**Information Systems – Spring 2019**

**HW3: Webscraping (25 points)**

**SOLUTIONS**

**Due: Jan 31, 2019**

Objective:

* Learn additional webscraping techniques by following an online tutorial
* Practice webscraping by continuing the NFL example we started in class

Submission:

* Submit your **Jupyter notebooks to the D2L dropbox by Jan 31 at 1600**. You should have two separate notebooks, one for each part. The notebooks must have all the cells executed (with output shown).
* Turn in a **hardcopy of the Part I worksheet with your answers written in**, in class on Jan 31.

Collaboration:

You may collaborate on this assignment with students currently enrolled in this course. You are encouraged to post and answer questions on Piazza related to this assignment. You may use code presented in class or posted on the class website as examples. You may use the Internet as a resource for code documentation and ideas. Direct copying from the Internet (or anyone) is never allowed, and you must be able to cite and explain any code you have incorporated within your own program. If in doubt, ask your instructor.

Grading:

In addition to correctness of your answers, you will be graded on completeness and organization of your notebook. Your code should be well commented and include markdown cells to indicate what you are doing.

**Part I – Weather webscraping tutorial**Your submission for Part I will be a Jupyter Notebook named “LastName\_HW2\_Part1” and this completed worksheet.

1. Complete the following tutorial: <https://www.dataquest.io/blog/web-scraping-tutorial-python/>

Take your time reading through this tutorial and follow along in your own Jupyter notebook. Use markdown cells where appropriate – this will be a great reference for you to use in the future.

1. Briefly explain the difference between the find and find\_all methods.

find returns the first instance of a html tag whereas find\_all returns all instances of a html tag.

1. Consider this line of code from the tutorial:

forecast\_items = seven\_day.find\_all(class\_="tombstone-container")   
How many elements are in forecast\_items? Does this make sense? Explain.  
There are 9 elements in forecast\_items. Yes, this makes sense because there are forecast items for tonight, and the next four days (day and night) for a total of 9 forecast items.

1. Provide the code necessary to extract the information shown in the green box below, saving it to a dictionary in this form:

{'Barometer': 'NA',

'Dewpoint': '28°F (-2°C)',

'Humidity': '38%',

'Last update': '02 Jan 11:43 am PST',

'Visibility': 'NA',

'Wind Speed': 'NA'}

Note: your information will look different than the one shown as I took this screenshot several weeks ago.

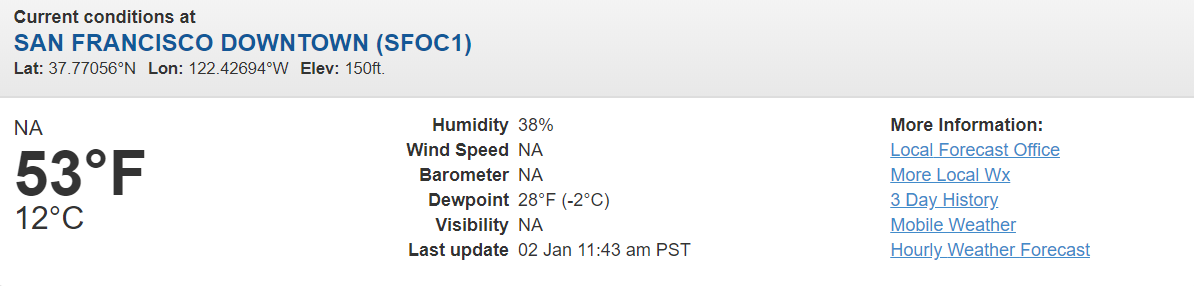


table = soup.find('table')

tds = table.find\_all('td')

wxDict = {}

for i in range(0,len(tds),2):

wxDict[tds[i].find('b').get\_text()] = tds[i+1].get\_text().strip()

wxDict

**Part II – NFL passing data**Your submission for Part II will be a Jupyter Notebook named “LastName\_HW2\_Part2”.

For Part II, you will continue working on the NFL webscraping example we started in class. Download the HW2\_Part2\_Blank.ipynb file on D2L and follow the instructions given.   
see Jupyter notebook for solutions