Star Wars (4-6) Text Analysis of 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

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

sw_tokens
```

4 Check which sentiment lexicon categorizes most words

```r
lexicons <- c("bing", "afinn", "loughran", "nrc")

df <-lexicons %>%
    map(~left_join(sw_tokens, get_sentiments(.), by = "word"))

names(df) <- lexicons


for (lexicon in lexicons){
  missing <- sum(is.na(df[[lexicon]][[5]]))
  print(str_glue("The lexicon {lexicon} has {missing} uncategorised words"))
}
```

```
## The lexicon bing has 6625 uncategorised words
## The lexicon afinn has 6714 uncategorised words
## The lexicon loughran has 7340 uncategorised words
## The lexicon nrc has 5101 uncategorised words
```