Analysing Structured Data with Pandas and ElementTree

Week 2: ElementTree

Documentation: https://docs.python.org/3.5/library/xml.etree.elementtree.html
Introduction: https://www.datacamp.com/community/tutorials/python-xml-elementtree#intro

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 tutorials:
Follow along 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?

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

For an extra challenge, try working MARCXML, Dublin Core XML, or MODS XML data, commonly used in libraries, archives, and museums:
https://data.nls.uk/data/metadata-collections/
https://about.biodiversitylibrary.org/tools-and-services/developer-and-data-tools/

Go Further

Example: Analyze MARCXML (a library metadata standard in XML format) from a file:
https://data.nls.uk/tools/jupyter-notebooks/exploring-the-national-bibliography-of-scotland/

Example: Analyze ISAD(G) XML (an archival metadata standard in XML format) from a URL:
https://github.com/thegoose20/eula41/blob/master/LHavens_NLS-Internship_Archives-Coll41.ipynb

Learning about XML: W3Schools XML Tutorial ("XML Home" through "XML Quiz"):
https://www.w3schools.com/xml/default.asp

Learning about ElementTree: video 13.4 Parsing XML:
https://www.coursera.org/lecture/python-network-data/13-4-parsing-xml-xVcE1

Learning about accessing data online from URLs: section 2. Accessing the Internet (6 videos):
https://www.linkedin.com/learning/python-xml-json-and-the-web/introducing-urllib?u=50251009