Analysis of the original Star Wars Trilogy

Breckin Hadley

This mini project uses text analysis and sentiment scoring to analyze the scripts of the original Star Wars trilogy (Episodes 4–6). By breaking up the dialogue and analyzing it using differnt sentiments (Bing, NRC, and AFINN), we can see the different trends in how emotions, word choice, and character dialogue changes throughout the original trilogy.

The project focuses on several questions: What words in the movies wave the highest positive or negative connotation? How are specific emotions, like fear and trust, expressed throughout the trilogy? Which characters have the most lines, and what words do they use the most? How does the emotional tone change throughout each movie, and how do the three movies compare to each other?

```
#Readin Star Wars data
library(readr)
SW <- read_csv("SW_Episode4-6.csv")
```

```
Rows: 2523 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): name, line

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.
```

```
#Tidy Data and adding the movie column
SW_tidy <- SW |>
  mutate(movie = case_when(
    row_number() <= 1001 ~ "new_hope",
    row_number() >= 1002 & row_number() <= 1894 ~ "empire_strikes",
    row_number() >= 1895 ~ "return_jedi"
  )) |>
  unnest_tokens(output = word, input = line)
```

```
SW_tidy
```

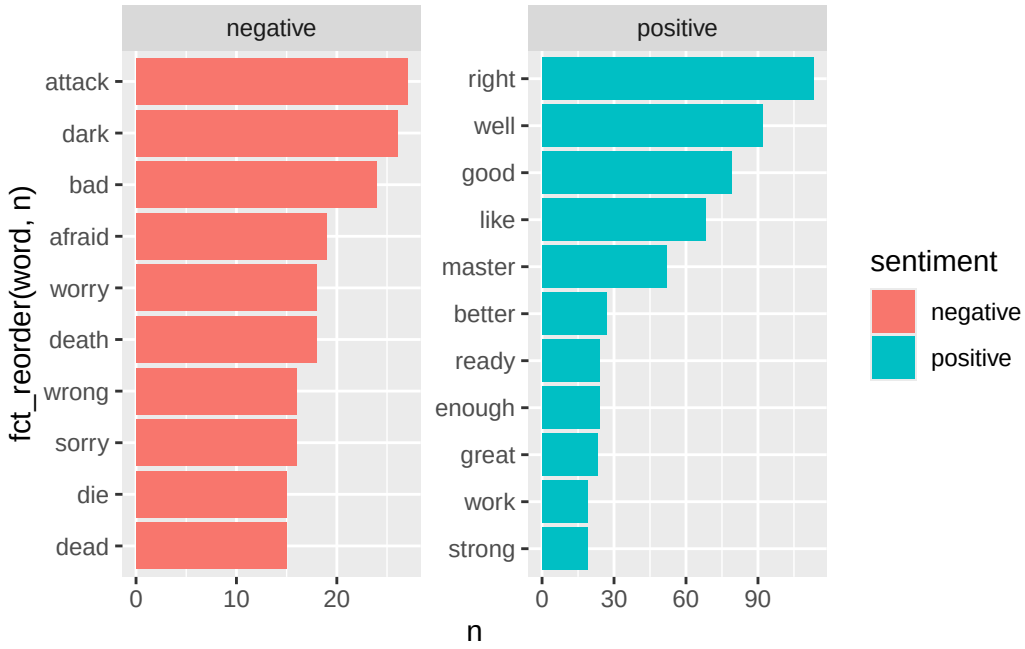
```
# A tibble: 25,937 x 3
   name     movie    word
   <chr>    <chr>    <chr>
 1 THREEPIO new_hope did
 2 THREEPIO new_hope you
 3 THREEPIO new_hope hear
 4 THREEPIO new_hope that
 5 THREEPIO new_hope they've
 6 THREEPIO new_hope shut
 7 THREEPIO new_hope down
 8 THREEPIO new_hope the
 9 THREEPIO new_hope main
10 THREEPIO new_hope reactor
# i 25,927 more rows
```

Negative and positive sentiment scores

```
SW_tidy|>
  inner_join(bing_sentiments) |>
  count(sentiment, word, sort = TRUE) |>
  group_by(sentiment) |>
  slice_max(n, n = 10) |>
  ungroup() |>

  ggplot(aes(x = fct_reorder(word, n), y = n, fill = sentiment)) +
    geom_col() +
    coord_flip() +
    facet_wrap(~ sentiment, scales = "free")
```

```
Joining with `by = join_by(word)`
```



For this plot I want to know the top 10 most frequent positive words and top 10 most frequent negative words used in Star Wars Episodes 4-6. This code above produces a faceted bar chart to illustrate the results.

