**Deadline: 3/13, 11:59pm**

**5% of final course grade**

**The goal of the project proposal is to outline your plan for the final project in order to get instructor approval. You should turn in a 1 page document plus references (single spaced; 1 inch margins, Times New Roman, 12pt font) describing your initial plan. You should describe the scope and goal of your project and explain why it is an important/interesting problem.**

**Make sure to include:**

* the goal of the project
* the problem you intend to solve
* 2-3 references of prior work that relates to your topic
  + <http://www.aclweb.org/anthology/W16-0204> (Using NLP to distinguish the gender of speaker in Film Dialogue, words, also the gender they speak to)
  + <https://unora.unior.it/retrieve/handle/11574/177394/41377/9788899982768.pdf#page=87> (Gender stereotypes in Movie Dialogue)
  + <https://www.jstor.org/stable/pdf/41679911.pdf?casa_token=0mVEhV9Uy68AAAAA:lBL41qx0f9B3c-AftVV0n9RsPolWLxjaetZ6Byib6_uqlJSf8Oy-pwZiqVcnuovqJbkO2za99R53V1nJZzr1qfegQKgZwdA6MOO9zLgCufNw07Z3rQ> (been to demonstrate how techniques from corpus linguistics might be employed to uncover some of the prototypical stylistic characteristics of dialogue in blockbust)
  + <http://cs224d.stanford.edu/reports/aashna.pdf>
  + <http://www.cs.cornell.edu/~cristian/papers/chameleons.pdf> (useful but I don’t quite understand yet, they also mentioned the difference between gender in initiator and responder)
  + The pudding analysis of movie dialogue: <https://pudding.cool/2017/03/film-dialogue>

https://github.com/matthewfdaniels/scripts

* + The superman speaks, the wonder woman keeps quiet: <http://www.diva-portal.org/smash/get/diva2:1290736/FULLTEXT01.pdf> (linguistic markers are useful)
  + Automatic Identification of Character Types (Character profiling):

<https://www.tandfonline.com/doi/pdf/10.1080/08839514.2017.1289311?needAccess=true>

* Bechdel test

http://bechdeltest.com/

* data you plan to use
* overview of the proposed approach
* evaluation plan (how you will measure the success of your work?)
* timeline for project completion

**Idea 1:**

Analyze gender in movies dialogs

<https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html>

<https://www.kaggle.com/c/movie-review-sentiment-analysis-kernels-only/>

Variables:

* Movie Genre
* Release Year
* Lines/dialogue (I have confusions how to define a dialogue when it ends?)
* The movie ratings (discover common features in famous movies). (should we consider the number of votes as a weight of rating?)
* Characters involved in the dialog:
  + Gender.
  + Gender of another person in conversation.
* Length.
* Topic.

Questions:

* How are the dialogs distributed among the characters?
* Do movies of a specific genre have common dialog styles?
* Sentiment analysis for dialogues (or we can pair this dataset with a movie review dataset, and see if specific dialogues trigger bad/good reviews?).
* Explore the average length of words in a specific genre, the average length of sentences and their connection to how sophisticated the scripts of the movies are? And how it changed with time, genre, rating score and character genders?

Hypothesis:

1. Female characters dialogue topics have changed through time, it became more diverse. (define topics, how you will define, and how you are classifying diverse)
2. Does the dialogue of the female characters differ across genres? For this study, we will specifically look at if the number of times they start a dialogue vs respond to someone else is different across genres. We will also examine if the length (DEFINE LENGTH) of their dialogues differ across genres.
3. Does the number of lead roles that are female vs male differ across genre?
   1. (This will be answered with same analysis as above.)
4. Change with time (Older movies would have more sophisticated scripts and dialogs than newer movies)
   1. Define sophisticated
5. Do movies with higher ratings tend to have more males with leading roles than females with leading roles?
   1. Make sure to be clear in your paper that you know this assumes there aren’t other factors coming into play and note what those possible other factors are, such as genre. You could run a stat to assess if any of the extraneous variables are coming into play.

Plan:

* Explore the dataset and examine a few data points.
* Read the literature on analyzing this dataset or other similar dialogue datasets.
* Read the literature on analyzing human dialogue (not only in movies).

We suppose that movies are actually a projection of real life. Through the study of movie dialogues, we may get a hint of the trend of thought in different eras and different societies.

Exploratory studies?

Add March 8th, on the 5th hypothesis, we could discuss it in three aspects (per genre):

1. Female lead or Male lead
2. Considering over all characters in one movie, rating changes with (#female characters)/(#male characters + #female characters), (#male characters)/(#male characters + #female characters)
3. Considering over all characters in one movie, rating changes with (#lines of females)/(#lines of males + #lines of females), (#lines of males)/(#lines of males + #lines of females)

The 2) and 3) are discussing do high rated movies are more diverse in gender.

In the review part, we could discuss in two ways: previous study of female position in movies (may be with different technologies); previous study of movie dialogues (may not focus on gender gap, and may be more technical)

In the proposal part, we could describe our plan in two levels: basic and fancy. We could start our study with the basic way, and if we have spare time, we could upgrade our technology to do a more sophisticated study. For example

1. How do we define the lead role? The basic way is to use the person with the most lines, the fancy way is to do a graph with all characters and find the most important person (like speak to the largest number of characters, or linked/mentioned by the most times).
2. How do we find the gender of characters? The basic way is to use the labels in the database. There are two more ways: use the gender distinguish tech in the reference paper(<http://www.aclweb.org/anthology/W16-0204> ); use link the Cornell DB to IMDB database and find the actual gender for the person act this character.

