COMS 6998-7: Empirical Methods of Data Science

Final Project Proposal

Amal Alabdulkarim (aa4235), Jing Qian (jq2282)

Project Goal

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.
    a. Females tend to have more lines in dramatic and romantic movies and fewer lines in other genres than male characters.
    b. 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.