Analyzing Structured Data in Python

Week 2: ElementTree

23 October 2020

CDCS Python Course Series

with Lucy Havens

Course Topics

Week 1: Pandas for CSV data

Week 2: ElementTree for XML data

Assignment

Watch videos *1.2. Quick Overview of XML* and *6.3. The ElementTree API* in the course *Python: XML, JSON, and the Web*

https://www.linkedin.com/learning/python-xml-json-and-the-web/quick-overview-of-xml?u=50251009

Find or create your own XML file to parse and analyze with ElementTree! What questions can you ask about it using the methods and functions in ElementTree? Can you transform the XML into another data format and put it into a Pandas DataFrame?

How did it go?

What datasets did you use?

XML Data: A Quick Recap

E**x**tensible **M**arkup **L**anguage

For storing and sending data (or metadata!)

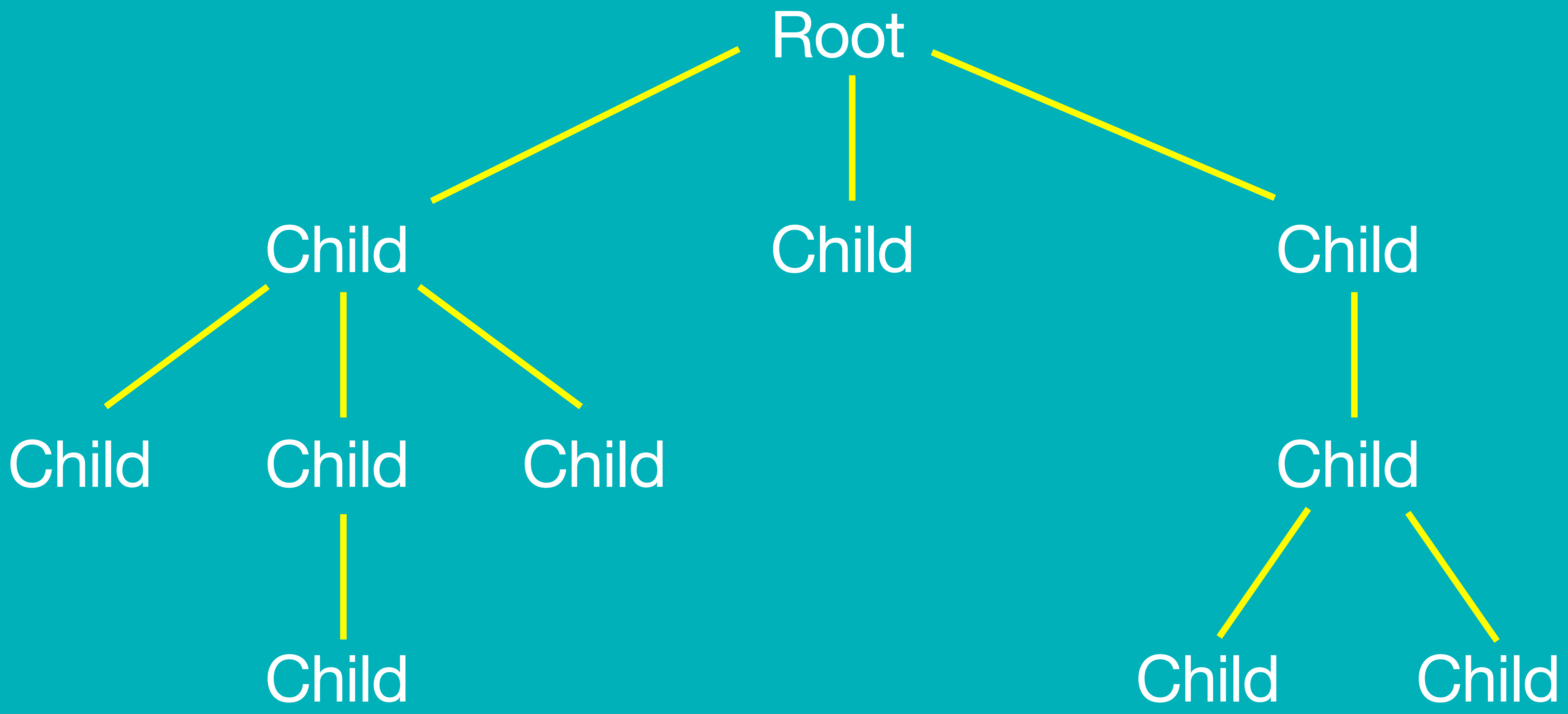
Hierarchical

Reference: https://www.datacamp.com/community/tutorials/python-xml-elementtree#intro

How does XML differ from HTML?

HTML has information about how to format, or display, data, and **XML does not.**

XML Data: A Quick Recap



XML Data: A Quick Recap

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
      <movie title="Indiana Jones: The raiders of the lost Ark">
        <format multiple="No">DVD</format>
        <year>1981</year>
        <rating>PG</rating>
        <description>
                'Archaeologist and adventurer Indiana Jones is hired
                by the U.S. government to find the Ark of the
                Covenant before the Nazis.'
        </description>
      </movie>
 </genre>
</collection>
```

XML Data: A Quick Recap

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
      <movie title="Indiana Jones: The raiders of the lost Ark">
        <format multiple="No">DVD</format>
        <year>1981</year>
        <rating>PG</rating>
        <description>
                'Archaeologist and adventurer Indiana Jones is hired
                by the U.S. government to find the Ark of the
                Covenant before the Nazis.'
        </description>
      </movie>
 </genre>
</collection>
```

XML Data: A Quick Recap

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
      <movie title="Indiana Jones: The raiders of the lost Ark">
        <format multiple="No">DVD</format>
        <year>1981</year>
        <rating>PG</rating>
        <description>
                'Archaeologist and adventurer Indiana Jones is hired
                by the U.S. government to find the Ark of the
                Covenant before the Nazis.'
        </description>
      </movie>
 </genre>
</collection>
```

XML Data: A Quick Recap

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
      <movie title="Indiana Jones: The raiders of the lost Ark">
        <format multiple="No">DVD</format>
        <year>1981</year>
        <rating>PG</rating>
        <description>
                'Archaeologist and adventurer Indiana Jones is hired
                by the U.S. government to find the Ark of the
                Covenant before the Nazis.'
        </description>
      </movie>
  </genre>
</collection>
```

ElementTree: A Quick Recap

- API for parsing and exploring XML data in Python
- Useful methods include:

```
.getroot()

.iter()

.iter(“TAG-NAME”)

.findall(“TAG-NAME”)

.findall(“PATH/TO/TAG/OR/ATTRIBUTE”)
```

Creating XML Data

- Build a tree of nodes
  - Manually, one by one
  - From lists
- Modify XML elements
- Export XML data (write to a file)

Reference: https://pymotw.com/2/xml/etree/ElementTree/create.html

DEMO

Thank you! Feedback?

For more intermediate Python courses, checkout the upcoming course on text analysis with NLTK!

More registration places have been opened