

World Class Machine Learning

Machine Learning Hub

Menu

ABOUT US

Menu

Python XML Parsing Example

Posted on June 9, 2018 by Jonathan Steele

Hello Everyone!

Today, I want to share how to parse through XML files in Python. We will use a simple example and then use a real-life example from my experience.

Resources

1. General Information about XML ([W3 Schools](#))
2. Tutorial I used for Python XML DOM parsing ([TutorialsPoint](#))
3. Setting up a Python Setup ([ML Hello World](#))
4. Example files ([W3 Schools](#))
 1. [Direct Note Link](#) (Alt-Click link to Download)
 2. [Direct CD Catalog Link](#) (Alt-Click link to Download)

Trivial Example

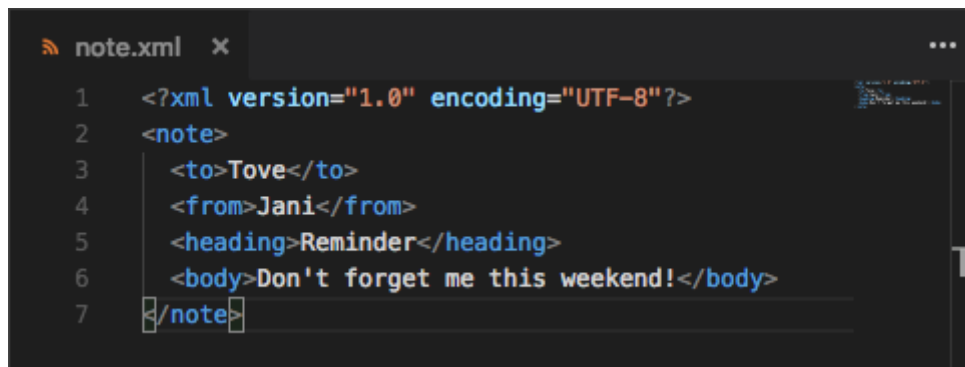
I'm assuming you have some things figured out already:

1. You know about XML generally (If not, follow Resource Link 1)

2. You already have Python working on your Computer (If not, follow Resource Link 3)

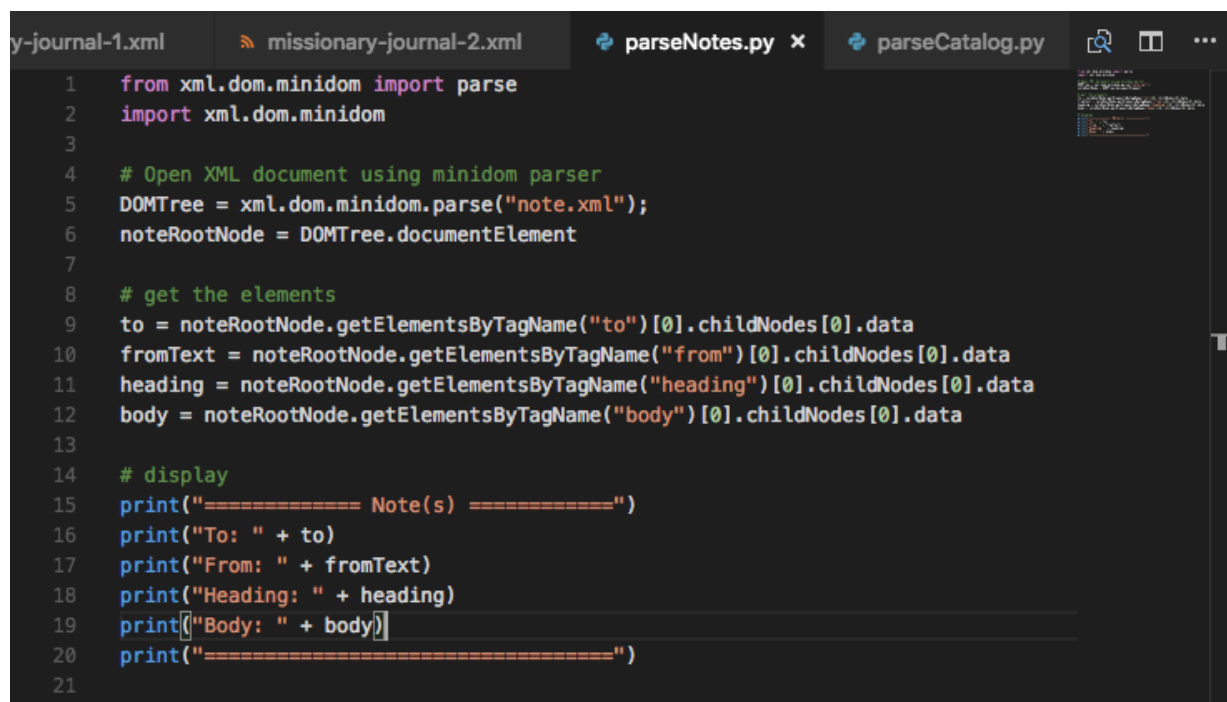
Now, let's start.

