Fall 2021 DASC 5301 Data Science
Programming Assignment 1
Due: October 8, Friday, 11:59pm

Problem Description

In this assignment, you are required to extract information about United States National Parks. Go to https://www.nps.gov/index.htm. From the drop-down list "By State..." (see screenshot 1 below), extract the links to all states. Follow the link to each state's page, you will find a list of parks in that state (see screenshot 2 for Texas's page.) From that page, extract each park's category (e.g., "NATIONAL MONUMENT"), name (e.g, "Alibates Flint Quarries"), and description (e.g., "13,000 years ago, this site ..."), as well as the link to the park's page. Follow the link to each park's page. You will see contact information at the bottom banner of the page (see screenshot 3 for the banner of Big Bend National Park.) From there extract the park's address, phone number, and various social media accounts. Store all extracted information into a CSV file, with the following columns in this particular order: Name, Category, Description, Street Address Line 1, Line 2, Line 3, City, State, Zip Code, Phone Number, and one column for each social media account: Facebook, Twitter, Instagram, YouTube, Flickr. Note that a park's information for some of the columns might be empty.

Below is the address of a park. Column Line 1 should have the value "Potomac Heritage NST Office", Line 2 is "National Park Service", and Line 3 is "1100 Ohio Drive SW", City is "Washington", State is "DC", and Zip Code is "20242". Note that some parks only have one or two lines in their street addresses, in addition to city, state, and zip code.

Potomac Heritage NST Office
National Park Service
1100 Ohio Drive SW
Washington, DC 20242