Fear and Trust sentiment scores

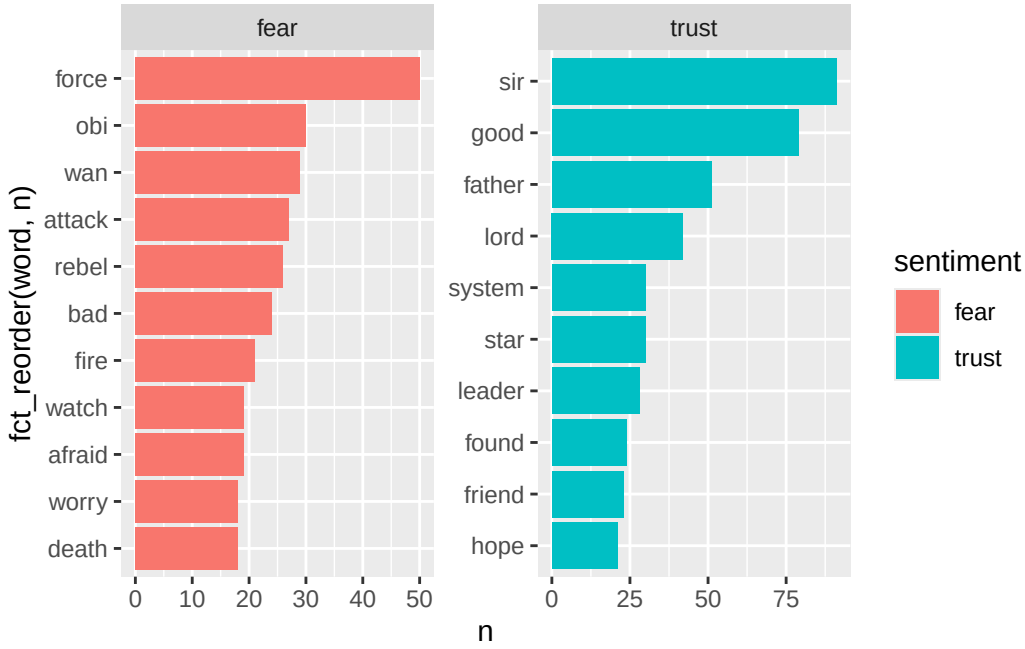
```
SW_tidy |>
  inner_join(nrc_sentiments) |>
  filter(sentiment %in% c("fear", "trust"))|>
  count(sentiment, word, sort = TRUE) |>
  group_by(sentiment) |>
  slice_max(n, n = 10) |>
  ungroup() |>

  ggplot(aes(x = fct_reorder(word, n), y = n, fill = sentiment)) +
    geom_col() +
    coord_flip() +
    facet_wrap(~ sentiment, scales = "free")
```

```
Joining with `by = join_by(word)`
```

```
Warning in inner_join(SW_tidy, nrc_sentiments): Detected an unexpected many-to-many relations
i Row 13 of `x` matches multiple rows in `y`.
```

```
i Row 7863 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

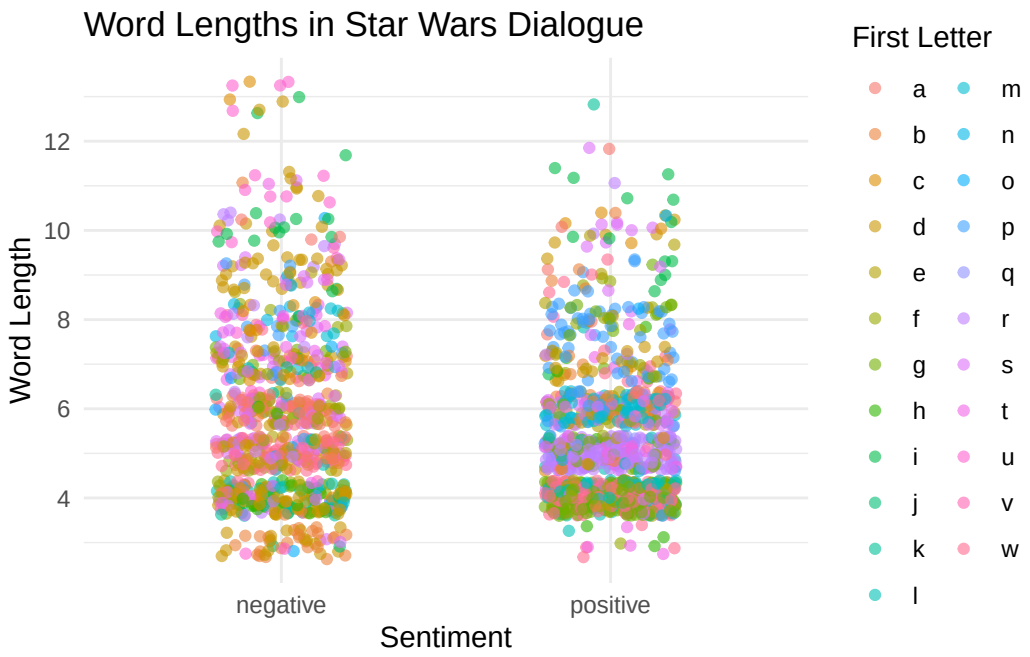


This plot produces similar results as the plot above but looks at the top 10 most frequent trusting words and top 10 most frequent fearful words used in Star Wars Episodes 4-6. This code also uses a faceted bar chart to showcase the results.

#Word Length vs Sentiment

