Text Analysis Visualization of Gender Representation in Hollywood Movies

2023-04-30

INTRODUCTION

We conducted text analysis of plot descriptions of movies to compare the use of words between films that have good and bad female representation. We used the Bechdel test as a measure of female representation, which looks at whether a movie has at least two named female characters who talk to each other about something other than a man. By analyzing the plot descriptions, we aimed to understand how movies with good and bad female representation differ in the words they use. Even though the Bechdel test isn't the sole indicator of female representation in films, this analysis can provide insights into how women are portrayed in movies and how their representation can impact the way we perceive and value women in society.

Data

```
setwd("C:/Users/ASUS/Desktop/dataviz final data")

# Import Data
raw_bechdel <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da
```

```
## Rows: 8839 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): imdb_id, title
## dbl (3): year, id, rating
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
movies <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20
```

```
## Rows: 1794 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (24): imdb, title, test, clean_test, binary, domgross, intgross, code, d...
## dbl  (7): year, budget, budget_2013, period_code, decade_code, metascore, im...
## num  (1): imdb_votes
## lgl  (2): response, error
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
data <- merge(raw_bechdel, movies, by = "title")

# Filter out 0 in ratings
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data_filtered <- data %>%
  filter(rating != 0)
```

Word Cloud

We chose 500 movies that pass the Bechdel Test and compare with 500 movies that failed the Bechdel test. We will analyse the text from "plot". We use the cleaning functions to remove unnecessary words (stop words), syntax, punctuation, numbers, white space, etc. We also creates a document-term-matrix, and provided word clouds of the most frequent words among the movies that pass and fail the Bechdel test.

Word Cloud for movie plots that passed the Bechdel Test

```
library(dplyr)
library(wordcloud2)
library(tidytext)
library(stringr)
library(plotrix)
library(wordcloud)
```

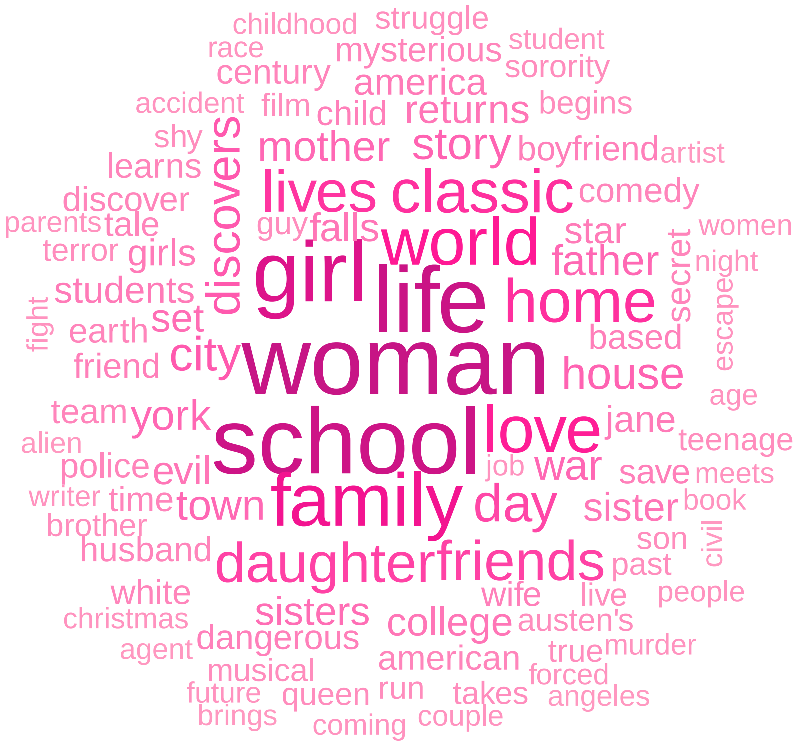
```
## Loading required package: RColorBrewer
```

```
library(tm)
```

```
## Loading required package: NLP
```