- Download the sample XML files from Resource Link 4.1 and 4.2.
- Get those sample XML files in a folder by themselves named "xml_example" or something like that.
- View the XML file to get a feel for what it's like.



```
note.xml x
1  <?xml version="1.0" encoding="UTF-8"?>
2  <note>
3    <to>Tove</to>
4    <from>Jani</from>
5    <heading>Reminder</heading>
6    <body>Don't forget me this weekend!</body>
7  </note>
```

- In that same folder, create a file named "parseNotes.py" and open it in a text editor that you prefer.
- Enter in this code: This will be a little rote your first time, but we will use it a little differently in a second and you can always examine the documentation and other tutorials to get a deeper feel.



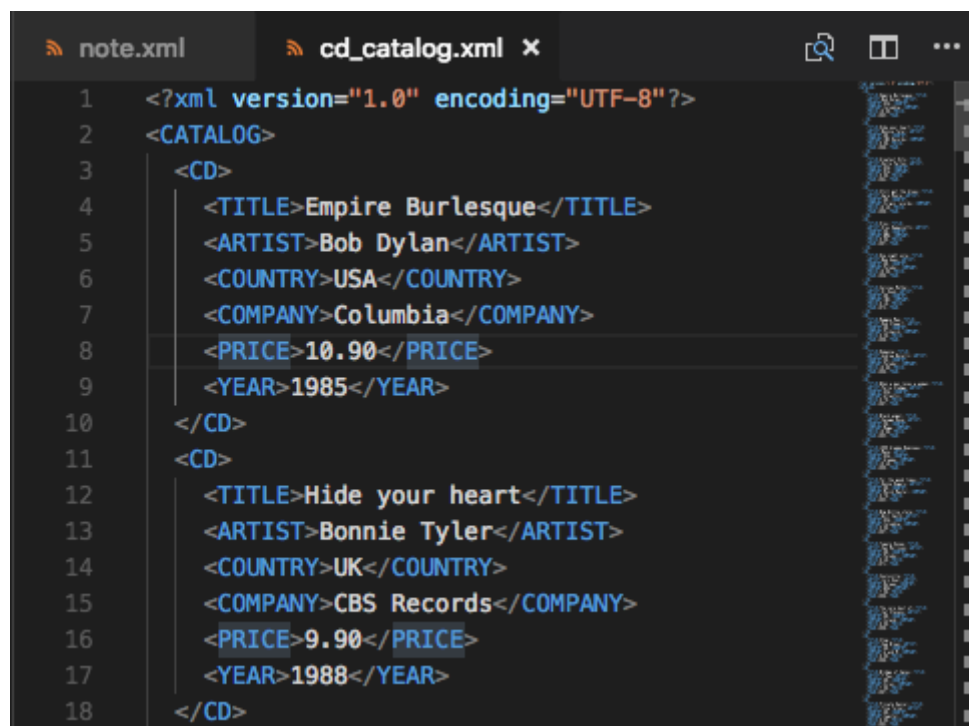
```
y-journal-1.xml  missionary-journal-2.xml  parseNotes.py x  parseCatalog.py
1  from xml.dom.minidom import parse
2  import xml.dom.minidom
3
4  # Open XML document using minidom parser
5  DOMTree = xml.dom.minidom.parse("note.xml");
6  noteRootNode = DOMTree.documentElement
7
8  # get the elements
9  to = noteRootNode.getElementsByTagName("to")[0].childNodes[0].data
10 fromText = noteRootNode.getElementsByTagName("from")[0].childNodes[0].data
11 heading = noteRootNode.getElementsByTagName("heading")[0].childNodes[0].data
12 body = noteRootNode.getElementsByTagName("body")[0].childNodes[0].data
13
14 # display
15 print("===== Note(s) =====")
16 print("To: " + to)
17 print("From: " + fromText)
18 print("Heading: " + heading)
19 print("Body: " + body)
20 print("=====")
21
```

Intermediate Example

If you've skipped the Trivial Example and jumped to the Intermediate, make sure to grab the "cd_catalog.xml" file from Resource Link 4.2 and get your stuff in a project folder.

