CMPS 184 **Spring 2019**

Assignment 5: Part A
Web Scraping (7 points)
Due Date: May 2-6, Thursday-Sunday[1], 11:59pm

The objective of this assignment is to demonstrate that you understand the concepts presented in Chapters 1 and 2 of the Book titled "Web Scraping with Python" by Ryan Mitchell, and Advanced Tutorial on Web Scraping Movie Website IMDB (covered in class).

1. Create one ipynb file with comments that capture the code for Chapters 1 and 2 of Ryan Mitchell's Book paying particular attention to Exception Handling (in Chapter 1) and Beautiful Soup commands such as prettify, find_all with various arguments, navigating trees, regular expressions, and lambda expressions.

(2 points)
(i) For page 3 "Totally Normal Gifts", add the following code: create and print a dataframe that displays the item title and cost with the headers.

(ii) For Russian Nesting Dolls, given the image, capture the following information (and only the following information) in a list and display: (in the description, only capture the bold part; in cost part, remove the decimals)
[Russian Nesting Dolls, 8 entire dolls per set! Octuple the presents!, $10,000]

2. (4 points) Capture the Dataquest Advanced Tutorial (required to use time module to pace your requests and also warning module to monitor your requests) in a ipynb file with the following changes:
Restrict your search for movies using advanced search starting 01-01-2010 till 12-31-2018, sorted in descending order by number of votes, include all titles, show all tiles, exclude adult titles:
(i) (1 point) capture and display the distribution of ratings as described in the tutorial,
(ii) (4 points) capture the gross collection by the movies in an additional column and create and display two scatter plots : one with x-axis as imdb scores and y-axis with Gross collections
and second with x-axis as metacritic scores and y-axis with Gross collections

(ii) Number of requests sent to IMDB site should be between [40,80] to be determined as follows: use last two digits of your Student ID. IF in the range, done. If less than range, multiply by 2 and convert to an integer till in the range. If more than range, multiply by .75 till in the range. For example, if last two digits is 16 or 34, you will end up with 64 or 68. If last two digits is 94, then you will end up with 71.

---

[1] I am concerned about overwhelming the IMDB website with web scraping requests. Ideally, there should be no more than 12 students per day bombarding IMDB website with these requests. Given that the assignment is due over a period of 4 days, hopefully random submissions will take care of it. It will also allow some students to give higher priority to presentation assignment.

Submission Requirements

1. Create a directory called “Assignment 5A”. Place the ipynb files inside this directory. Call them Assignment 5A.1.ipynb and Assignment 5A.2.ipynb. Submit the zipped directory.