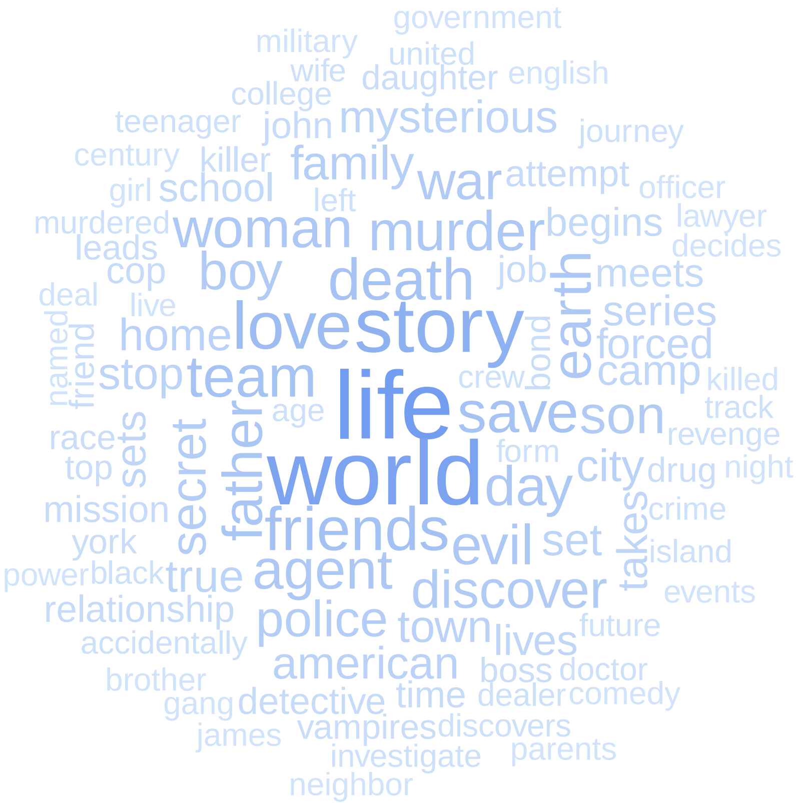
```
# Select the 500 movies that pass the Bechdel Test and 500 movies that failed the Bechdel test
pass <- data_filtered %>%
  arrange(desc(binary)) %>%
  head(500)

fail <- data_filtered %>%
  arrange(binary) %>%
  head(500)
```

```
# Combine pass and fail into one dataframe
all_bechdel <- bind_rows(pass,fail)

# Clean the text for PASS
clean_text_pass <- pass %>%
  select(plot) %>%
  unnest_tokens(word, plot) %>%
  mutate(word = str_replace_all(word, "[^[:alnum:]']", "")) %>%
  filter(!word %in% stop_words$word) %>%
  filter(!str_detect(word, "^\\d+$")) %>%
  mutate(word = str_to_lower(word)) %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 100)

# Create a Document-Term-Matrix
movie_dtm <- DocumentTermMatrix(pass)

# Define a custom color palette of different shades of pink
pink_palette <- colorRampPalette(c("#FFC0CB", "#FF69B4", "#FF1493", "#C71585"))

# Create a word cloud with the custom pink color palette
wordcloud(words = clean_text_pass$word, freq = clean_text_pass$n, scale = c(3, 0.5),
          random.order = FALSE, colors = pink_palette(length(clean_text_pass$word)))
```



Word Cloud for movie plots that failed the Bechdel Test

```r
# Clean the text using the functions introduced in lecture
clean_text_fail <- fail %>%
  select(plot) %>%
  unnest_tokens(word, plot) %>%
  mutate(word = str_replace_all(word, "[^[:alnum:]']", "")) %>%
  filter(!word %in% stop_words$word) %>%
  filter(!str_detect(word, "^\\d+$")) %>%
  mutate(word = str_to_lower(word)) %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 100)

# Create a Document-Term-Matrix
movie_dtm <- DocumentTermMatrix(fail)

# Define a custom color palette of different shades of blue
blue_palette <- colorRampPalette(c("#E6F3FF", "#BFD6F8", "#99B9F2", "#739DF0"))

# Create a word cloud with the custom blue color palette
wordcloud(words = clean_text_fail$word, freq = clean_text_fail$n, scale = c(3, 0.5),
          random.order = FALSE, colors = blue_palette(length(clean_text_fail$word)))
```



Based on the word clouds, we found that the most common words in the plot of movies that passed the Bechdel test are woman, school, life, girl, family, home, world, love, classic and daughter. Common words in the plot of movies that failed the Bechdel test are life, world, story, save, friends, team, and death.

