Word Frequency

Kailin Xu

2025-03-19

```
csv1 <- read.csv("cleaned_vaxxhappened1.csv")
csv2 <- read.csv("cleaned_conspiracy1.csv")
csv3 <- read.csv("cleaned_HermanCainAward1.csv")
csv4 <- read.csv("cleaned_politics1.csv")

merged_data <- rbind(csv1, csv2, csv3, csv4)

write.csv(merged_data, file = "merged_data.csv", row.names = FALSE)
```

Word Frequency

```
library(tidytext)
```

```
## Warning:  'tidytext' R 4.4.3
```

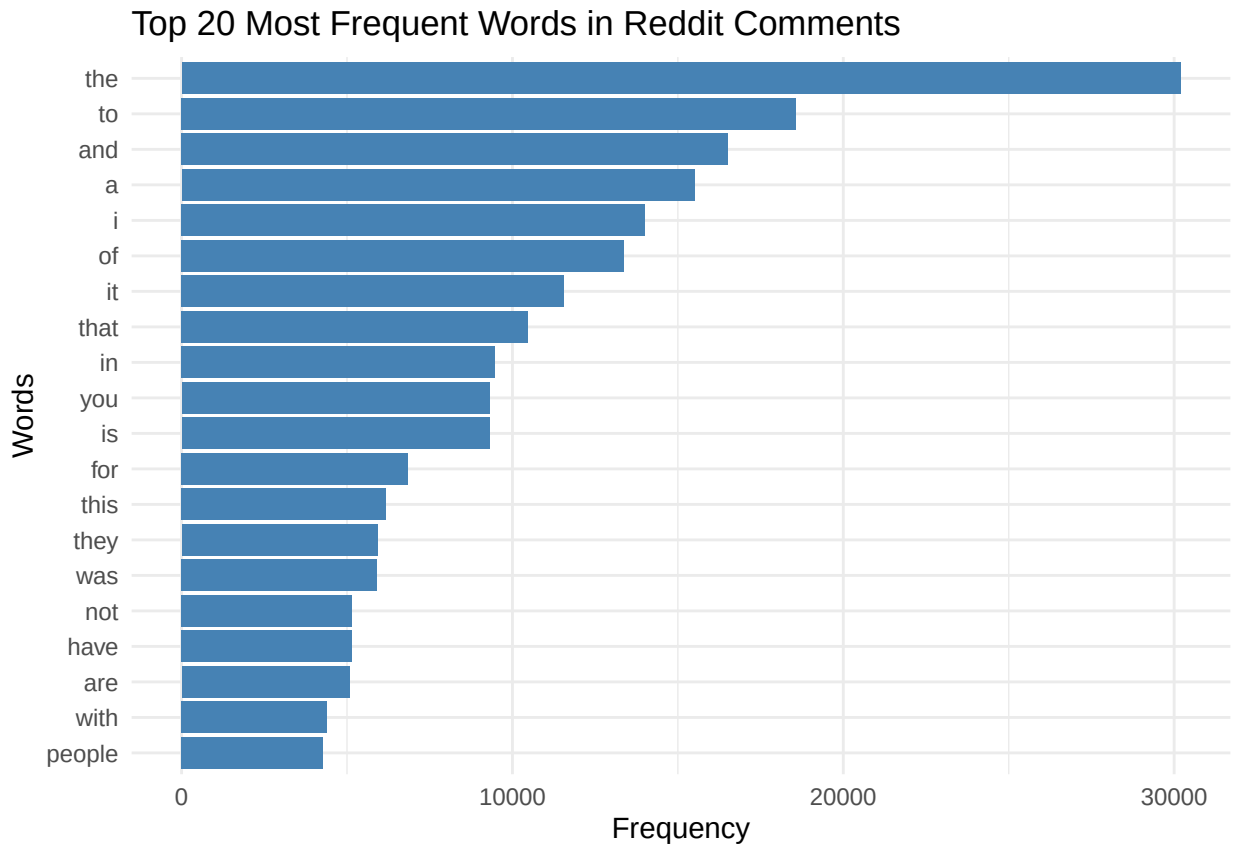
```
library(dplyr)
library(ggplot2)
library(stringr)
library(tidyr)
```

```
merged_data$comment <- as.character(merged_data$comment)

word_counts_raw <- merged_data %>%
  unnest_tokens(word, comment) %>%
  count(word, sort = TRUE)

head(word_counts_raw, 20)
```

```
##      word     n
## 1     the 30190
## 2      to 18560
## 3     and 16519
## 4       a 15504
## 5       i 14003
## 6      of 13383
## 7      it 11559
## 8    that 10461
## 9      in  9473
```

```
## 10    you  9307
## 11     is  9304
## 12    for  6841
## 13   this  6175
## 14   they  5923
## 15    was  5896
## 16    not  5157
## 17   have  5156
## 18    are  5077
## 19   with  4394
## 20 people  4270
```

```
top_words <- word_counts_raw %>%
  slice_max(n, n = 20)  # top 20 words

ggplot(top_words, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "steelblue") +
  coord_flip() +  # rotate graph
  labs(title = "Top 20 Most Frequent Words in Reddit Comments",
       x = "Words", y = "Frequency") +
  theme_minimal()
```



