ECON498/900 Machine Learning and Big Data(Spring 2019)
Problem Set 1
Due: 16th April 2019 (In class)

Name: _____
UID: _____

This exercise involves two parts: Scraping a website and performing machine learning using the dataset obtained in scrapping.

1. For the webscraping part, you should choose to work on one of the following tasks:
   (a) Collect the historical data of all the cryptocurrencies on coinmarketcap.com, including the opening price, closing price, day high, day low, volume, market Cap. Together with the characteristics (like those I have collected in class) of all the cryptocurrencies which is on coinmarketcap.com.
   (b) Collect information of least 10000 GitHub users using the GitHub API.
   (c) Pick a particular route on a particular date, scrape kayak.com/expedia.com/travelocity.com/alike for the prices and characteristics of a sample of the flights available. You should collect the data for at least a week.
   (d) You can also choose to scrape it only once, but then you need to pick many many routes.
   (e) Collect the data about name, ratings, voter number, prices of boardgames on boardgamegeek.com.
   (f) Collect data from other websites/APIs which are interesting. If you choose this option you need to email me (tomlam@g.clemson.edu) about the website you want to scrape, what data you plan to get and for how long. I will decide whether the task is difficult and interesting enough for the requirement of this exercise.

   If you are ECON498 students, you have one more option:
   (a) Collect the characteristics of all the cryptocurrencies on coinmarketcap.com, including name, symbol, market cap, price, circulating Supply, and trading volume. Collect these data every 2 hours and for several days. The code will be very very similar to the code that I taught in class, but you need to change the scraping frequency (of course), and you need to collect all the cryptocurrencies instead of just the top 100.

2. Describe the data you got from part 1 and perform some machine learning using the data. It can be supervised learning or unsupervised learning. DO NOT just run the program at paste the results. You need to write about what questions you are asking, and how the machine learning program is helping you to answer the question you were asking. There is no requirement in the length of your answers but if anyone really really wants a rule of thumb, write 3 pages.

You must hand in your homework via Github. Please send your GitHub id to Liuna (lissagh@clemson.edu). Create a repository named "ECON498_ps1" or "ECON900_ps1". You should include everything you want to hand in inside your repository, including the code, the data downloaded, and the answer for part 2. You should also include a file named README, which includes a step-by-step instruction of how to run your Python code to collect the data and perform machine learning. This is especially important if you have multiple Python files (And you should!).