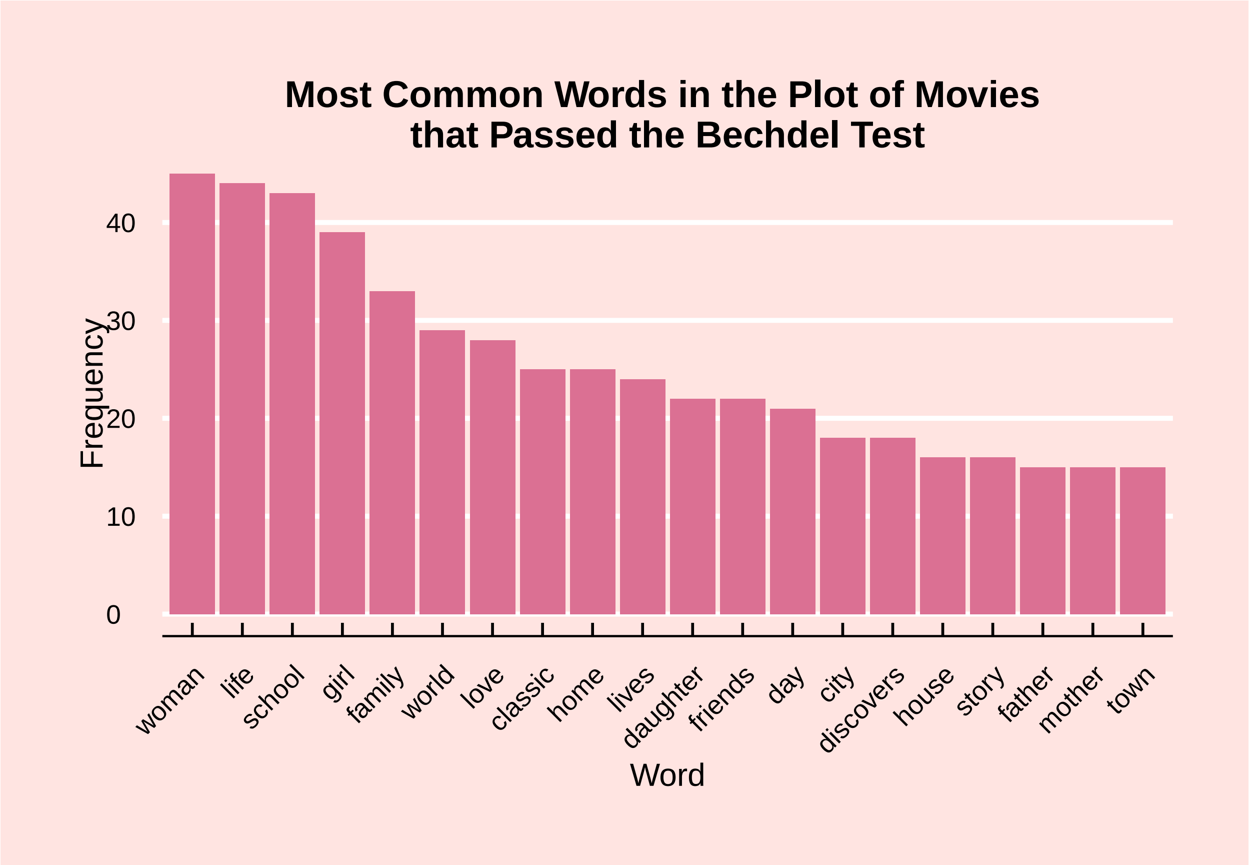
In the next stage of our analysis, we presented bar graphs displaying the frequency of the top 20 most commonly used words, along with their respective counts, to provide further details about the distribution of these words. Our analysis includes three visual representations: bar graphs for both passed and failed Bechdel Test categories, as well as a pyramid plot that compares the most common words used in movies that passed and failed the Bechdel Test. By presenting these visualizations side by side, we can easily compare and contrast the frequency and type of words used in the plot descriptions of movies that passed and failed the test.

