Assignment 1

This assignment is open book, and open access. The total credit is 100. At the submission, please submit your codes and the associated files as noted in each question.

Q1. Data Collection

Suppose we want to collect data from Top 200 movies in 2023. This is the link: https://www.boxofficemojo.com/year/world/2023/.

1.1. Prepare a web crawler in Python to download a list of the IMDb Top 200 movies, and save a file, **TopMoviesBoxOffice.txt**; In the file, please include the variables, rank, movie_name, movie ID, Worldwide_Boxoffice, Domestic_Boxoffice. The list needs to include the headers (variable names) and the 200 movies info.
For example, one row of record is
"2,The Super Mario Bros. Movie,gr2226213381,1361973409,574934330" (15 points)

1.2. Prepare a web crawler in Python to collect the information of the introduction info. of the Top 200 movies, and save in a file, **TopMoviesIntro.txt**; In the file, please include the variables, movie_name, movie ID, Intro. The list needs to include the headers (variable names) and the 200 movies info.
For example, one row of record is "The Super Mario Bros. Movie, gr2226213381, A plumber named Mario travels through an underground labyrinth with his brother Luigi, trying to save a captured princess." (20 points)

1.3. Prepare a web crawler in Python to download the poster image of the top 50 movies, and organize them in a folder, named "Images/movie_id.jpg" (15 points)

**Hint: First collect all movie IDs. Then for each movie, visit the movie introduction page using the movie id. Finally, locate the poster and introduction elements in the web page.**

**To better evaluate your work, your submission should include:**

1. **code.py: your codes to complete the 3 tasks**
2. **readme.txt: briefly introduce the code blocks/functions corresponding to each task**
3. **TopMoviesBoxOffice.txt: collected data**
4. **TopMoviesIntro.txt: collected data**
5. **Images.zip: images of top 50 movies**

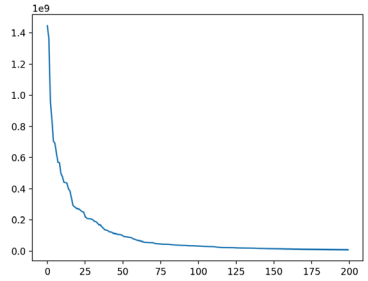
Q2. Data Cleaning

2.4 Please go through the text cleaning process in the tutorial and clean the **TopMoviesIntro.txt.** Create a new column named “cleaned_intro” and save the new data to **TopMoviesIntro_clean.txt**. The variables are movie_name,movie ID,Intro,cleaned_intro (15 points)

Q3. Data Analyses

3.5 Please plot a line plot of the Worldwide_Boxoffice variable in the **TopMoviesBoxOffice.txt,** from top 1 to top 200. (15 points)

The result is similar to:




3.6 Please take the introduction of all the Top 200 movies, and make a word cloud in a shape that you like, saved in an image named, top200 customized. (20 points)

**Hint: To do Q2, Q3, you can use jupyter notebook to complete the work.**

**To better evaluate your work, your submission should include:**

1. **code.ipynb: your codes to complete the 3 tasks**
2. **TopMoviesIntro_clean.txt: cleaned data**

**The figures you plot should be presented in the notebook.**