Assignment 11

Allison Tessman

2024-03-29

Question 1. Working with a text dataset containing quotes from the TV Show Friends. Do the following:

1. Plot the word frequency of the text data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidytext)

#Create list of tokens
df = read_csv('https://bryantstats.github.io/math475/assignments/friends_quotes.csv')
```

```
## Rows: 60291 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): author, episode_title, quote
## dbl (3): episode_number, quote_order, season
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df = df %>%
  select(quote) %>%
  rename(text = quote)

stop_word2 = tibble(word = c(letters, LETTERS, "oh", 'just'))

# list of tokens/words
```

```
df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  anti_join(stop_word2)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```
## # A tibble: 341,291 x 1
##    word
##    <chr>
##  1 nothing
##  2 tell
##  3 guy
##  4 work
##  5 c'mon
##  6 going
##  7 guy
##  8 gotta
##  9 something
## 10 wrong
## # i 341,281 more rows
```

```
# Count token frequency
df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  anti_join(stop_word2)%>%
  count(word, sort = TRUE)
```

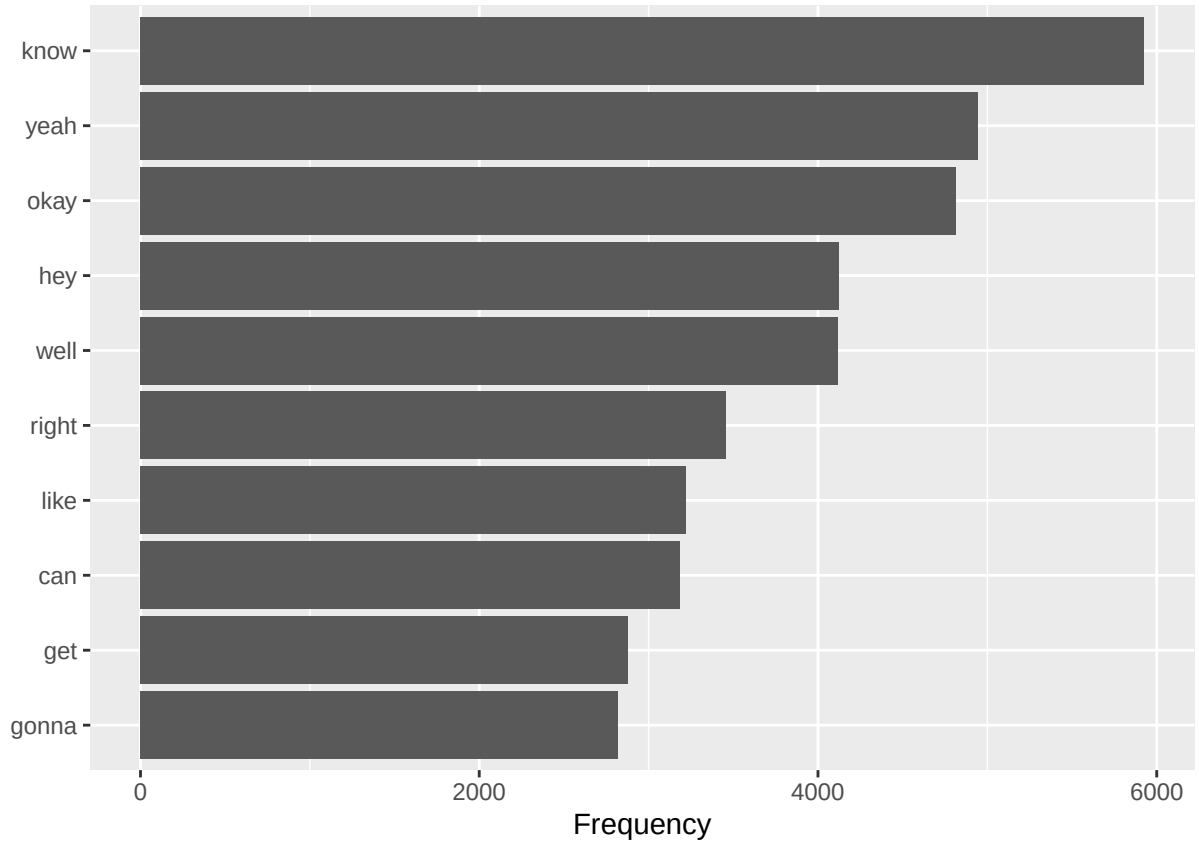
```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```
## # A tibble: 16,877 x 2
##    word      n
##    <chr> <int>
##  1 know   5926
##  2 yeah   4946
##  3 okay   4817
##  4 hey    4121
##  5 well   4119
##  6 right  3459
##  7 like   3220
##  8 can    3187
##  9 get    2877
## 10 gonna  2818
## # i 16,867 more rows
```