Extra:

Older movies would have more sophisticated scripts and dialogs than newer movies.

Define sophisticated, Change with time

Useful References:

COMS 6998-7: Empirical Methods of Data Science

**Final Project Proposal**

Amal Alabdulkarim (aa4235), Jing Qian (jq2282)

# **Project Goa**l

Movie dialogues are a projection of people’s interactions in real life. We can use these dialogues to understand the actual human interactions and get a hint of the trend of thought and its change through time. In this project, we want to understand the female character roles in movies through studying these dialogues. We are going to analyze the change of the dialogues through time, genre and rating of movies.

# 

# **The Problem**

To achieve the goal of the project we want to answer the following questions:

* How do female character topics of dialogues change with time?
* Does the dialogue of the female characters differ across genres?
* Does the number of lead roles that are female vs male differ across genre?
* Do movies with higher ratings tend to have male-leading roles than female-leading roles?

## Hypotheses:

1. Female characters dialogue topics became more diverse (using the Bechdel test) in recent movies compared to older ones.
2. Female characters have a more powerful appearance in specific genres.
   1. Females tend to have more lines in dramatic and romantic movies and fewer lines in other genres than male characters.
   2. Females tend to have more significant roles in dramatic and romantic movies and fewer in other genres.
3. Movies with higher rating tend to have a stronger male presence.

# **Prior Work**

There are plenty of gender studies of movie dialogues. In 2012, Mclntyre performed a comprehensive analysis of the prototypical stylistic characteristics of dialogue in blockbuster movies using multiple techniques from corpus linguistics. Schofield and Mehr (2016) used NLP to distinguish the gender of the speaker in film dialogues in the same dataset we will use and found differences between single-gender and two-gender conversations and gendered speech. Basili, Nissim, and Satta (2017) studied the gender stereotype in movie dialogues and concluded that movie languages portray the stereotype that men and women talk on recognizable traits attached to femininity and masculinity.

# **Dataset**

For this project, we will be using the Cornell Movie Dialogs Corpus in Danescu-Niculescu-Mizil (2011). This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts:

* 220,579 conversational exchanges between 10,292 pairs of movie characters
* Involves 9,035 characters from 617 movies
* In total 304,713 utterances

Movie metadata included:

* Genres
* Release year
* IMDB rating
* Number of IMDB votes
* Character metadata included:
  + gender (for 3,774 characters)
  + position on movie credits (3,321 characters)

Because our hypotheses are gender-related, we will extract only the 3,774 characters with gender data and their utterances.

# **Method and** Evaluation Plan

For the first hypothesis, to test the diversity of the female character dialogues, we define diversity in dialogue in two main aspects. The first aspect is when two female characters are talking to each other, they will be talking about something other than men. The second aspect is whether the female character conversations span across different topics. To test the first aspect, we are going to use the Bechdel test, which has this in its third criteria and shows which movies pass this test. To evaluate the second aspect, we will look more detailed into the dialogues, using the latent Dirichlet allocation (LDA) for topic modeling and comparing the topics in female-male, male-male, female-female and male-female dialogs and female monologues.

The second hypothesis has two sub-hypotheses and the both can be tested with similar statistical analysis. We will specifically look at if the number of times female character start a dialogue vs responding to someone else is different across genres. We will also examine if the length of their dialogues (measured by the number of words and number of lines) differ across genres. We will also identify lead characters in those movies by the length of their dialogues and their mentions in other characters’ dialogues.

For the third hypothesis, we will measure how the rating of the movie relate to the female character role (identified using the previous statistical analysis for the second hypothesis), the ratio of female to male characters in the movies and ratio of female to male characters dialogues lengths (measured by number of words and number of lines).

In evaluating these hypotheses, we will be testing for statistical significance. And we will also consider the other factors and extraneous variables that may affect the findings of the analysis and make sure we address them correctly. We will use graphs and exploratory visualization for the results and the data to discover the relations and patterns and explain the findings.

# 

# **Timeline**

3/13 Submit a project proposal

3/20 Download dataset and take a first exploration of the whole data

3/31 Finish literature review

4/15 Test Hypothesis 1 and Hypothesis 2, and get plots of corresponding results

4/25 Test Hypothesis 3 and get plots of corresponding results

4/28 Analyse result and finish the first draft of the project report

4/30 Finish the presentation slides

5/1 In class presentation and modify report according to the response from Professor

Levine and classmates

5/3 Submit final project

# **References**

Dan McIntyre (2012). Prototypical Characteristics of Blockbuster Movie Dialogue: A Corpus Stylistic Analysis. *Texas Studies in Literature and Language, 54*, 402-425.

Schofield, Alexandra & Mehr, Leo (2016). Gender-Distinguishing Features in Film Dialogue. *Texas Studies in Literature and Language, 54*, 32-39.

Busso, L., & Vignozzi, G. (2017). Gender Stereotypes in Film Language: A Corpus-Assisted Analysis. CLiC-it 2017 11-12 December 2017, Rome, 71.

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (pp. 76-87). Association for Computational Linguistics.