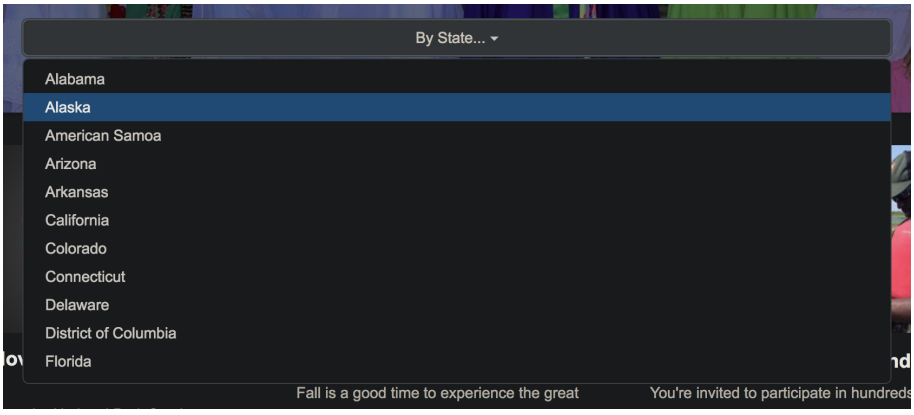

Figure 1: Screenshot 1: Links to all states

Academic Honesty

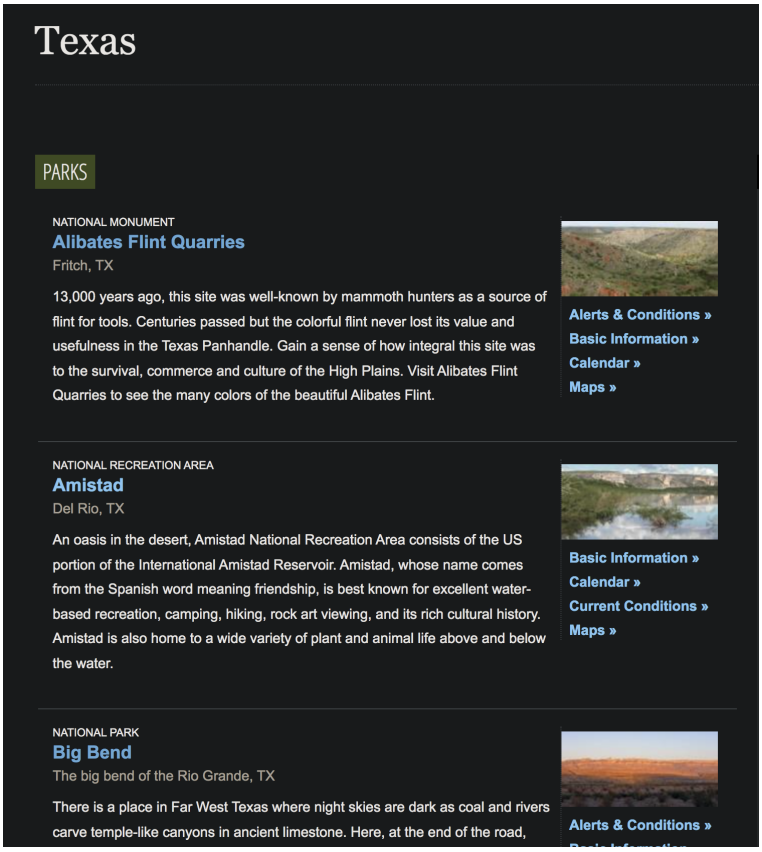

Figure 2: Screenshot 2: List of parks in a state

- This assignment must be done individually and independently. You must implement the whole assignment by yourself. Academic dishonesty is not tolerated.
- You can discuss topics related to the assignment with your fellow students. But you are not allowed to discuss/share your solution and code.

Python Version

We will test your code under the particular version of Python 3. Make sure you develop your code using this version.

What to Submit

You are required to submit a Python file named national_parks.py to the Programming Assignment 1 entry in Canvas.

Grading Rubrics

Your program will be evaluated on correctness, execution efficiency, and code quality. Make sure to thoroughly understand the following grading rubrics.

(1) Correctness: 60 points

You will be evaluated on whether you can accomplish the given tasks, i.e., extract the required information and store it in a CSV file. The correctness score is further decomposed into the following components.

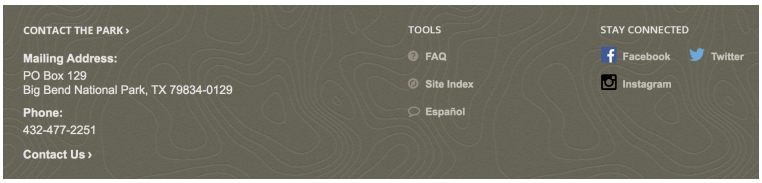

Figure 3: Screenshot 3: Contact information of a park

6 points: Retrieve the page with the drop-down list (page 1)
6 points: Retrieve the links to all states from page 1
6 points: Retrieve each state's page (page 2)
6 points: Retrieve the list of all parks from page 2 for each state
12 points: Retrieve each park's name, category, and description from the above list of parks in each state.
6 points: Retrieve each park's page (page 3)
12 points: Retrieve each park's contact information from page 3 for each park.
3 points: Results correctly stored in a CSV file.
3 points: The CSV file correctly follows the required schema (attribute names and ordering).

(2) Execution efficiency: 20 points (The following are guidelines for representatives scores. A student's score might take other values on the scale.)
20 points: your code is highly efficient and is among the fastest in the submissions.
15 points: your code is faster than majority of the submissions.
10 points: your code's execution efficiency is average.
5 points: your code is slower than majority of the submissions.
0 points: your code is highly inefficient and is among the slowest in the submissions, OR yours is mostly incorrect implementation which makes efficiency evaluation not meaningful.

(3) Quality—clarity, organization, modularity, comments: 20 points (The following are guidelines for representatives scores. A student's score might take other values on the scale.)
Follow good coding standards to make your program easy to understand and easy to maintain/extend. Provide sufficient comments in your code and make it self-explaining. Use functions when appropriate to make your code modularized.
Excellent : 20 points
Very Good : 15 points
Good : 10 points
Fair : 5 points
Poor: 0 points

(4) Total score: 100 points
Your score will be calculated from the individual break-ups using the following equation:
Correctness + Efficiency + Quality