

XML Web Scraping Assignment

This assignment has to do with scraping data in XML format from the NOAA web site below. This assignment statement provides considerable detail to help you construct your code.

- Use Google Chrome for the initial steps of investigating the NOAA site and determining the URL structure.
- Manually go to the web page noted below and choose the parameters as are, also, noted below before left-clicking on the Update button:
 - <http://www.ncdc.noaa.gov/temp-and-precip/climatological-rankings/index.php>
 - Parameters:
 - Period(s): 6 month period
 - Parameter: Average Temperature
 - State/Region/Ag Belt: Virginia
 - Climate Division: All Divisions
 - Month: June
 - Year: 2017
- Left-click on the Update button.

The screenshot shows the NOAA website's 'Climatological Rankings' page. The search form is configured with the following parameters:

- Period(s): 6-Month Period
- Parameter: Average Temperature
- State / Region / Ag Belt: Virginia
- Climate Division: All Divisions
- Month: June
- Year: 2017

The 'Update' button is highlighted with a red box. Below the form, the results for 'Contiguous U.S. Average Temperature Rankings, June 2017' are displayed. The table shows the following data:

PERIOD	AVG TEMP	20 TH CENTURY AVERAGE	DEPARTURE	RANK	WARMEST/COLDEST SINCE	RECORD
Jun 2017 1-month period	70.34°F (21.30°C)	68.49°F (20.27°C)	1.85°F (1.03°C)	10th Warmest	Coldest since 2014	1907
Jun 1977				30th Warmest	Warmest since 2016	2016

- Right Click on “XML” button and “Open link in New Tab”
- Observe the URL specification in that new tab and how the search parameters are embedded in it.

- Write a Python program named `xml_scrape.py` to access Average Temperature Data in XML format using that URL structure found above by following these instructions/specifications:
 - Get average temperature data for:
 - `Period(s)`: 6 month period
 - `State/Region/Ag Belt`: Virginia
 - `Climate Division`: All Divisions
 - `Month`: August
 - `Year`: 2016
 - Create a string variable for each one of the parameters above
 - Assign those variables the appropriate values to embed into a URL in order to obtain data from the NOAA web site
 - Retrieve the XML data, parse it with the `lxml` package, and print each of these data items on a separate line, in this order, without any other printed text:
 - Your W&M username (this doesn't come from the web page)
 - `value`
 - `twentiethCenturyMean`
 - `lowRank`
 - `highRank`
 - Use the Python coding template that has been provided, which is entitled `xml_scrape.py`
 - Submit your Python code file. There are two options, where the first works only if you are on campus.
 - On Campus: Open a Windows File Explorer Window and paste this location into the address bar while substituting your W&M username for `your_username` and your Course Section Number for `X`:
\\jonesfiles.campus.wm.edu\acstore-classes\BUAD5012-X\student\your_username
 - Off Campus: FTP your file using the directions in the PowerPoint file from Blackboard named "FTP Access to Network Folder.pptx"
- Coding hint... use the string substitution method
 - Create a string template with the symbols '`%s`' as placeholders for where you will insert the values for month, year, etc., for example,
 - `template = "My name is %s, %s"`
 - Then you can substitute string values for the '`%s`' symbols using a statement like this:
 - `last_name = 'Bradley'`
 - `first_name = 'Jim'`
 - `print template % (last_name, first_name)`
 - This results in a printout of `'My name is Bradley, Jim'`
 - You can use the same approach for this assignment by creating a variable name for `Period(s)`, `State/Region/Ag Belt`, `Climate Division`, `Month`, and `Year` and substituting those values into a string template for the NOAA web page.