Counting Words

```r
# load packages
library(tidyverse)
library(tidytext)
```

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 transcripts 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.

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

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

1. Examine the two data sets and identify which is structured, unstructured, or semi-structured. List and identify the types of variables for each data set.

2. Choose either of the datasets. Use the `tidytext` package to tokenize either data sets, count the words in the text variable for each season/book/movie, filter out the stopwords, and visualize the top 15 words using a barplot. Below, is an example code for you to get started.

```r
### [BEGIN] - CHANGE DATA SET HERE
df <- atla
### [END] - CHANGE DATA SET HERE

### [BEGIN] - WORK ON YOUR DATA WRANGLING HERE
# tokenize text and count words
df_counts <- df %>%
  # group by season/book/movie
  group_by(book) %>%
  # automatically tokenizes the text and puts them into the word variable
  unnest_tokens(word, full_text) %>%
  # use the count function to get the frequency of each word
  count(word) %>%
  # filter out the stopwords using the stopwords list defined earlier
  filter(!word %in% swd_list) %>%
  # count the frequency of each word in each group
  arrange(desc(n))
### [END] - WORK ON YOUR DATA WRANGLING HERE
```

```r
### [BEGIN] - WORK ON YOUR GGPLOT PIPELINE HERE

### [END] - WORK ON YOUR GGPLOT PIPELINE HERE
```