Joyce Woznica  
IST 652, Week 6

Activity 6.4: XML, DOM, and Element Tree

Code:

rss1 = 'http://feeds.bbci.co.uk/news/world/rss.xml'

rss2 = 'http://rss.cnn.com/rss/edition\_world.rss'

rss3 = 'https://www.yahoo.com/news/world/rss'

import urllib.request

xmlstring = urllib.request.urlopen(rss1).read().decode('utf8')

len(xmlstring)

xmlstring[:500]

import xml.etree.ElementTree as etree

import io

xmlfile = io.StringIO(xmlstring)

tree = etree.parse(xmlfile)

type(tree)

root = tree.getroot()

type(root)

root.tag

root.attrib

root.text

len(root)

for child in root:

print(child)

firstchild = root[0]

type(firstchild)

firstchild.tag

firstchild.attrib

firstchild.text

len(firstchild)

firstgrandchild = firstchild[0]

firstgrandchild.tag

firstgrandchild.attrib

firstgrandchild.text

len(firstgrandchild)

itemlist = firstchild.findall('item')

len(itemlist)

firstitem = itemlist[0]

firstitem.attrib

firstitem.text

len(firstitem)

for element in firstitem:

print(element.tag, element.attrib, element.text)

firstitem = firstchild.find('item')

firstitem

alltitles = tree.findall('.//title')

len(alltitles)

for title in alltitles[:6]:

print(title.text)

Output:

﻿rss1 = 'http://feeds.bbci.co.uk/news/world/rss.xml'

rss2 = 'http://rss.cnn.com/rss/edition\_world.rss'

import urllib.request

xmlstring = urllib.request.urlopen(rss1).read().decode('utf8')

len(xmlstring)

Out[23]: 14681

xmlstring[:500]

Out[24]: '<?xml version="1.0" encoding="UTF-8"?>\n<?xml-stylesheet title="XSL\_formatting" type="text/xsl" href="/shared/bsp/xsl/rss/nolsol.xsl"?>\n<rss xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:content="http://purl.org/rss/1.0/modules/content/" xmlns:atom="http://www.w3.org/2005/Atom" version="2.0" xmlns:media="http://search.yahoo.com/mrss/">\n <channel>\n <title><![CDATA[BBC News - World]]></title>\n <description><![CDATA[BBC News - World]]></description>\n <link>https://www.bbc'

import xml.etree.ElementTree as etree

import io

xmlfile = io.StringIO(xmlstring)

tree = etree.parse(xmlfile)

type(tree)

Out[25]: xml.etree.ElementTree.ElementTree

root = tree.getroot()

type(root)

Out[27]: xml.etree.ElementTree.Element

root.tag

Out[28]: 'rss'

root.attrib

Out[29]: {'version': '2.0'}

root.text

Out[30]: '\n '

len(root)

Out[31]: 1

for child in root:

print(child)

<Element 'channel' at 0x11b208dd0>

firstchild = root[0]

type(firstchild)

Out[34]: xml.etree.ElementTree.Element

firstgrandchild = firstchild[0]

firstgrandchild.tag

firstgrandchild.attrib

firstgrandchild.text

len(firstgrandchild)

Out[35]: 0

itemlist = firstchild.findall('item')

len(itemlist)

firstitem = itemlist[0]

firstitem.attrib

firstitem.text

len(firstitem)

for element in firstitem:

print(element.tag, element.attrib, element.text)

title {} Coronavirus: US to borrow record $3tn as spending soars

description {} The sum is more than five times the previous quarterly record, in the 2008 financial crisis.

link {} https://www.bbc.co.uk/news/business-52537938

guid {'isPermaLink': 'true'} https://www.bbc.co.uk/news/business-52537938

pubDate {} Mon, 04 May 2020 22:03:11 GMT

firstitem = firstchild.find('item')

firstitem

Out[38]: <Element 'item' at 0x11b363e90>

alltitles = tree.findall('.//title')

len(alltitles)

Out[40]: 29

for title in alltitles[:6]:

print(title.text)

BBC News - World

BBC News - World

Coronavirus: US to borrow record $3tn as spending soars

Coronavirus: World leaders pledge billions for vaccine fight

Hundreds of sheep flock to locked down city

Malaria 'completely stopped' by microbe