Let's go!

- Create a file name "parseCatalog.py"
- Add the necessary imports, as seen above.
- Read in the file "cd_catalog.xml" using MiniDom, as seen above.
- View the XML file to get a feel for how we will need to parse it.



```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <CATALOG>
3   <CD>
4     <TITLE>Empire Burlesque</TITLE>
5     <ARTIST>Bob Dylan</ARTIST>
6     <COUNTRY>USA</COUNTRY>
7     <COMPANY>Columbia</COMPANY>
8     <PRICE>10.90</PRICE>
9     <YEAR>1985</YEAR>
10  </CD>
11  <CD>
12    <TITLE>Hide your heart</TITLE>
13    <ARTIST>Bonnie Tyler</ARTIST>
14    <COUNTRY>UK</COUNTRY>
15    <COMPANY>CBS Records</COMPANY>
16    <PRICE>9.90</PRICE>
17    <YEAR>1988</YEAR>
18  </CD>
```

- As you can see, there is a root node "CATALOG" that has a list of "CD"s
 - and each "CD" has a
 - TITLE
 - ARTIST
 - COUNTRY
 - COMPANY
 - PRICE
 - YEAR
- So let's make sure to parse our stuff in that way. But let's add one thing. Let's use a for loop to access the list of "CD"s so that we can read in as many or as few CDs as there are in the Catalog.
- Test out a For Loop on your CDs list by using code kind of like this:

```

1  from xml.dom.minidom import parse
2  import xml.dom.minidom
3
4  # Open XML document using minidom parser
5  DOMTree = xml.dom.minidom.parse("cd_catalog.xml")
6  catalogRootNode = DOMTree.documentElement
7
8  # get the list
9  cds = catalogRootNode.getElementsByTagName("CD")
10 numOfRows = 0
11 for cd in cds:
12     numOfRows += 1
13
14 print("The Catalog has " + str(numOfRows) + " CDs in it. Counted via a FOR Loop.")
15 print("The Catalog has " + str(len(cds)) + " CDs in it. Counted via a LEN command.")
16
17
18
19
20
21

```

- Once you've figured out how to use the For Loops for parsing the XML lists, then you're ready to look at the whole list.

```

1  from xml.dom.minidom import parse
2  import xml.dom.minidom
3
4  # Open XML document using minidom parser
5  DOMTree = xml.dom.minidom.parse("cd_catalog.xml")
6  catalogRootNode = DOMTree.documentElement
7
8  print("===== CD(s) =====")
9  # get the list
10 cds = catalogRootNode.getElementsByTagName("CD")
11 for cd in cds:
12     # collect data
13     title = cd.getElementsByTagName("TITLE")[0].childNodes[0].data
14     artist = cd.getElementsByTagName("ARTIST")[0].childNodes[0].data
15     country = cd.getElementsByTagName("COUNTRY")[0].childNodes[0].data
16     company = cd.getElementsByTagName("COMPANY")[0].childNodes[0].data
17     price = cd.getElementsByTagName("PRICE")[0].childNodes[0].data
18     year = cd.getElementsByTagName("YEAR")[0].childNodes[0].data
19
20     # display
21     print("=====")
22     print("Title: " + title)
23     print("Artist: " + artist)
24     print("Country: " + country)
25     print("Company: " + company)
26     print("Price: " + price)
27     print("Year: " + year)
28
29 print("=====")
30
31

```

- And that's all there is to it! You've parse the whole XML file!

Conclusion

So far, we've done a couple trivial examples of how to parse XML files.

Next time, we can have a real-life example for parsing and sorting a journal file that was kept in XML. Until then!

LANGUAGE

English ▾

RECENT POSTS

Block Chain Killer Apps – Article Review

What is the Difference between MAR and MCAR?

Review of Andrew Ng's Machine Learning Course

F Scores

Single Numeric Evaluation Metric

CATEGORIES

Blockchain (1)

Chinese (12)

Chinese Programming Words (12)

Machine Learning (13)

Ideas for ML Side Projects (6)

Software Development (2)

Intro to Python (2)

Uncategorized (2)

PAGES

About Us

 English

© 2019 World Class Machine Learning | WordPress Theme by Superbthemes