

## Final Project Midpoint Check-in

The GitHub link: [https://github.com/ErinLR/IS452B\\_FinalProject](https://github.com/ErinLR/IS452B_FinalProject)

- **What've been done so far:**

- 1) I've been struggling with pulling out the data from IMDB database these days, and I've tried API (which is not working all the time); so I changed to another way: using Python Web Scraping with BeautifulSoup. However, I luckily found out a list consisting of all U.S. released movies from 1972 to 2016 can be directly exported as csv file (around 10,000 rows) from IMDB official site.

([http://www.imdb.com/list/ls057823854/?ref\\_=login](http://www.imdb.com/list/ls057823854/?ref_=login))

Thus, I finally got the data that I will play with. (I hope this way is not considered as "hard-code"; if so, I can go back to use web scraping way to pull out the data.)

- 2) Since the original file is a csv file that I'm not very familiar with, I looked up how to do stuff with it in Python, and then came up with some codes. However, the codes are not complete yet, and I'm still in the process of exploring it.

- 3) For now, the lists of movies by its released year have been extracted. Similarly, the lists by its genre/ratings/director can be obtained with very similar method. However, there might be problems when obtaining the list of movies by director, since there are a large number of names in the data. The most possible way I can come up with for now, is to first get a list of all the directors' names and then sort out the movies based on this list.

- **What else still needs to be done:**

Obtaining the following lists & Completing the narrative file

- 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