Introduction to Python II (Exercises 03)

Simple web scraping

1) Using the beautifulsoup module, write a program that does the following:

   Open the URL www.meteomedia.com

   Find how many <a> tags are there in the html code
   Print all href links inside the <a> tags.

2) Use the beautifulsoup module to fetch the contents from:

   www.groupce.com/python/html/thejourney.html

   Print all <a> tags. How many are there?
   Print all <tr> tags. How many are there? (note: <tr> are tags to specify rows in an html table)
   For every <tr> tag, find all <td> tags. Print all <td> tags.
   Print the second <td> tag found in every <tr> tag .

3) Use the beautifulsoup module to fetch the contents from:

   www.groupce.com/python/html/thejourney.html

   This is the first step in your journey.

Akuna	35	Akuna way
Samara	20	Samara way
Transxer	15	Transxer way
Romanor	42	Romanor way
Xlander	20	Xlander way
Picamont	5	Picamont way
Nederham	88	Nederham way
Sheemba	3	Sheemba way
Tomatoid	52	Tomatoid way
Serviante	63	Serviante way

   Add up all the integers in the second column of the table.

4) (JSON exercise) Fetch the contents of the following URL (using urllib):

   http://www.groupce.com/python/json/json_comments.json

   and parse it using the json standard library.

   Print all the names that start with an 'A' and
   print the 'count', and the running total for the 'count'.

   Hint: take a look at the json file so that you get an idea of its format.

5) (JSON exercise) Create a list of dictionaries with 5 countries, their capital, and approx. population(of the capital)

   Sample entry:
   {'Country': 'Canada', 'Capital':'Ottawa', 'Population': 883,391}

   Write the dictionary to a json format file.

   Check the contents of the file in notepad or other text editor.

   Read the file back into your script and parse with json.

   Print the contents from the resulting dictionary. (key and value)

6) (BONUS Scraping exercise) Using beautifulsoup, start by fetching the contents of:

   www.groupce.com/python/html/thejourney.html

   Prompt the user for 2 numbers (ie. 2 integers) (between 1-4)

   Starting from the first html file: thejourney.html

   Find the table row corresponding to the first integer, extract its link (href)
   Follow the extracted "href" to the corresponding html page. Once in the page,
   use the second number to extract the corresponding row in the html.
   Once again follow the link (href) and find the new page. In the new page you need to find the title of the page. Inside the title, you will get a string containing a mathematical operation. Extract it and print the result of the operation (hint. Use eval() to evaluate a string containing some kind of algebraic operation).