Bar graph for movie plots that passed the Bechdel Test

```
library(ggplot2)
```

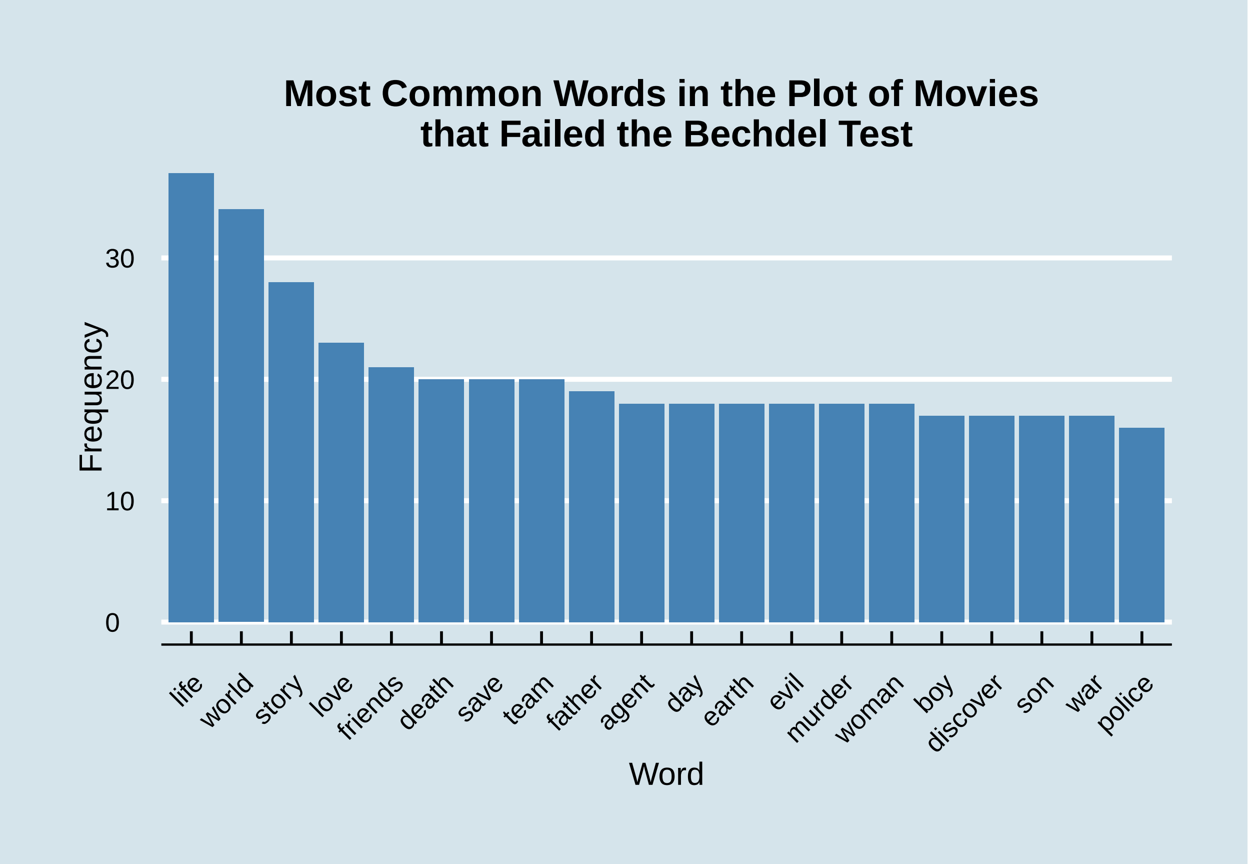
```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##     annotate
```

```
library(ggthemes)

# Create a data frame of the most common words in the plots of movies that passed the Bechdel Test
clean_text <- pass %>%
  select(plot) %>%
  unnest_tokens(word, plot) %>%
  mutate(word = str_replace_all(word, "[^[:alnum:]']", "")) %>%
  filter(!word %in% stop_words$word) %>%
  filter(!str_detect(word, "^\\d+$")) %>%
  mutate(word = str_to_lower(word)) %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 20)

# Create a bar graph of the most common words
ggplot(clean_text, aes(x = reorder(word, -n), y = n)) +
  geom_col(fill = "palevioletred") +
  labs(x = "Word", y = "Frequency", title = "Most Common Words in the Plot of Movies \nthat Passed the B
  theme_economist() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1, size = 10),
        axis.text.y = element_text(size = 10),
        axis.title = element_text(size = 12),
        plot.title = element_text(size = 14, hjust = 0.5),
        plot.margin = unit(c(1, 1, 1, 1), "cm"),
        panel.background = element_rect(fill = "mistyrose"),
        plot.background = element_rect(fill = "mistyrose"))
```



