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

Contents

1 Dependencies **1**

2 Reading in clean data **1**

3 Tokenize and remove stop words **1**

4 Check which sentiment lexicon categorizes most words **1**

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

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

sw_tokens
```

4 Check which sentiment lexicon categorizes most words

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

# vector of 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){

  # 3rd 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
```