```
# Plot token frequency
df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
```

```
  anti_join(stop_word2)%>%
  count(word, sort = TRUE)%>%
  head(10) %>%
  ggplot(aes(x = n, y = reorder(word, n))) +
  geom_col() +
  labs(y = '', x = 'Frequency')
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```



2. Plot a word cloud of the text data

```
# Plot word cloud
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  anti_join(stop_word2)%>%
  count(word, sort = TRUE)%>%
  with(wordcloud(word, n, random.order = FALSE,
                 max.words = 50, colors=brewer.pal(8,"Dark2")))
```

```
## Joining with 'by = join_by(word)'
```

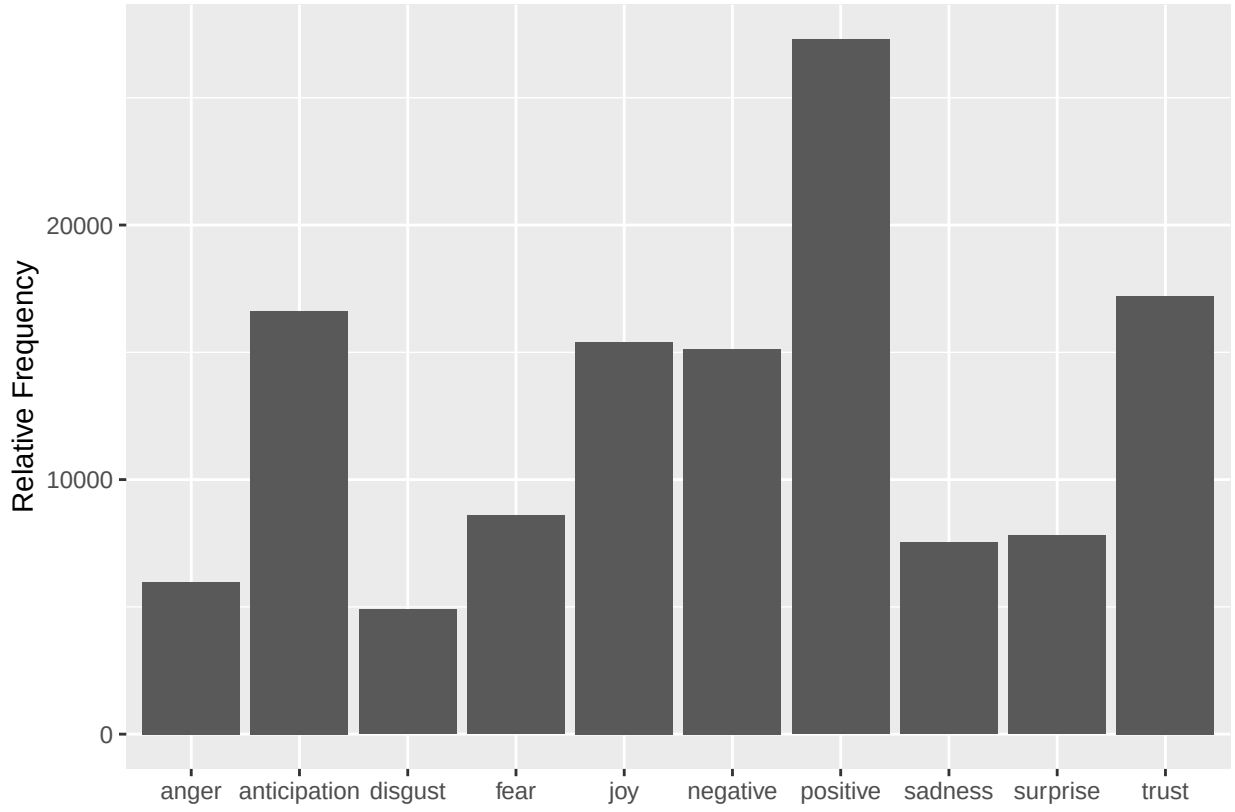
```
## Joining with 'by = join_by(word)'
```

3. Run a sentiment analysis on the data

```
# Sentiment Analysis Using nrc Lexicon

df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  anti_join(stop_word2) %>%
  inner_join(get_sentiments("nrc")) %>%
  filter(!is.na(sentiment)) %>%
  count(sentiment, sort = TRUE) %>%
  ggplot(aes(sentiment, n))+geom_col()+
  labs(y='Relative Frequency', x ='')
```

```
## Joining with 'by = join_by(word)'
## Joining with 'by = join_by(word)'
## Joining with 'by = join_by(word)'
```



