MIS 532: ADVANCED BUSINESS PROGRAM DEVELOPMENT
Fall 2018
Project 1 (Due September 20, 6:40 pm)

Project Description:

Box Office Mojo (http://boxofficemojo.com) is a website that tracks a variety of information on movies released each year (such as genre, studio, box office revenues, etc.). You will write scripts that collect data on more than 700 movies for the year 2017 that are listed on the site. Information on movies that were released in 2017 is available at:

http://www.boxofficemojo.com/yearly/chart/?yr=2017&p=.htm

You will also write scripts to collect data from the page People Index (also on the Box Office Mojo site) on all movie actors (ordered by Total Gross) listed on the page (there are more than 800 actors currently):

http://boxofficemojo.com/people/?view=Actor&pagenum=1&sort=sumgross&order=DESC&&p=.htm

There will be four main steps involved.

Step 1:
Create a database to store the data you will collect. Note that you will need two tables to store the data (one table for the movies, and the other table for actors and their rank). Please specify appropriate data type for each attribute in your tables. For example, Movie Name should be defined as text, and Movie ID could be defined as integer.

Step 2:
Collect the following information for each movie and save in a text file named **movies.txt**:

i. Movie Name.
ii. Movie ID. Note that this is not displayed on the webpage, but appears in the source code for the page. For example, for movie "Star Wars: The Last Jedi", the Movie ID is "starwars8".
iii. Studio that produced the movie.

Step 3:
Collect the following information for each actor from the page People Index (also on the Box Office Mojo site) and save in a separate text file named **actors.txt**:

i. Rank (Row number displayed on the page).
ii. Actor Name.
iii. Actor ID (it is in the source code of the webpage).

Step 4:

You will load the text files created in Steps 2 and 3 into the corresponding database tables. Note that an alternative approach is to insert the data into the tables directly as you collect data on each movie/actor (i.e., directly in Steps 2 and 3).

Deliverables:

1. A copy of your python code.
2. The **movies.txt** and the **actors.txt** files you created in Steps 2 and 3.
3. A copy of your database in a self-contained file with .db as the extension name.
4. A data dictionary for your database. The dictionary should list the names of all the tables, and the names and data types of all attributes in each table.
5. Which studio produced the most number of movies in 2017? How many movies did the studio produce? Provide a list of all the movies produced by the studio. Also provide the SQL queries to get these answers.

Hint:

Dump database and structure (schema): http://www.sqlitetutorial.net/sqlite-dump/

Please note that in future assignments you will add to the data collected in this assignment. That will require you to use data collected in this assignment to collect additional data, and then to conduct analysis on that data. Therefore, it is very important that you ensure all data are correctly captured in your database.