```
SW_tidy |>
  inner_join(bing_sentiments, by = "word") %>%
  mutate(
    word_length = str_length(word),
    first_letter = str_sub(word, 1, 1),
    word_upper = str_to_upper(word)
  ) |>
  ggplot(aes(x = sentiment, y = word_length, color = first_letter)) +
  geom_jitter(width = 0.2, alpha = 0.6) +
  labs(
    title = "Word Lengths in Star Wars Dialogue",
    x = "Sentiment",
    y = "Word Length",
```

```
    color = "First Letter"
  ) +
  theme_minimal()
```



This plot shows how the word length compares to the sentiment while also showing what letter the word starts with. It seems that words that start with f,g,h, and i tend to have a more positive connotation while words starting with s,t,u,v have a more negative connotation.

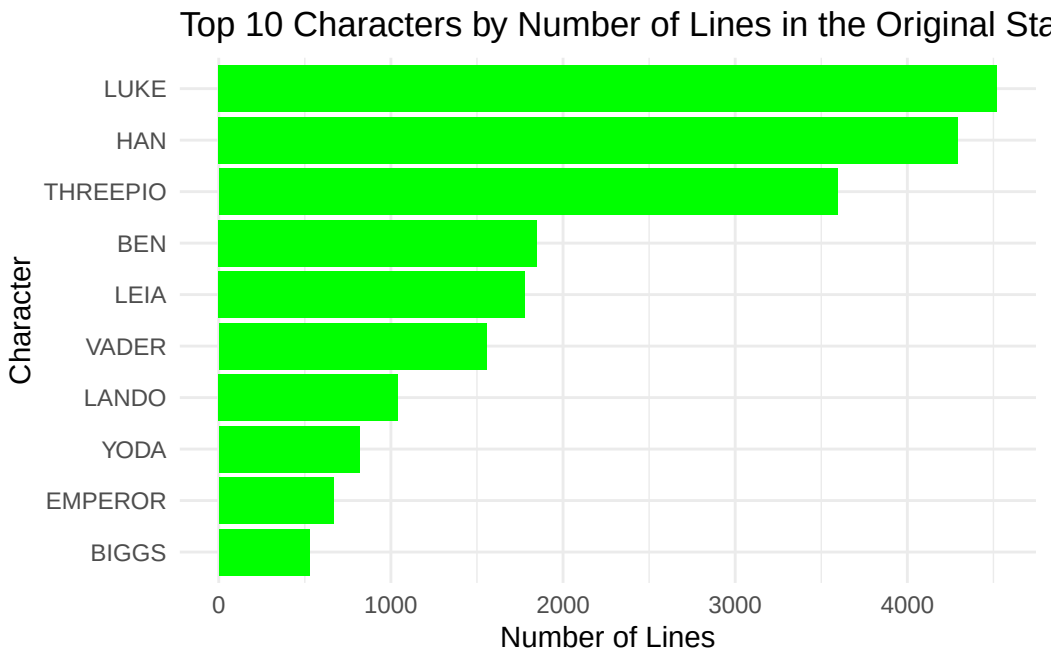
**Who Says the Most lines*

```
#Who Says the Most lines
SW_tidy |>

  count(name, name = "num_lines", sort = TRUE) |>
  slice_max(order_by = num_lines, n = 10) |>

  ggplot(aes(x = reorder(name, num_lines), y = num_lines)) +
    geom_col(fill = "green") +
    coord_flip() +
    labs(
      title = "Top 10 Characters by Number of Lines in the Original Star Wars Trilogy",
      x = "Character",
      y = "Number of Lines"
```

