Sentiment Analysis of Star Wars (4-6) Movie Scripts

Contents

1 Dependencies

```
library(tidyverse)
library(tidytext)
```

2 Reading in clean data

```
sw_scripts <- read_csv("clean_data/original_sw_trilogy.csv")
```

3 Tokenize and remove stop words

I will use single words as my token as I am interested in the sentiments of words.

```
sw_tokens <- sw_scripts %>%
  unnest_tokens(
    word,
    dialogue
  ) %>%
  anti_join(stop_words)

sw_tokens
```

4 Check which sentiment lexicon categorizes most words

Before analysing the sentiment of the text, I want to check which lexicon is able to categorise/rate most words.

Available lexicons in `tidytext::get_gentiments()`:

- **Bing**
  - 2 categories: positive or negative
- **AFINN**
  - 11 ratings: integer between -5 (negative) and +5 (positive)
- **Loughran**
  - 6 ratings: negative, positive, litigious, uncertainty, constraining, or superfluous
- **NRC**
  - 10 categories: 8 basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, or disgust) and 2 sentiments (negative or positive)

```
unique_words <- sw_tokens %>% distinct(word)

# available lexicons in tidytext::get_sentiments()
lexicons <- c("bing", "afinn", "loughran", "nrc")

# create list of joined datasets with available lexicons
nested_df <- lexicons %>%
  map(~left_join(unique_words, get_sentiments(.), by = "word"))

# attach lexicon names to list
names(nested_df) <- lexicons


for (lexicon in lexicons){

  # 2nd element is sentiment category or rating
  sentiments <- nested_df[[lexicon]][[2]]

  # count all values without attached sentiment
  missing <- sum(is.na(sentiments))

  print(str_glue("{lexicon}: {missing} uncategorised words"))

}
```

```
## bing: 1930 uncategorised words
## afinn: 2018 uncategorised words
## loughran: 2170 uncategorised words
## nrc: 1690 uncategorised words
```

Seems like the NRC lexicon is able to categorise most words, so I will use it for my sentiment analysis.

5 Visualisations of sentiments

5.1 Set theme for all plots

```
theme_set(theme_minimal() +
            theme(
              strip.text = element_text(size = 10, face = "bold"),
              axis.text = element_text(size = 8, face = "bold"),
              axis.title = element_text(size = 12, face = "bold")
              )
          )
```

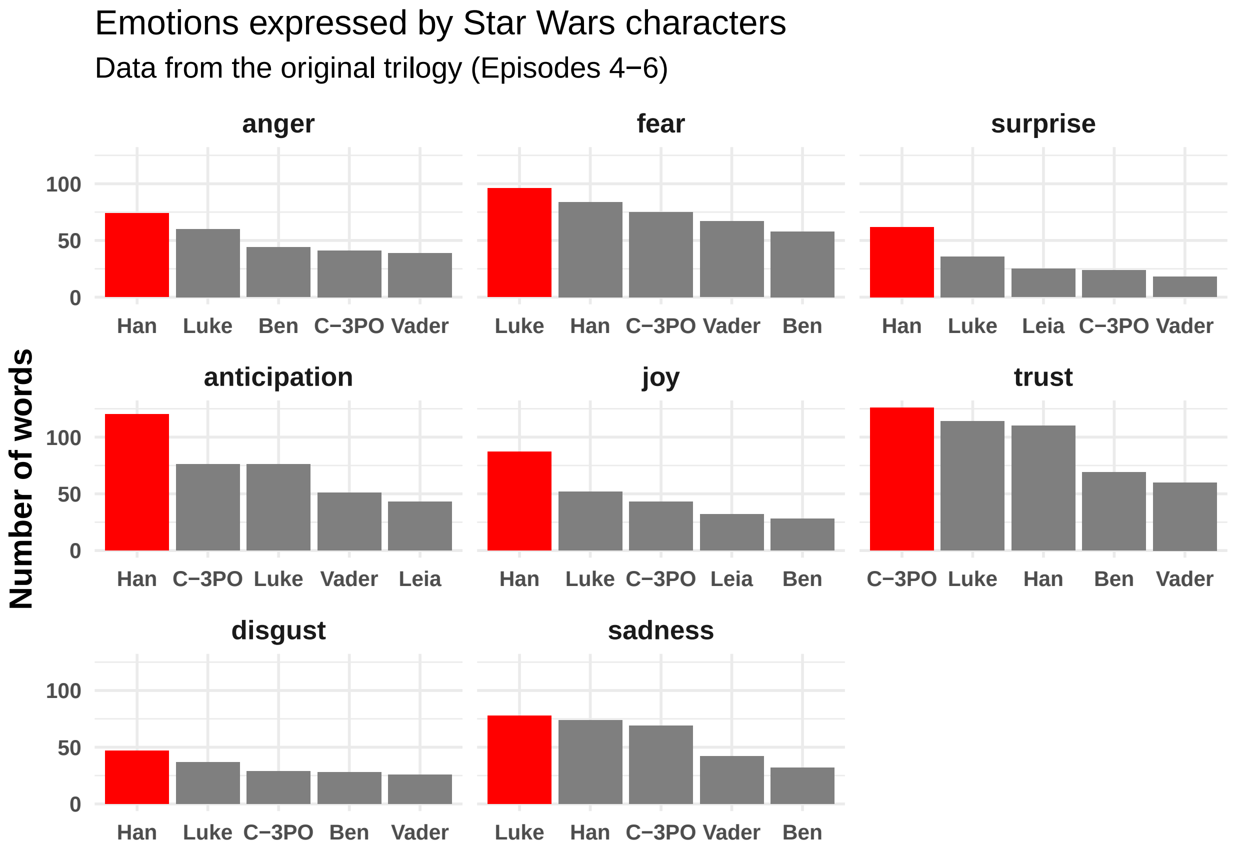
5.2 Create dataset with all NRC-categorised script words

```
sw_sentiments <- sw_tokens %>%
  inner_join(get_sentiments("nrc"), by = "word")
```

