



Scraping HTML Data with BeautifulSoup

In this assignment you will write a Python program to use **urllib** to read the HTML from the data files below, and parse the data, extracting numbers and compute the sum of the numbers in the file

This course uses a third-party tool, Scraping HTML Data with BeautifulSoup, to enhance your learning experience. The tool will reference basic information like your name, email, and Coursera ID.



I, **Ishaan Narula**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

[Learn more about Coursera's Honor Code](#)

[↗ Open Tool](#)

Done



Welcome Ishaan Narula from Using Python to Access Web Data

Scraping Numbers from HTML using BeautifulSoup In this assignment you will write a Python program similar to <http://www.py4e.com/code3/urllink2.py>. The program will use **urllib** to read the HTML from the data files below, and parse the data, extracting numbers and compute the sum of the numbers in the file.

We provide two files for this assignment. One is a sample file where we give you the sum for your testing and the other is the actual data you need to process for the assignment.

- Sample data: http://py4e-data.dr-chuck.net/comments_42.html (Sum=2553)
- Actual data: http://py4e-data.dr-chuck.net/comments_400226.html (Sum ends with 23)

You do not need to save these files to your folder since your program will read the data directly from the URL.

Note: Each student will have a distinct data url for the assignment - so only use your own data url for analysis.

Data Format

The file is a table of names and comment counts. You can ignore most of the data in the file except for lines like the following:

```
<tr><td>Modu</td><td><span class="comments">90</span></td></tr>
<tr><td>Kenzie</td><td><span class="comments">88</span></td></tr>
<tr><td>Hubert</td><td><span class="comments">87</span></td></tr>
```

You are to find all the `` tags in the file and pull out the numbers from the tag and sum the numbers.

Look at the sample code provided. It shows how to find all of a certain kind of tag, loop through the tags and extract the various aspects of the tags.

```
...
# Retrieve all of the anchor tags
tags = soup('a')
for tag in tags:
    # Look at the parts of a tag
    print 'TAG:',tag
    print 'URL:',tag.get('href', None)
    print 'Contents:',tag.contents[0]
    print 'Attrs:',tag.attrs
```

You need to adjust this code to look for **span** tags and pull out the text content of the span tag, convert them to integers and add them up to complete the assignment.

Sample Execution

Done

```
$ python3 solution.py
Enter - http://py4e-data.dr-chuck.net/comments_42.html
Count 50
Sum 2...
```

Turning in the Assignment

Enter the sum from the actual data and your Python code below:

Sum: (ends with 23)

Python code: