# Project Proposal

Project Abstract:

The program involves the text file processing and string manipulation which the class has taught us. It lets the user input the name of some actor or actress, like ‘Ziyi Zhang’, and writes the the information of the highest and lowest rated movies of that actor/actress on Rotten Tomatoes into different txt files including the **movie’s title, introduction, director, cast, rating, genre.** Then it generates a new txt file which contains the critic reviews and audience reviews of each movie. Finally, it creates an output txt file to analyze the word frequency in the movie comments and reviews. The program tries to find if there is some pattern in the comments of a movie star’s best and worst works.

The program can also generate a html file which shows the network of the movie star’s collaboration with other actors or actresses. Or this html file is to show the network representing the weak and strong relationships between the actor/actress and various movies genres.

Expected Deliverables:

A txt file including the information of the highest rated movies of that actor.

A txt file including the information of the lowest rated movies of that actor.

A txt file containing the reviews of the highest rated movies of that actor.

A txt file containing the reviews of the lowest rated movies of that actor.

A txt file analyzing the word frequency of the comments of the actor’s best works.

A txt file analyzing the word frequency of the comments of the actor’s worst works.

A html file to show the collaboration relationships of the actor or his movie genres as a network.

Steps:

1. Get the actor/actress tag from the raw\_input and match it with the url like ‘https://www.rottentomatoes.com/celebrity/zhang\_ziyi’.
2. Analyze the html structure and obtain the links of the best movies of that actor/actress. Go to each movie’s web page using the links and read it. Then open a new txt file to grab the movie information.
3. Write the comments into an another txt file while looping each movie’s web page.
4. Do the same thing for the actor’s worst movies.
5. Choose the appropriate method to analyze the word frequency of the movies’ comments. Generate a txt file to show the results.
6. While processing the web page of each movie, count the cast or the genres. Make a network data visualization as an output of html format.

Challenges:

1. How many movies I need? How many comments for each movie I should take? I would try some name tags and different parameters to find the most proper number.
2. How to write the function to calculate the word frequency? I would take a look at others’ open sourced code to learn their algorithms or may borrow their ideas.
3. How to deal with the structure of web age? I need to learn some libraries like urllib, BeautifulSoup by myself.