CS 112 Assignment 13

Submit your finished code to the Dropbox. You will need to pass off the assignments with the TA or tutor during their office hours (their information is in the CS 112 Course Information section of the Table of Contents), or the instructor in class. Follow the formatting guide which is also found in the CS 112 Course Information section.

Read the instructions carefully. Make sure your output matches the example run.

Objectives:

In this assignment you will learn how to do the following:

- Become familiar with parsing html and extracting information from a web page.
- Libraries Needed:
    - Request
    - bs4

Install the beautifulsoup4 module using PIP

For this assignment you will need to install a third-party module using pip. Follow the instructions for installing third party modules found in Appendix A of the book or the instructions for installing a module from the Python Software Foundation at https://docs.python.org/3/installing/.

Pip is run from the command console in Windows. To find the command console, click the Windows Start button and look in either Accessories (Windows 7) or Windows System (Windows 10). To run Command Prompt as an administrator, right click on the Command Prompt icon and choose 'run as administrator'. Inside the Command window, you will need to change the directory to where pip is located. Type cd and then the path where pip is located.

From there, simply type

>pip install beautifulsoup4

Program: Save elements of a web page and their URLs to a file.

Write a program that goes to the Deseret News site and extracts the headlines and their URLs and then save them to a file. The file should contain the Title of the headline as well as a **complete** URL to the story.

Key Hints:

- The book shows the wrong syntax for creating a beautifulsoup object.

  - o Book example (WRONG):
    res = requests.get('http://nostarch.com')
    noStarchSoup = bs4.BeautifulSoup(res.text)

  - o CORRECT syntax:
    res = requests.get('http://nostarch.com')
    NoStarchSoup = bs4.BeautifulSoup(res.text,"html.parser")

- The book does not cover all of the methods available for a beautifulsoup object. For complete documentation on beautifulsoup go to the following link:
  https://www.crummy.com/software/BeautifulSoup/bs4/doc/

- One very helpful method for BeautifulSoup is the find_all() method. See the BeautifulSoup documentation at the link above for details.

- The elements you are looking for are **rotator-item__story-headline**

- The BeautifulSoup find_all() method takes two parameters and returns a list of all the elements that match the class specified in the second parameter.

  Syntax:
  elementList = NoStarchSoup.find_all('a',class_='rotator-item__story-headline')

- Use print statements to make sure your code is doing what you think it is doing. (See example run)

Example Run

Screen Scraping 'http://deseretnews.com'

Loading the online page 'http://www.deseretnews.com' into the dnWeb BeautifulSoup object...

Getting the elements from the dnWeb BeautifulSoup object.

The element[0]'s getText() value is: Feds agree not to prosecute UTA but continue investigation
The element[0]'s href is: /article/865677092/Feds-agree-not-to-prosecute-UTA-but-continue-investigation.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677092/Feds-agree-not-to-prosecute-UTA-but-continue-investigation.html
writting the element to the file....

The element[1]'s getText() value is: Jazz notes: Utah without three starters against Portland tonight
The element[1]'s href is: /article/865677104/Jazz-notes-Utah-without-three-starters-against-Portland-tonight.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677104/Jazz-notes-Utah-without-three-starters-against-Portland-tonight.html
writting the element to the file....

The element[2]'s getText() value is: Jazz notes: Utah without three starters against Portland tonight
The element[2]'s href is: /article/865677104/Jazz-notes-Utah-without-three-starters-against-Portland-tonight.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677104/Jazz-notes-Utah-without-three-starters-against-Portland-tonight.html
writting the element to the file....

The element[3]'s getText() value is: How much a 1844 LDS hymnal was appraised for on TV might surprise you
The element[3]'s href is: /article/865677023/1844-LDS-hymnal-appraised-for-40000-50000-on-Antiques-Roadshow.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677023/1844-LDS-hymnal-appraised-for-40000-50000-on-Antiques-Roadshow.html
writting the element to the file....

The element[4]'s getText() value is: How much a 1844 LDS hymnal was appraised for on TV might surprise you
The element[4]'s href is: /article/865677023/1844-LDS-hymnal-appraised-for-40000-50000-on-Antiques-Roadshow.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677023/1844-LDS-hymnal-appraised-for-40000-50000-on-Antiques-Roadshow.html
writting the element to the file....

