ECON498/900 Data Scrapping and Machine learning(Spring 2020) Midterm Exam 1

Name: _____

UID: _____

Due: 10th April 2020 (11:59pm (EDT))

This exercise involves two parts: Scrapping a website and cleaning the dataset:

1. For the webscrapping part, it is more straightforward if you follow the steps listed below to do deep scrapping. Also you can choose to finish the bonus part and gain bonus points:
   (a) Collect the historical daily box office data of the movies in American cinemas on https://www.boxofficemojo.com/daily/2019/?view=year. Each of you is required to download the data for a different 7-year period. You will receive your assigned years via separate emails. Here, we use the year of 2019 as an EXAMPLE:
   (b) Step 1, you can get the link of each day's box office data in the year 2019 on https://www.boxofficemojo.com/daily/2019/?view=year. Parse the links of each day and save them to a CSV file for the use of the following steps.
   (c) Step 2, go to every link you obtained from step 1. There are approximately 2555 links in total($7 * 365$). Collect the data of all the movies from each link, including the movie name, daily gross box office, number of theaters, gross box office to date, number of days in release, distributers along with the website of that movie(finding it by ['href']), which will be saved to a CSV file for the use of following steps.
   (d) Step 3, download every movie's website you obtained from step 2 and parse the data of opening money, release date, MPAA, running time(use hour as the unit, 2 h 30 min = 2.5 h), Genres, in release period(use days as the unit), widest release, the ID of IMDbPro(You can get the id from the link of IMDBpro using the same method in the hint). Hint: You may want to use a unique filename(the movie's id) to save each movie so you can check whether you have already downloaded this movie or not before downloading each movie to avoid downloading repeatedly, which can save you a lot of time. You may use the regular expression to find the movie's id from its link and define the filename as its id using this: "filename = re.findall(r"[a-zA-Z]2[0-9]1,20", link)[0]".
   (e) Step 4, now you have two datasets. One is the daily box office data of each movie from step 1 and the other is the data of each movie's characteristics from step 3. Finally, merge the two datasets into a dataset with each movie's characteristics and daily box office for seven years. The format of your dataset should look like the sample dataset.

   Your answer should contains 3 parts:

   - Your python source code (py files)
   - The dataset you downloaded (csv files)
   - A README file which contains the instructions of explaining your source code and how to use your python program

   All 3 parts are important and your exam score depends on all of them. You should set up a git repository and upload your answer to Github before the deadline. If you are not familiar with Github, you may want to upload it earlier to make sure you will not miss the deadline. Use a private repository and invite Tom(misotomlam) and Yujie(deng2yj) as collaborators. Since this is an exam, you should work on it without communicating with your classmates and Yujie will not teach you how to do it. However, if you have difficulty uploading your answer to Github, you can email Yujie and me for assistance.