Bar graph for movie plots that failed the Bechdel Test

```
library(ggplot2)
library(ggthemes)

# Create a data frame of the most common words in the plots of movies that passed the Bechdel Test
clean_text <- fail %>%
  select(plot) %>%
  unnest_tokens(word, plot) %>%
  mutate(word = str_replace_all(word, "[^[:alnum:]']", "")) %>%
  filter(!word %in% stop_words$word) %>%
  filter(!str_detect(word, "^\\d+$")) %>%
  mutate(word = str_to_lower(word)) %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 20)

# Create a bar graph of the most common words
ggplot(clean_text, aes(x = reorder(word, -n), y = n)) +
  geom_col(fill = "steelblue") +
  labs(x = "Word", y = "Frequency", title = "Most Common Words in the Plot of Movies \nthat Failed the
  theme_economist() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1, size = 10),
        axis.text.y = element_text(size = 10),
        axis.title = element_text(size = 12),
        plot.title = element_text(size = 14, hjust = 0.5),
        plot.margin = unit(c(1, 1, 1, 1), "cm"))
```

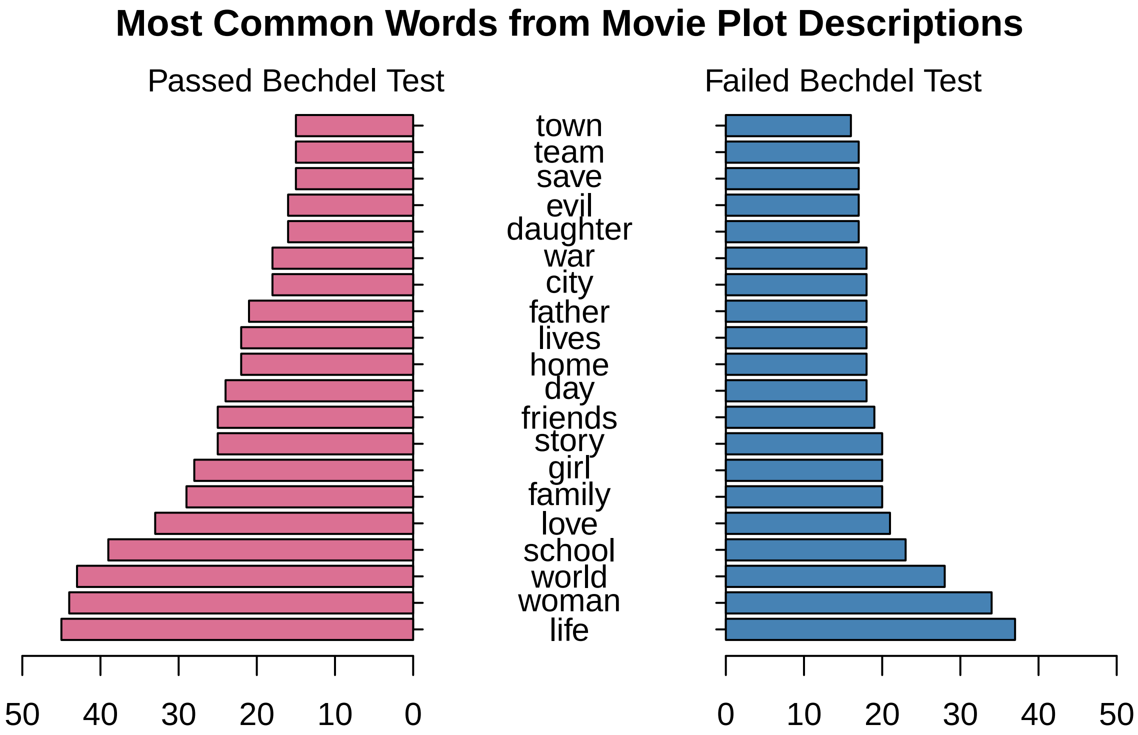


Pyramid Plot

We provide a pyramid plot to show how the words between passing and failed Bechdel test movies differ in frequency. A selection of 20 top words are chosen.

```
library(dplyr)
library(tidytext)
library(stringr)
library(plotrix)

pass <- data_filtered %>%
  arrange(desc(binary)) %>%
  head(500)

fail <- data_filtered %>%
  arrange(binary) %>%
  head(500)

# Combine pass and fail into one dataframe
all_bechdel <- bind_rows(pass,fail)

# Clean the 20 words
clean_text <- all_bechdel %>%
  select(plot) %>%
  unnest_tokens(word, plot) %>%
```

```r
  mutate(word = str_replace_all(word, "[^[:alnum:]']", "")) %>%
  filter(!word %in% stop_words$word) %>%
  filter(!str_detect(word, "^\\d+$")) %>%
  mutate(word = str_to_lower(word)) %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 20)

# Create a data frame of the most common words in the plots of movies that passed the Bechdel Test
clean_text_pass <- pass %>%
  select(plot) %>%
  unnest_tokens(word, plot) %>%
  mutate(word = str_replace_all(word, "[^[:alnum:]']", "")) %>%
  filter(!word %in% stop_words$word) %>%
  filter(!str_detect(word, "^\\d+$")) %>%
  mutate(word = str_to_lower(word)) %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 20)


# Create a data frame of the most common words in the plots of movies that passed the Bechdel Test
clean_text_fail <- fail %>%
  select(plot) %>%
  unnest_tokens(word, plot) %>%
  mutate(word = str_replace_all(word, "[^[:alnum:]']", "")) %>%
  filter(!word %in% stop_words$word) %>%
  filter(!str_detect(word, "^\\d+$")) %>%
  mutate(word = str_to_lower(word)) %>%
  count(word, sort = TRUE) %>%
  slice_head(n = 20)

# Create the pyramid plot
par(mar=c(5,5,2,2))
pyramid.plot(clean_text_pass$n, clean_text_fail$n,
             labels=clean_text$word,
             main="Most Common Words from Movie Plot Descriptions",
             lxcol= "palevioletred", rxcol= "steelblue", gap=20,
             top.labels = c("Passed Bechdel Test", " ", "Failed Bechdel Test"),
             xlim=c(0,50),
             laxlab = seq(from = 0, to = 50, by = 10),