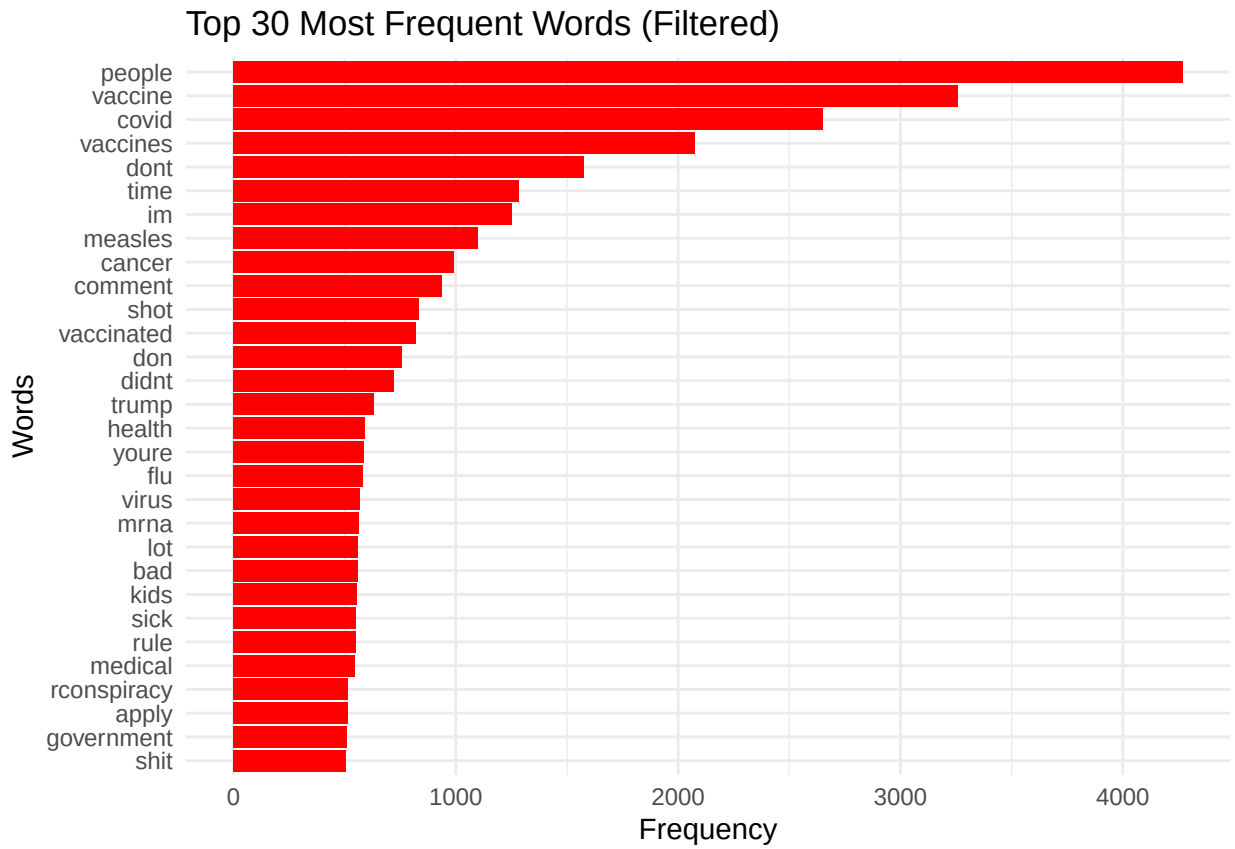
stop words

```
# remove stop words
data("stop_words")
```

```r
word_counts <- merged_data %>%
  unnest_tokens(word, comment) %>%
  anti_join(stop_words, by = "word") %>%
  count(word, sort = TRUE)

head(word_counts, 20)
```

```
##          word    n
## 1      people 4270
## 2     vaccine 3258
## 3       covid 2653
## 4    vaccines 2074
## 5        dont 1578
## 6        time 1282
## 7          im 1251
## 8     measles 1100
## 9      cancer  993
## 10    comment  938
## 11       shot  833
## 12 vaccinated  822
## 13        don  757
## 14      didnt  724
## 15      trump  632
## 16     health  590
## 17      youre  588
## 18        flu  583
## 19      virus  568
## 20       mrna  563
```

```r
top_words_stop <- word_counts %>%
  slice_max(n, n = 30)

ggplot(top_words_stop, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "red") +
  coord_flip() +
  labs(title = "Top 30 Most Frequent Words (Filtered)",
       x = "Words", y = "Frequency") +
  theme_minimal()
```



add more stop words

```
custom_stop_words <- c( "don","im","comment","didnt","youre","time","people","dont")

top_words_custom <- word_counts %>%
  filter(!word %in% custom_stop_words) %>%
  slice_max(n, n = 20)

top_words_custom
```

```
##           word    n
## 1      vaccine 3258
## 2        covid 2653
## 3     vaccines 2074
## 4      measles 1100
## 5       cancer  993
## 6         shot  833
## 7   vaccinated  822
## 8        trump  632
## 9       health  590
## 10         flu  583
## 11       virus  568
## 12        mrna  563
## 13         lot  559
## 14         bad  558
## 15        kids  555
```

```
## 16        sick  551
## 17        rule  550
## 18     medical  548
## 19       apply  517
## 20 rconspiracy  517
```

```
ggplot(top_words_custom, aes(x = reorder(word, n), y = n)) +
  geom_col(fill = "orange") +
  coord_flip() +
  labs(title = "Top 20 Most Frequent Words (cleaned)",
       x = "Words", y = "Frequency") +
  theme_minimal()
```