Analyzing Structured Data in Python

WEEK 1: ELEMENTTREE

21 March 2022

CDCS Python Course Series

Instructor: Lucy Havens

COURSE TOPICS

Week 1: Pandas for CSV data

Week 2: ElementTree for XML data

Please let me know if there are specific topics you'd like to cover!

Raise your hand if...

You know what XML data looks like

RAISE YOUR HAND IF...

You know what XML data looks like

You’ve worked with XML data

Raise your hand if…

You know what XML data looks like

You’ve worked with XML data

You’ve worked with ElementTree

WHAT IS XML?

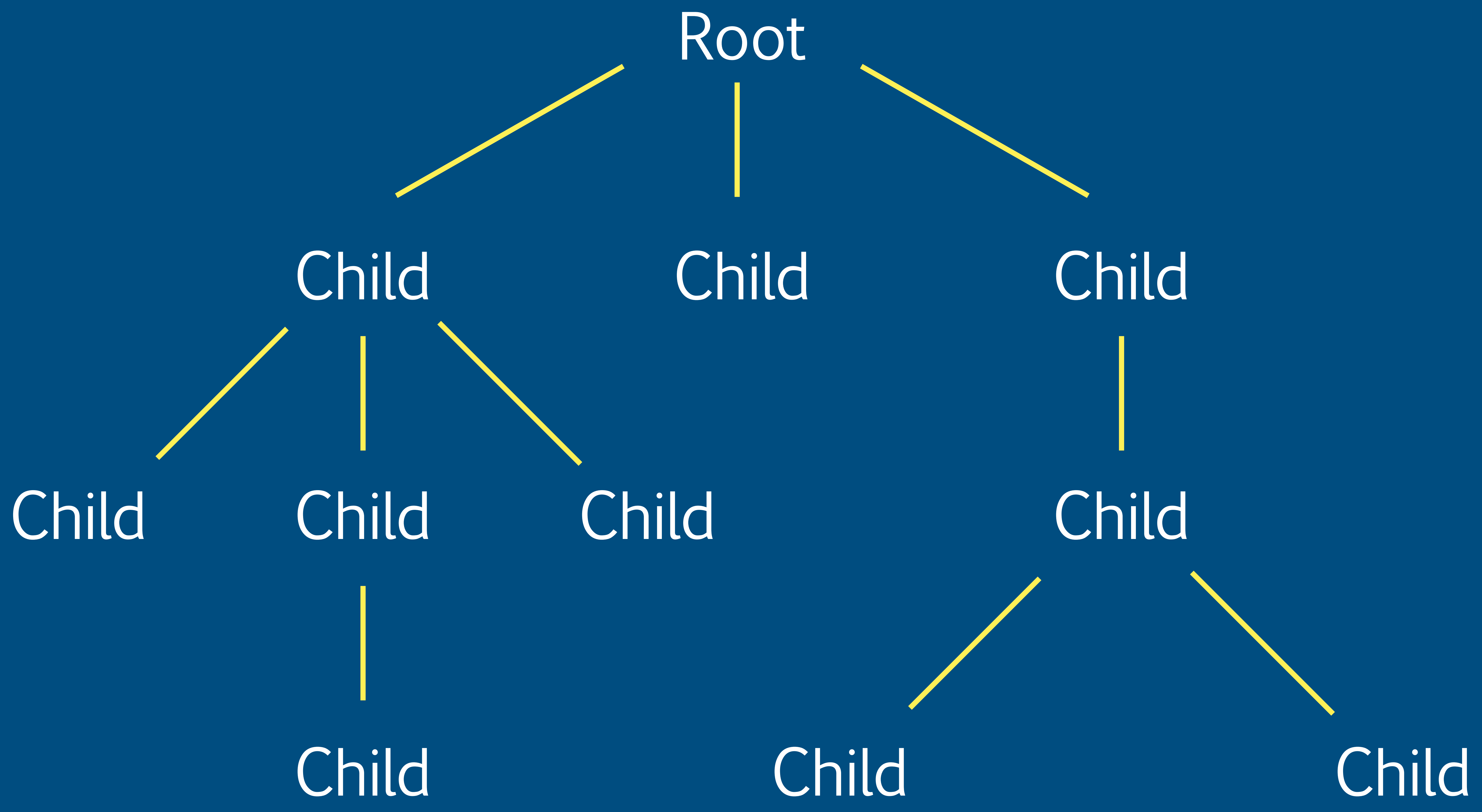
E**x**tensible **M**arkup **L**anguage

Similar to HTML (**H**ypertext **M**arkup **L**anguage)

Organizes data **hierarchically**

Reference: DataCamp's Python XML Tutorial with ElementTree: Beginner's Guide

XML VISUALIZED



Root

Child

Child

Child

Child

Child

Child

Child

Child

Child

Child

XML EXAMPLE

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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>
```

Reference: [*https://data.mendeley.com/datasets/rth2kr5hxf/2*](https://data.mendeley.com/datasets/rth2kr5hxf/2)

XML Example

What's the root?

```xml
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 EXAMPLE

COLLECTION!

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example

Can you name a child?

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example

Genre, movie, format…

```xml
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 EXAMPLE

IF THIS IS A START TAG, WHERE'S THE END TAG?

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example: Tags

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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>
```

Here's the end tag!

XML Example: Tags

```xml
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example: Element

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example: Another Element

```xml
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example: Attribute

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 EXAMPLE: ATTRIBUTE

CAN YOU NAME OTHERS?

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example: Attribute

Category, Favorite, Title

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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 Example: Text

```
<?xml version="1.0"?>
<collection>
  <genre category="Action">
    <movie favorite="True" 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>
```

LET'S CODE!

XML in the Real World

- Web publishing
- Business applications (sending data between different technology systems)
- Digital metadata formats
- Harvesting data through APIs
- Downloading data dumps

Reference: https://www.ibm.com/docs/en/i/7.3?topic=introduction-uses-xml

ASSIGNMENT

Watch the videos below from LinkedIn Learning's "Python: XML, JSON, and the Web" course:

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

6.3 The ElementTree API:
https://www.linkedin.com/learning/python-xml-json-and-the-web/the-elementtree-api?u=50251009

Complete the following online tutorial in your own Jupyter Notebook:

Turn XML data into CSV data: https://www.geeksforgeeks.org/xml-parsing-python/

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 extract some of the XML data and put it in a DataFrame using Pandas?

If you're not sure where to find XML data, you could try one of these:

- *https://www.oldbaileyonline.org/browse.jsp?foo=bar&path=sessionsPapers/17800628.xml&div=t17800628-33&xml=yes*
- *sample.xml at https://data.mendeley.com/datasets/rth2kr5hxf/2*

THANKS EVERYONE!

Next course meeting: Friday, 10:00-11:00 AM BST

Office hours available on Wednesday

To schedule, please message me on Teams!

Feedback Survey: https://forms.office.com/r/YYNrqvuNr8