SI 618 Lab 1:  Fetching and parsing structured content.

The purpose of this lab is to have you practice using some powerful

modules for fetching and parsing important Web content types:

    urllib3 or urlopen : for fetching the content of a URL (e.g. HTML page)

    BeautifulSoup : for parsing HTML and XML pages

    json : for JSON reading and writing

As in earlier labs, you should fill in the code for the functions provided.

main() is already set up to call the functions with a few different inputs,

printing 'OK' when each function is correct.

A. get\_title (2 points)

The get\_title function should process the HTML page stored in the global

variable html\_doc, and return the title of the page.

get\_title() should return u'Three Little Pigs'

B. process\_json (2 points)

The process\_json function should load the dictionary stored as a JSON string

in global variable json\_str, and return the sum of the values in this dictionary.

process\_json() should return 8 because 3+2+1+1+1 = 8

C. get\_pigs (3 points)

The get\_pigs function should process the HTML page stored in the global variable

html\_doc, and return the three pigs listed below 'there were three little pigs'

in a JSON string.

Note that it should return a string, not a list.

get\_pigs() should return '["Pig A", "Pig B", "Pig C"]'

D. get\_story\_headings (3 points)

The get\_story\_headings function should process the HTML page stored in the global variable

html\_doc, and return the three story headings in a JSON string.

Note that it should return a string, not a list.

get\_story\_headings() should return '["Story 1", "Story 2", "Story 3"]'

E. get\_houses (3 points)

The get\_houses function should process the HTML page stored in the global variable

html\_doc, and return information in the house table in a JSON string.

Note that it should return a string, not a list.

get\_houses() should return '[["Pig A", "Straw"], ["Pig B", "Stick"], ["Pig C", "Brick"]]'

HINT: contruct a list of tuples first, and then convert it to a JSON string.

F. get\_links (3 points)

The get\_links function should process the HTML page stored in the global variable

html\_doc, and return all url links in the page in a JSON string.

Note that it should return a string, not a list.

get\_links() should return '["http://en.wikipedia.org/wiki/Three\_Little\_Pigs", "http://en.wikipedia.org/wiki/Big\_bad\_wolf"]'

G. treasure\_hunting (4 points)

The treasure\_hunting function should first visit http://cbudak.com/618/Example\_Domain.htm, and

then find the only url link on that page, and then visit that url link.

On this page, there is a table under 'Test IDN top-level domains'. In the first

column (Domain), there are a list of foreign characters.

You need to fetch the content of the cell in column 1 and row 3, and return it.

treasure\_hunting() should return the Unicode string u'\u6d4b\u8bd5' corresponding

to the characters 测试  (the code points U+6D4B U+8BD5)