The element[5]'s getText() value is: 10 things learned from BYU's spring football camp
The element[5]'s href is: /article/865677078/10-things-learned-from-BYUs-spring-football-camp.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677078/10-things-learned-from-BYUs-spring-football-camp.html
writting the element to the file....

The element[6]'s getText() value is: 10 things learned from BYU's spring football camp
The element[6]'s href is: /article/865677078/10-things-learned-from-BYUs-spring-football-camp.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677078/10-things-learned-from-BYUs-spring-football-camp.html
writting the element to the file....

The element[7]'s getText() value is: Got power? New Rocky Mountain Power plan includes more wind, more solar, new transmission line
The element[7]'s href is: /article/865677084/Got-power-New-Rocky-Mountain-Power-plan-includes-more-wind-more-solar-new-transmission-line.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677084/Got-power-New-Rocky-Mountain-Power-plan-includes-more-wind-more-solar-new-transmission-line.html
writting the element to the file....

The element[8]'s getText() value is: Feds agree not to prosecute UTA but continue investigation
The element[8]'s href is: /article/865677092/Feds-agree-not-to-prosecute-UTA-but-continue-investigation.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677092/Feds-agree-not-to-prosecute-UTA-but-continue-investigation.html
writting the element to the file....

The element[9]'s getText() value is: Morgan teen nearly dies after overdose linked to 'pink' drug
The element[9]'s href is: /article/865677083/Morgan-teen-nearly-dies-after-overdose-linked-to-pink-drug.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677083/Morgan-teen-nearly-dies-after-overdose-linked-to-pink-drug.html
writting the element to the file....

The element[10]'s getText() value is: Morgan teen nearly dies after overdose linked to 'pink' drug
The element[10]'s href is: /article/865677083/Morgan-teen-nearly-dies-after-overdose-linked-to-pink-drug.html
Modifying the href into a active link yields: www.deseretnews.com/article/865677083/Morgan-teen-nearly-dies-after-overdose-linked-to-pink-drug.html
writting the element to the file....

The data has been written to the DeseretNewsHeadlines.txt file
>>>

Output File Example

Feds agree not to prosecute UTA but continue investigation,www.deseretnews.com/article/865677092/Feds-agree-not-to-prosecute-UTA-but-continue-investigation.html

Jazz notes: Utah without three starters against Portland tonight,www.deseretnews.com/article/865677104/Jazz-notes-Utah-without-three-starters-against-Portland-tonight.html

Jazz notes: Utah without three starters against Portland tonight,www.deseretnews.com/article/865677104/Jazz-notes-Utah-without-three-starters-against-Portland-tonight.html

How much a 1844 LDS hymnal was appraised for on TV might surprise you,www.deseretnews.com/article/865677023/1844-LDS-hymnal-appraised-for-40000-50000-on-Antiques-Roadshow.html

How much a 1844 LDS hymnal was appraised for on TV might surprise you,www.deseretnews.com/article/865677023/1844-LDS-hymnal-appraised-for-40000-50000-on-Antiques-Roadshow.html

10 things learned from BYU's spring football camp,www.deseretnews.com/article/865677078/10-things-learned-from-BYUs-spring-football-camp.html

10 things learned from BYU's spring football camp,www.deseretnews.com/article/865677078/10-things-learned-from-BYUs-spring-football-camp.html

Got power? New Rocky Mountain Power plan includes more wind, more solar, new transmission line,www.deseretnews.com/article/865677084/Got-power-New-Rocky-Mountain-Power-plan-includes-more-wind-more-solar-new-transmission-line.html

Feds agree not to prosecute UTA but continue investigation,www.deseretnews.com/article/865677092/Feds-agree-not-to-prosecute-UTA-but-continue-investigation.html

Morgan teen nearly dies after overdose linked to 'pink' drug,www.deseretnews.com/article/865677083/Morgan-teen-nearly-dies-after-overdose-linked-to-pink-drug.html

Morgan teen nearly dies after overdose linked to 'pink' drug,www.deseretnews.com/article/865677083/Morgan-teen-nearly-dies-after-overdose-linked-to-pink-drug.html