             raxlab = seq(from = 0, to = 50, by = 10),
             mtext(" ", side = 1, line = 5, col = "black", cex = 1.2))
```



```
## 50 50
```

```
## [1] 5 5 2 2
```

FINAL ANALYSIS

Based on our analysis of movie plots, we compared the frequency of words used in movies that passed and failed the Bechdel Test as a measure of female representation. The Bechdel Test is a standard to evaluate female representation in movies that requires a movie to have at least two named female characters who talk to each other about something other than a man. The Bechdel Test has become a widely recognized standard for evaluating the representation of female characters in movies, providing a simple yet effective metric to evaluate whether women are depicted as fully formed characters who have conversations about topics other than men. However, it is important to note that a movie that fails the Bechdel Test should not be automatically labeled as anti-feminist or problematic. Instead, the test serves as a tool for critical analysis and a starting point for further examination of gender representation in media. By using the Bechdel Test as a measure of female representation in movies, we can better understand the patterns and biases that may exist in media and work towards creating more equitable and diverse representation of women in film.

We found that movies with better female representation (passed Bechdel Test) had more common words in the plot like “woman”, “school”, “life”, “girl”, “family”, which implies themes of domesticity. In contrast, movies with less female representation (failed Bechdel Test) had plots with common words like “life”, “world”, “story”, and “team”, which implies more adventure-driven plots. These findings suggest that there are differences in the types of movies that pass or fail the Bechdel Test, which may be related to the representation of female characters in the movies. The themes that we observed in the movies that passed or failed the

Bechdel Test reflect certain gender norms that have been perpetuated in media. The use of words like "woman", "school", "life", "girl", and "family" in the plots of movies that passed the Bechdel Test suggests that these movies may be more oriented towards domestic themes that are traditionally associated with women. On the other hand, movies that failed the Bechdel Test tended to have more adventure-driven plots with words like "life", "world", "story", and "team", which aligns with traditional masculine gender norms.

This type of analysis is important because it helps to shed light on potential differences in how women are represented in media, and how this representation might differ based on whether a movie passes or fails the Bechdel Test. By understanding these differences, we can start to identify potential biases and patterns that may exist in media and work towards creating more equitable representation of women in movies. Additionally, this analysis can contribute to a broader conversation around gender and media representation, and help us to better understand the ways in which media shapes our perceptions and expectations of gender roles.