Ran LI
UIN_655284071

Final Project Proposal

- ***Project abstract***

  In order to better apply what I've learned in this class into practice or gain more practical experiences, I plan to choose the *large programming project*. Since watching films is one of my hobbies, I plan to focus on analyzing a dataset of movies (expected to include around 10,000 movies but depending on how large the dataset I can collect) collected from IMDB that consists of titles, directors, ratings, genres, release year, etc. My current goal is to export a series of lists categorized by release years, genres, directors, and ratings. (saved in separate files) In addition to sorting the movies into different categories, I also plan to further analyze the data; for example, I might try to find out the top 50 directors whose movies are mostly rated over 8 out of 10. By doing this, I could not only practice Python; but also bring convenience for myself, like it's easier for me to choose a film to watch when I feel boring later.

- ***List of deliverables***
  - Python file to obtain the results
  - The list of movies categorized by release years (separate files)
  - The list of movies categorized by genres (separate files)
  - The list of movies categorized by ratings (separate files)
  - The list of movies categorized by directors (separate files)
  - The file consisting of top 100 directors whose movies are mostly rate over 8/10
  - The narrative about my project

- ***Deliverables due at the midpoint check-in***

  Until the midpoint check-in, the following files are expected to be completed:

  - The full dataset of movies
  - The list of movies categorized by release years (separate files)
  - The list of movies categorized by genres (separate files)
  - The list of movies categorized by ratings (separate files)
  - The completed Python codes for the files above
  - The half-completed Python codes for slicing out the movies by the directors

- ***Steps to complete the project***

  1) Collecting data from IMDB website (expected to include around 10,000 films)
  2) Doing some research on reading & writing csv files with Python
  3) (Converting the csv file into txt file if necessary)
  4) Writing codes to slice out the movies based on its release years
  5) Writing codes to slice out the movies based on its genres
  6) Writing codes to slice out the movies based on its ratings
  7) Writing codes to slice out the movies based on its directors
  8) Writing codes to find out top 100 directors whose movies are mostly rate over 8/10
  9) Double check if any error
  10) Complete the narrative

- ***Possible challenges***
  - *Collecting data*

    It might take time to collect the dataset that I'm going to use; thus, I have to figure out a way to complete the very first step. I will try to read the instruction shown on the official site, if still not working, I might manually add movies into a watchlist and then export it in a csv file (This way could definitely work but waste a lot of time, I will try my best not use it.)

  - *Csv files in Python*

    Since the dataset collected from IMDB official site can only be csv file and I'm not very familiar with dealing with this type of file in Python, it might cause possible challenges to write codes. If I find it very difficult to write codes for csv file, I might choose to convert it into txt file and then write relevant codes.

  - *Possible problems occurred during coding*

    Since I'm still in the process of learning Python, many possible problems are likely to happen when I write the codes especially when I might have to deal with csv files for this project. If the problem does occur, I will try to search online to figure out a way to solve it or go back to double check my structure to see if anything goes wrong. If I couldn't find any solution online, I will reach out to classmates or professor to ask for suggestions.