ANSWER KEY

INFO 102 Lab 9: Web Scraping

We will learn about **Web Scraping** on the following platform for today's activity: go.illinois.edu/102web
Login using the following username: **{{Username}}** Your password is: **{{Password}}**

Please follow the instructions online and complete the online activities first. You can discuss with your partner. After finishing the online tutorial, complete this worksheet with your partner. Your attendance will be based on completing the activities online and this worksheet.

Q1: *The find_all method is used in Plan 3.* What does the find_all method do? Give an example of when you would use the find_all method over the find method.

The find_all method in Beautiful Soup finds and returns a list of all matching elements based on the given criteria.
One possible answer: When looking for all bb.q chicken locations instead of just the first one listed .

Q2a: Fill in the blanks to complete the code for getting the tag's **text**.

```
tag = soup.find('a')
info = tag.text
print(info)
```

Q2b: Fill in the blanks to complete the code for getting the tag's **link**.

```
tag = soup.find('a')
info = tag.get('href')
print(info)
```

Q3: Look at the following HTML:

```
<h2 class="major-heading">Information Sciences</h2>
```

a) What would you see **in your browser** if you go to a website that had this HTML piece?
A heading that says 'Information Sciences'

b) What is the **text** for this HTML element?
Information Sciences

c) What is the **tag description** for this HTML element?
'h1', class_='heading'

Q4: The website https://myillini.illinois.edu/Programs contains a list of **headings** to the pages for every major at UIUC. **Circle the three plans** needed to scrape the names of every major at UIUC and print them.

a. [✓ Plan 1: Get a soup from a URL]
b. Plan 2: Get a soup from multiple URLs
c. [✓ Plan 3: Get info from all tags of a certain type]
d. Plan 4: Get info from a single tag
e. [✓ Plan 5: Print info]

Q5: Write the HTML tag you would use to show the content on the left.
Hint: The HTML tags you will use in this exercise are: ***<p>, <h1>, <li>, <ul>, <a>, <img>***

Content to show on webpage	**HTML tag to use**
To show an image:	**img**
To write a heading:	**h1**
To write a paragraph:	**p**
To create an unordered list:	**ul**
To add a new item to the list:	**li**
To show a link to another webpage:	**a**

Q6: Write down your response for the **Code explaining activity** on the tutorial you completed.

- Accesses Jane Doe's faculty webpage
- Extracts href links from each news-row__link and stores them in collect_info
- Combines the collected links with a new base university url to access each news article
- Finds all the paragraph tags, extracts text content from the paragraphs, and stores them in collect info
- Print all of the collected paragraph texts

Q7a: Think of **one task** that would require you to collect information from a webpage. You can look at the examples in the tutorial for inspiration, but try to write down something different.

One example: Get list of products from Amazon

Q7b: *Discuss your partner's response on Q7a.* Can you achieve **their task** with webscraping? Write down which plans you would need to use, or explain a challenge you would encounter when completing this task.

Potential answer: you need to get a soup from the URL, get all tags, and print them

When you're done, check out with a TA or CA, and hand over this completed worksheet.
Bye!