Advanced Web Development and Web Scraping
Spring 2020
Exam II Practice

Format: the exam will involve Python coding and web scraping questions similar to the homeworks and to the practice problems below.

Outline

I. **Python – you should be able to write code that uses the following Python concepts**
   a. Input/output
   b. Basic arithmetic
   c. Strings, Lists, and slicing
   d. Dictionaries
   e. For loops (e.g., to iterate through each element of a list or dictionary)
   f. *If* statements

II. **Web scraping with requests and BeautifulSoup**
   a. Retrieving a web page using the requests package, and specifying an appropriate header, and checking for a valid response
   b. BeautifulSoup
      i. Parsing a web page to create a Beautiful Soup object
      ii. Getting the first element of a particular type
      iii. Finding the first element or all elements by type, id, or class name
      iv. Extracting the text of an element
      v. Finding the children of an element

Practice Problems

1. Write code that asks the user to enter two numbers, then finds the sum of the numbers, and outputs the result in the form: *The sum of 3 and 5.6 is 8.6*.

2. Create a list containing the following strings: *dog, cat, bird, kangaroo*
   Iterate through the list to print out all animals in order, and in reverse order, as follows:

```
The animals in order are:
dog
cat
bird
kangaroo


The animals in reverse order are:
kangaroo
bird
cat
dog
```

3. Print out all animals that contain a 'd' in the name

4. Print out all animals in uppercase, if they are 3 characters long

5. Create a word count dictionary that counts the number of times each word occurs in the following string:

   This is a test to see which word occurs the most. What will the most common word be?

   First, use the *replace* function to remove the period and question mark from the string, and convert each word to lowercase. Then use the *split* method to create a list of words from the string.

   In order to count words, start with an empty dictionary, e.g., *d = {}*. Then for each word, check if the word is already in the dictionary. If not, then add the word to the dictionary (the word is the key), and the value is 1. If the word is already in the dictionary, then increase the counter. Output all the words in the dictionary and the corresponding word counts. Also output the word that occurs most often (if more than one word occurs the most, you can output any of those words).

6. Write a webscraper to extract and output the director, writers, and stars for the following movie https://www.imdb.com/title/tt1502407/?ref_=nv_sr_2
   Note: you should extract the text of each link, so for writers, you will output "3 more credits".

7. From Eastern's Wikipedia page (https://en.wikipedia.org/wiki/Eastern_Connecticut_State_University), extract and output the following:

    - The main heading ("Eastern Connecticut State University")
    - The secondary headings ("History", "Academics", etc)
    - From the table on the right hand side, the date established, and the total number of students.

8. Create a data frame that consists of the school name, year founded, and number of students for Eastern Connecticut State University, Central Connecticut State University, Western Connecticut State University, and Southern Connecticut State University from the appropriate Wikipedia pages. Sleep for 1 second after making a request to each page. Then construct a bar graph showing the number of students at each University. The label for the bar graph should consist of just the first word for each University (e.g., "Eastern" rather than "Eastern Connecticut State University".