Question 2. Do Question 1 on your own text dataset.

```
#CNN News Text

# Create list of tokens
library(tidyverse)
library(tidytext)
df = read_csv('~/Applied Analystics SAS Prog/mymath475/CNNtext.csv')
```

```
## Rows: 11490 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): id, article, highlights
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df = df %>%
  select(highlights) %>%
  rename(text = highlights)

# list of tokens/words
df %>%
```

```
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords())
```

```
## Joining with `by = join_by(word)`
```

```
## # A tibble: 381,752 x 1
##    word
##    <chr>
##  1 experts
##  2 question
##  3 packed
##  4 planes
##  5 putting
##  6 passengers
##  7 risk
##  8 u.s
##  9 consumer
## 10 advisory
## # i 381,742 more rows
```

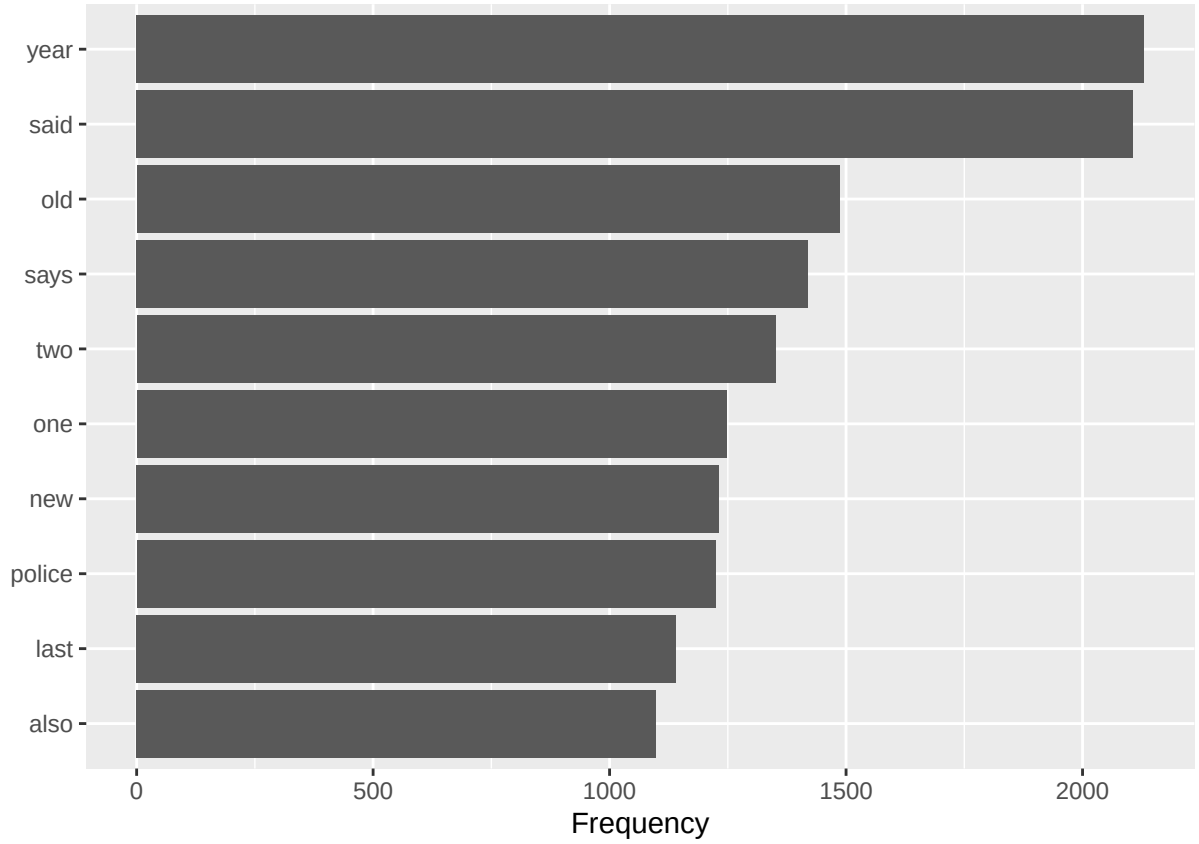
```
# Count word frequency
df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  count(word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
```

```
## # A tibble: 37,424 x 2
##    word       n
##    <chr>  <int>
##  1 year    2130
##  2 said    2106
##  3 old     1486
##  4 says    1419
##  5 two     1351
##  6 one     1247
##  7 new     1231
##  8 police  1224
##  9 last    1141
## 10 also    1098
## # i 37,414 more rows
```

```
# Plot word frequency
df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  count(word, sort = TRUE)%>%
  head(10) %>%
  ggplot(aes(x = n, y = reorder(word, n))) +
  geom_col() +
  labs(y = '', x = 'Frequency')
```

```
## Joining with `by = join_by(word)`
```



