SI 618 Project Proposal

**Introduction**

This project explores what we associate to women and men in our “dreams”, films. It tries to find out what kind of film plots are related to femininity and what are related to masculinity. If 60% or more characters in a film are female, we categorize it as a “woman’s film”, and if 60% or more are male, we categorize it as a “man’s film”. We find out the plots of these films and see what kind of keywords are related to woman’s film and what kind are related to man’s across time. This project will help us understand femininity and masculinity in films, an important kind of mass media in human culture, and how they evolve throughout the history.

**Data sources**

The first dataset I intend to use is the IMDB 5000 movie dataset (chuansun76, 2016) from Kaggle.com. It contains data from IMDB.com including movie titles, movie links, plot keywords, and years made. The second data set I intend to use is two files from Polygraph's Film Dialogue Dataset (Daniels, 2016). One is character\_list5.csv, which contains script id of a film, IMDB character names, and gender of the characters. The other is meta\_data7.csv, which contains script ids, the corresponding IMDB ids, titles of the films, and years.

**Data manipulation**

First of all, I intend to count the percentage of female characters for all films in the Polygraph’s Film Dialogue Dataset. I will do this by importing the csv data into Python, count the number of characters that are marked as female for each script id. I will use this to determine which films are “female” and which are “male”. Next, I will look up the IMDB ids of the films in meta\_data7.csv with script ids. I will use the IMDB ids to find the corresponding plot keywords for each film in the IMDB 5000 movie dataset by looking up each film through their IMDB links, since the IMDB links contain their IMDB ids. Finally, I will count the occurrence of keywords in different years for woman’s and man’s films respectively and find out the most popular one for each gender every five years.

**Visualization**

I intend to do a movie plot timeline. The x-axis represents time, and plot keywords for women and men are inside shapes with different colors in the corresponding time period. The number of occurrences of a plot keyword determines the sizes of shapes. Depending on the final result, I may shorten or lengthen the time interval. The final result may look like this: <http://music-timeline.appspot.com>.

References

Chuansun76. (2016). Retrieved from https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

Daniels. (2016). GitHub - matthewfdaniels/scripts. Retrieved from https://github.com/matthewfdaniels/scripts