5.3 Sentiment words spoken by characters

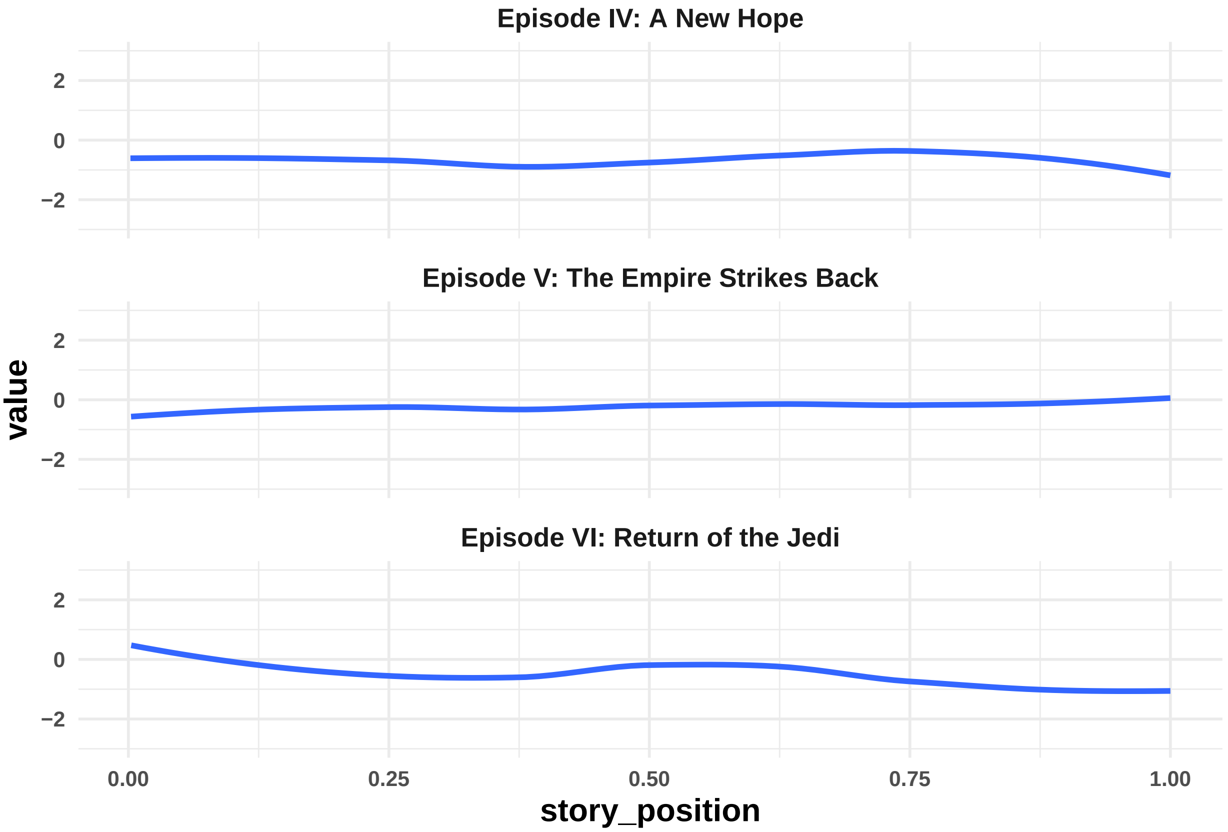
```
# NEED TO ACCOUNT FOR HOW MUCH CHARACTERS SPEAK
# TAKE TOP 20? SPEAKING CHARACTERS AND CALCULATE PERCENTAGES?


sw_sentiments %>%
  filter(!(sentiment %in% c("positive", "negative"))) %>%
  group_by(character, sentiment) %>%
  summarise(count = n()) %>%
  ungroup() %>%
  arrange(desc(count)) %>%
  group_by(sentiment) %>%
  slice_max(count, n = 5) %>%
  mutate(is_max_count = count == max(count),
         character = recode(character, Threepio = "C-3PO")) %>%
  ggplot(aes(
    x = reorder_within(character, -count, sentiment),
    y = count,
    fill = is_max_count
    )) +
  geom_col() +
  scale_fill_manual(
    values = c(`TRUE` = "red", `FALSE` = "grey50"), guide = F) +
  labs(
    x = NULL,
    y = "Number of words",
    title = "Emotions expressed by Star Wars characters",
    subtitle = "Data from the original trilogy (Episodes 4-6)"
  ) +
  facet_wrap(~sentiment, scales = "free_x", dir = "v") +
  scale_x_reordered()
```



6 Sentiment arcs

```
sw_tokens %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(episode) %>%
  mutate(
    word_n = row_number()
  ) %>%
  mutate(
    story_position = word_n/max(word_n) # all books on scale from 0 to 1
  ) %>%
  mutate(episode = as.character(episode),
      episode = recode(episode,
                       "4" = "Episode IV: A New Hope",
                       "5" = "Episode V: The Empire Strikes Back",
                       "6" = "Episode VI: Return of the Jedi")) %>%
ggplot() +
aes(x = story_position, y = value) +
geom_smooth(se = FALSE) +
facet_wrap(~episode, ncol = 1) +
coord_cartesian(ylim = c(-3, 3))
```


Episode IV: A New Hope
Episode V: The Empire Strikes Back
Episode VI: Return of the Jedi
value
story_position