```
  ) +
    theme_minimal()
```



I wanted to compare the number of lines differnt characters had in Star Wars episodes 4-6. I decided to pick the top 10 characters with the most lines and display them on a horizonatal histogram.

Alt Text:

This plot is a horizontal bar chart titled "Top 10 Characters by Number of Lines in the Original Star Wars Trilogy." On the x-axis is the lists characters making the top 10, and on the y-axis we have the number of lines that character speaks through out the trilogy. The horizontal bars are colored green (like Lukes Lightsaber) and the characters are ordered from most lines at the top to least lines at the bottom. The character with the most lines in the trilogy is Luke and the character with the tenth most lines is Biggs.

Most said word by each character

```
SW_tidy |>
  anti_join(stop_words, by = "word") |>
  group_by(name, word) |>
  summarise(line_count = n(), .groups = "drop") |>
  group_by(name) |>
```

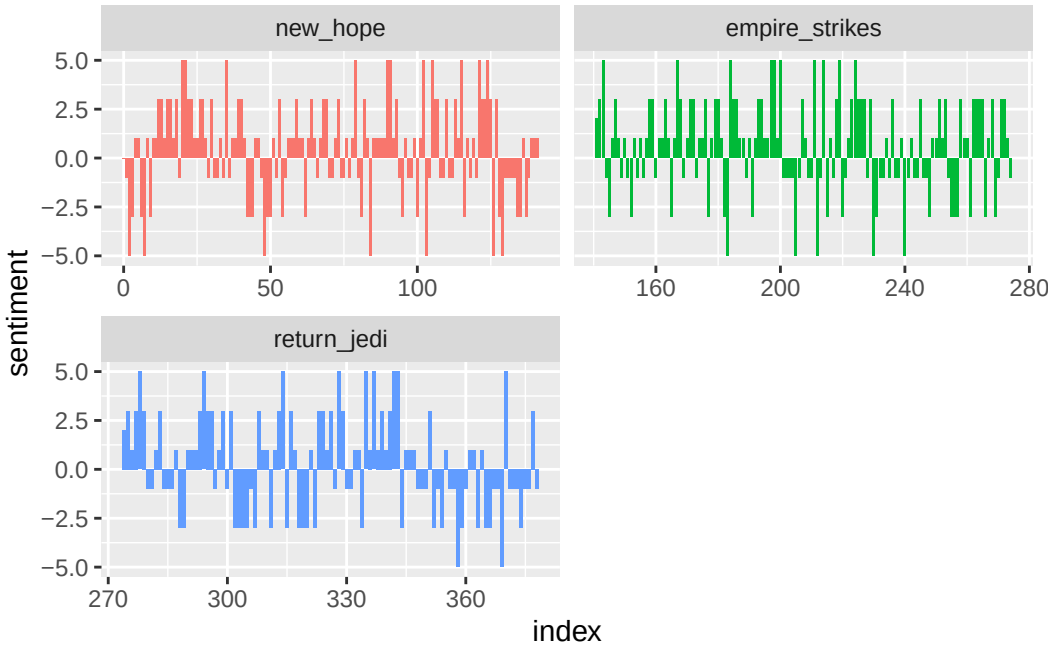
```
  slice_max(order_by = line_count, n = 1, with_ties = FALSE) |>
  arrange(desc(line_count))
```

```
# A tibble: 122 x 3
# Groups:   name [122]
   name     word    line_count
   <chr>    <chr>        <int>
 1 THREEPIO artoo           57
 2 HAN      chewie          46
 3 LUKE     artoo           24
 4 BEN      luke            22
 5 LEIA     luke            16
 6 PIETT    lord            15
 7 YODA     force           14
 8 BIGGS    luke            13
 9 VADER    master          12
10 EMPEROR  friends          7
# i 112 more rows
```

For this code I wanted to find which word was said the most by each character. I had to account for stop_word (the,and,I,me, etc) but this were easily filtered out using anti_join.

Theme throughtout the Trilogy

```
SW_tidy |>
  inner_join(bing_sentiments, by = "word") |>
  count(movie, index = row_number() %/% 5, sentiment) |>
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) |>
  mutate(sentiment = positive - negative,
         movie = factor(movie,
                        levels = c("new_hope",
                                   "empire_strikes",
                                   "return_jedi"))) |>
  ggplot(aes(x = index, y = sentiment, fill = movie)) +
    geom_col(show.legend = FALSE) +
    facet_wrap(~movie, ncol = 2, scales = "free_x")
```



This plot was one of my favorites to make. This is a sentiment trajectory plot for each Star Wars movie which basically shows the emotional trajectory for each of the films. All of the films seem to be relativley similar to one another with one not being dratically differnt from the others.