Kyleigh Harris (Student ID 301571408)

Yolanda Setiawan (Student ID 301589647)

**LING 450 - Computational Linguistics**

**Project Proposal**

**Description of project:**

This project will extract dialogue from movie scripts to see how different genres compare regarding whether they pass the Bechdel test or not. The Bechdel test evaluates the representation of women in movies by asking the following question: “Does the movie feature any scenes in which two women are talking to each other about anything other than men?” We will implement some of the approaches from Assignment 1 such as utilizing spaCy matcher to look for patterns in the movie’s dialogue and the Gender Gap Tracker to compare and identify which provided the most accurate output. The genres that will be evaluated include: adventure, action, comedy, drama, and thriller, in which each genre will contain 5 movies each. The movies will be selected from IMDb and each genre will contain 1 movie from each of the following decades: 1980s, 1990s, 2000s, 2010s, and 2020s.

**Project Objectives: A List of Tasks and Deliverables**

For this project, we are hoping to run each movie script through our code from Assignment 1 to extract dialogue from only the female characters in the movie. The movie scripts will be extracted from an internet movie script database (https://imsdb.com/). We will take this text and use spaCy and regular expressions to identify sentences that *do not* contain references to men such as by searching for sentences that do not contain male characters’ names or contain the pronouns “he”, “him”, or “his”. If there is any dialogue that match these specifications, then we will be able to deduce that the movie passes the Bechdel test. 

**Schedule for the Tasks and Deliverables:**

We are still fine-tuning what our schedule will look like in terms of tasks and deliverables but a general idea of how this project will unfold is as follows: (1) Compiling movie scripts into datasets for each genre - much like how we had to compile five text files into a single variable to be read in Approach 1 of Assignment 1; (2) using feature engineering, we will adjust the code for each genre to extract personal pronouns, using the context of sentences in which personal pronouns are used to find out which names are from male characters and which are from female characters; (3) using our indirect quotes code from Assignment 1, we will extract the dialogue from female characters in each movie script; (4) we will use regular expressions to find matches within female characters’ dialogue and search for dialogue that does not include any references to men or the male characters in the movie; (5) from this data, we will be able to see which movies pass the Bechdel test and how these different genres compare in terms of adequate female representation. 

**List of Resources:**

We are still fleshing out how we will be producing our deliverables but the resources we expect to use include our code from Assignment 1, text files of movie scripts, spaCy, and online resources for trouble-shooting and improving the accuracy of our code. 

**Anticipated Risks:**

Some risks that we anticipate include having to completely start from scratch on our code if the code is not able to extract what we need from the movie scripts due to the way they are formatted. We might also have to consider the fact that our code can incorrectly extract dialogue from male characters who share a name with a female character in a movie from the same genre. This would require us to rethink grouping all movies of the same genre into a single variable to be read. 

**Division of labor:**

For this project, we will be dividing tasks seen through the table below:

|  |  |
| --- | --- |
| Tasks | Name |
| (1) Compiling movie scripts into datasets for each genre | Yolanda and Kyleigh |
| (2) Adjusting the code for each genre to extract personal pronouns | Yolanda |
| (3) Extract the dialogue from female characters in each movie script | Kyleigh |
| (4) Use regular expressions to find matches within female characters’ dialogue and search for dialogue that does not include any references to men or the male characters in the movie | Yolanda and Kyleigh |
| (5) Analyze and compare data | Yolanda and Kyleigh |

**Communication plan:**

We will have clear and consistent communication throughout the project to ensure a collaborative completion. We will be communicating mainly on messages to help each other out for troubleshooting, and progress updates to ensure the project is completed on time. For the code, we will be creating a github to compile all of our work and plan on combining the notebooks at the end of our project to make it organized. For the write-up, we will collaborate on Google Docs to complete the project.