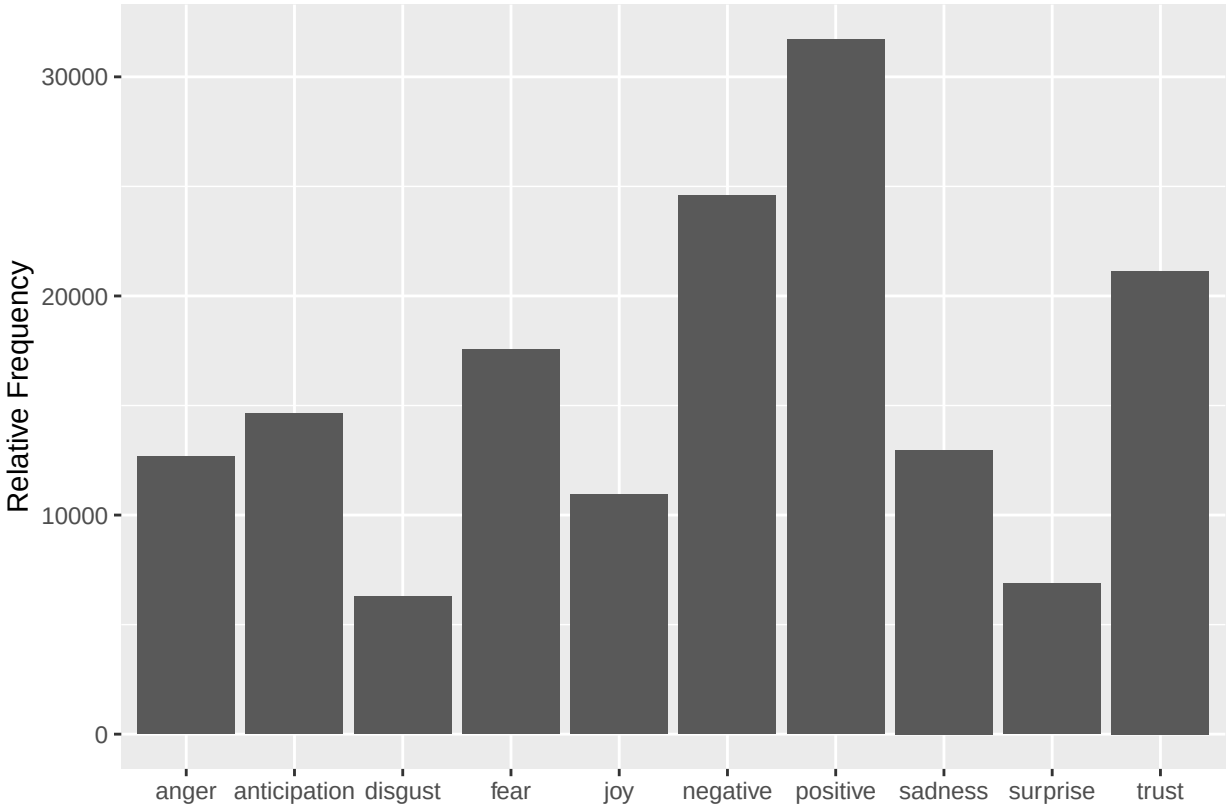

```
# plot word cloud
library(wordcloud)

df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  count(word, sort = TRUE)%>%
  with(wordcloud(word, n, random.order = FALSE,
                 max.words = 50, colors=brewer.pal(8,"Dark2")))
```

```
## Joining with `by = join_by(word)`
```


```
# Sentiment Analysis Using nrc Lexicon
df %>%
  unnest_tokens(input = text, output = word) %>%
  anti_join(get_stopwords()) %>%
  inner_join(get_sentiments("nrc")) %>%
  filter(!is.na(sentiment)) %>%
  count(sentiment, sort = TRUE) %>%
  ggplot(aes(sentiment, n))+geom_col()+
  labs(y='Relative Frequency', x ='')
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```


Relative Frequency
30000
20000
10000
0
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust