Text Subsets and Word Categorizations

```
# load packages
library(tidyverse)
library(tidytext)
library(textdata)
library(wordcloud)
```

Below is an R code chunk that load the data sets `avatar-the-last-airbender.csv` and the `starwars.csv` files. The file `avatar-the-last-airbender.csv` contains transcrripts of all three seasons of the show "Avatar: Last Airbender". The file `starwars.csv` contains the transcripts of the 6 Star Wars movies. The code below also loads the stopwords lexicon and NRC (National Research Council Canada) sentiments lexicon.

```
# load the stopwords lexicon
swd_list <- tidytext::stop_words %>%
  filter(lexicon == "SMART") %>%
  select(word)
swd_list <- swd_list$word

# load NRC emotion lexicons
emotions_6 <- textdata::lexicon_nrc()

# load the data sets
atla <- read_csv("avatar-the-last-airbender.csv")
sw <- read_csv("starwars.csv")
```

1. Below is an R code that takes a subset of the `atla` data frame where it chooses two characters and tokenizes the words in the `character_words` variable while grouping them by `book` and `character` variables. Your task is to modify the code as listed below:

    a. Include two other characters, and remove the stopwords.
    b. Choose the positive and negative words in the `emotions_6` data frame and use it to subset the previous data frame. Hint: Use the `left_join()` function.
    c. Create a standardized stacked barplot that compares the proportion of positive and negative words for each character in each book. Write a paragraph describing your observations.

```
### [BEGIN] - MODIFY MAGRITTR PIPELINE HERE
atla_sub <- atla %>%
  # pick characters
  filter(character %in% c("Aang","Zuko")) %>%
  # group by book and character
  group_by(book,character) %>%
  # tokenize the character's words and put it into the word variable
  unnest_tokens(word,character_words) %>%
  # count the words
  count(word)
### [END] - MODIFY MAGRITTR PIPELINE HERE
```

```
### [BEGIN] - WRITE YOUR GGPLOT PIPELINE HERE

### [END] - WRITE YOUR GGPLOT PIPELINE HERE
```

2. The code shown below uses the `sw` data frame and subsets it using a word or regular expression for each level in the `episode` variable. It then tokenizes each sentence. Your task is to modify the code as listed below:

    a. Define a word or a regular expression that matches a different name to subset the data.
    b. Use the `emotions_6` data frame to subset the words. Use 6 sentiment levels. Choose the top 15 most frequent words.
    c. Create a wordcloud and color the words according to its corresponding sentiment.
    d. Repeat a-c but with a different name and a custom regular expression. Write a paragraph describing your observations. Note that the method of subsetting here is not by the dialogues of a character but by each sentence mentioning the character/object, which includes the character's words.

```
### [BEGIN] - DEFINE A WORD OR A REGULAR EXPRESSION HERE
string_pattern <- "luke"
### [END] - DEFINE A WORD OR A REGULAR EXPRESSION HERE

# subset the dataframe by tokenizing by sentence
#   and pick sentences with the defined string
sw_sub <- sw %>%
  # group by episode
  group_by(episode) %>%
  # tokenize by sentence
  unnest_sentences(sentence, text) %>%
  # choose rows with the string pattern
  filter(grepl(string_pattern, sentence, ignore.case = TRUE))

### [BEGIN] - MODIFY MAGRITTR PIPELINE HERE
# get words in each sentence
sw_sub_tokens <- sw_sub %>%
  # tokenize words
  unnest_tokens(word,sentence) %>%
  # filter out the stopwords
  filter(!word %in% swd_list)
### [END] - MODIFY MAGRITTR PIPELINE HERE
```

```
### [BEGIN] - WRITE YOUR GGPLOT PIPELINE HERE

### [END] - WRITE YOUR GGPLOT PIPELINE HERE
```