self.genre = content
        elif self.CurrentData == "price":
            self.price = content
        elif self.CurrentData == "publish_date":
            self.publish_date = content
        elif self.CurrentData == "description":
            self.description = content


parser = xml.sax.make_parser()
parser.setFeature(xml.sax.handler.feature_namespaces, 0)
parser_object = ParseXML()
parser.setContentHandler(parser_object)
parser.parse("sample.xml")

""
```

```
--------Book--------
Id: bk101
Author: Gambardella, Matthew
Title: XML Developer's Guide
genre: Computer
price: 44.95
publish_date: 2000-10-01
description: An in-depth look at creating applications with XML.
description:
--------Book--------
Id: bk102
Author: Ralls, Kim
Title: Midnight Rain
genre: Fantasy
price: 5.95
publish_date: 2000-12-16
description: A former architect battles corporate zombies, an evil sorceress,
and her own childhood to become queen of the world.
description:
description:
```

[ ]: ''

…

#### Simplified version

[ ]:

```
import xml.sax

class ParseXML(xml.sax.ContentHandler):

    def __init__(self):
      self.CurrentData = ""

    def startElement(self, tag, attributes):
        self.CurrentData = tag
        if tag == "book":
            print("Book")
            book_id = attributes["id"]
            print(f"Id: {book_id}")

    def endElement(self, tag):
        print(f"{self.CurrentData}: {self.content}")

    def characters(self, content):
        self.content = content

parser = xml.sax.make_parser()
parser.setFeature(xml.sax.handler.feature_namespaces, 0)
parser_object = ParseXML()
parser.setContentHandler(parser_object)
parser.parse("sample.xml")

""
```

```
Book
Id: bk101
author: Gambardella, Matthew
title: XML Developer's Guide
genre: Computer
price: 44.95
publish_date: 2000-10-01
description: An in-depth look at creating applications with XML.
description:
Book
Id: bk102
author: Ralls, Kim
title: Midnight Rain
genre: Fantasy
```

```
price: 5.95
publish_date: 2000-12-16
description: A former architect battles corporate zombies, an evil sorceress,
and her own childhood to become queen of the world.
description:
description:
```

```
[ ]: ''
```