Analysis on COVID-19 Reddit Threads

Samar Dikshit

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(ggplot2)
library(ggraph)
library(igraph)
library(readr)
library(tidyr)
library(tidytext)
```

Read the dataset.

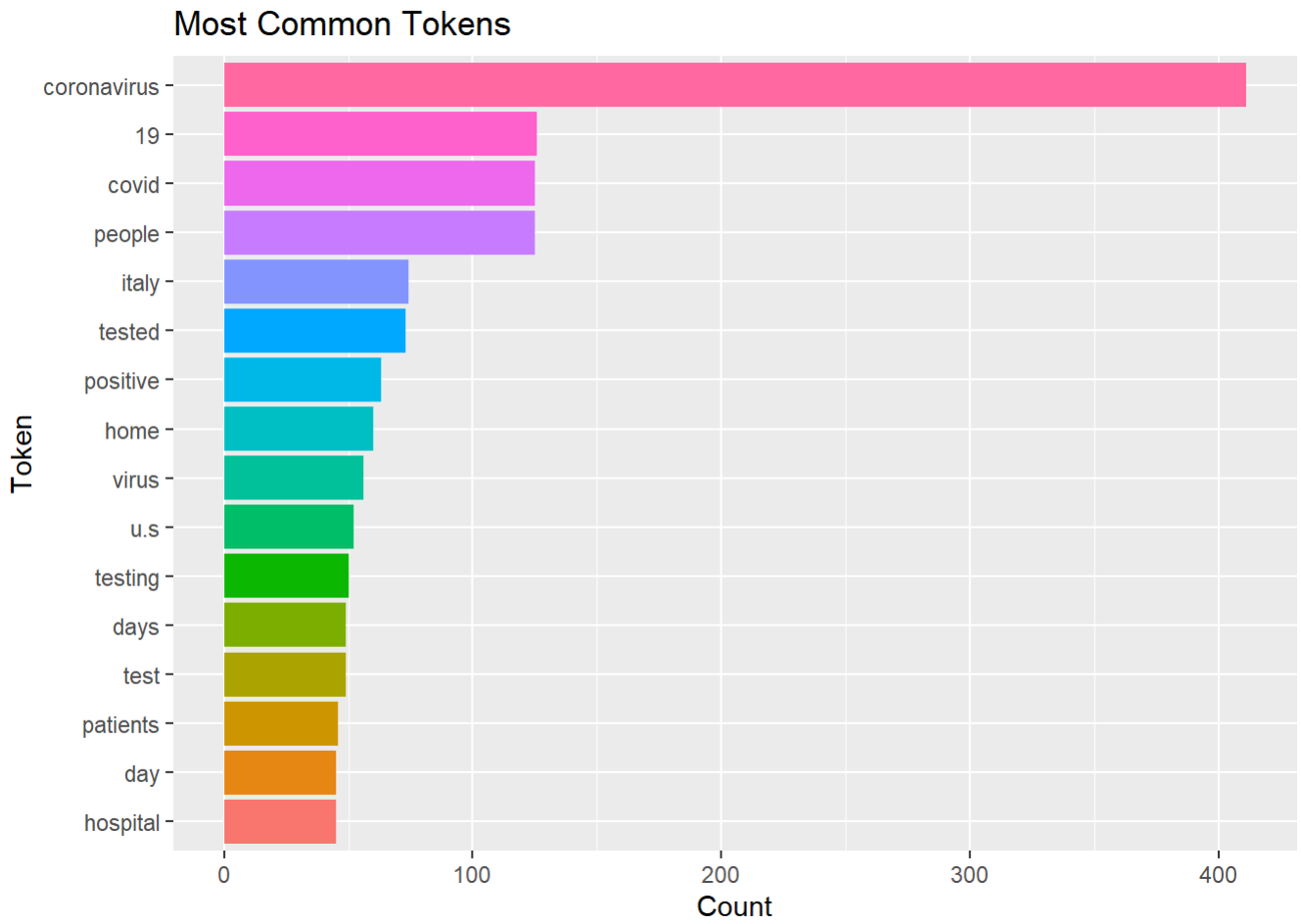
```
filePath <- paste(getwd(), "data.csv", sep = "/")
df <- read_csv(filePath)

df <- df %>% rename(id = X1)
```

After tokenising the thread titles and removing stop word, we obtain the tokens. We can also generate bigrams where each word in the bigram is not a stop word.

```
tokenisedData <- df %>% unnest_tokens(word, title) %>% anti_join(stop_words, by = "word")
bigramData <- df %>% unnest_tokens(bigram, title, token = "ngrams", n = 2) %>% separate(bigram,
 c("w1", "w2"), sep = " ") %>% filter(! w1 %in% stop_words$word, ! w2 %in% stop_words$word) %>%
 unite(bigram, w1, w2, sep = " ")
```

Finding the 15 most common tokens:

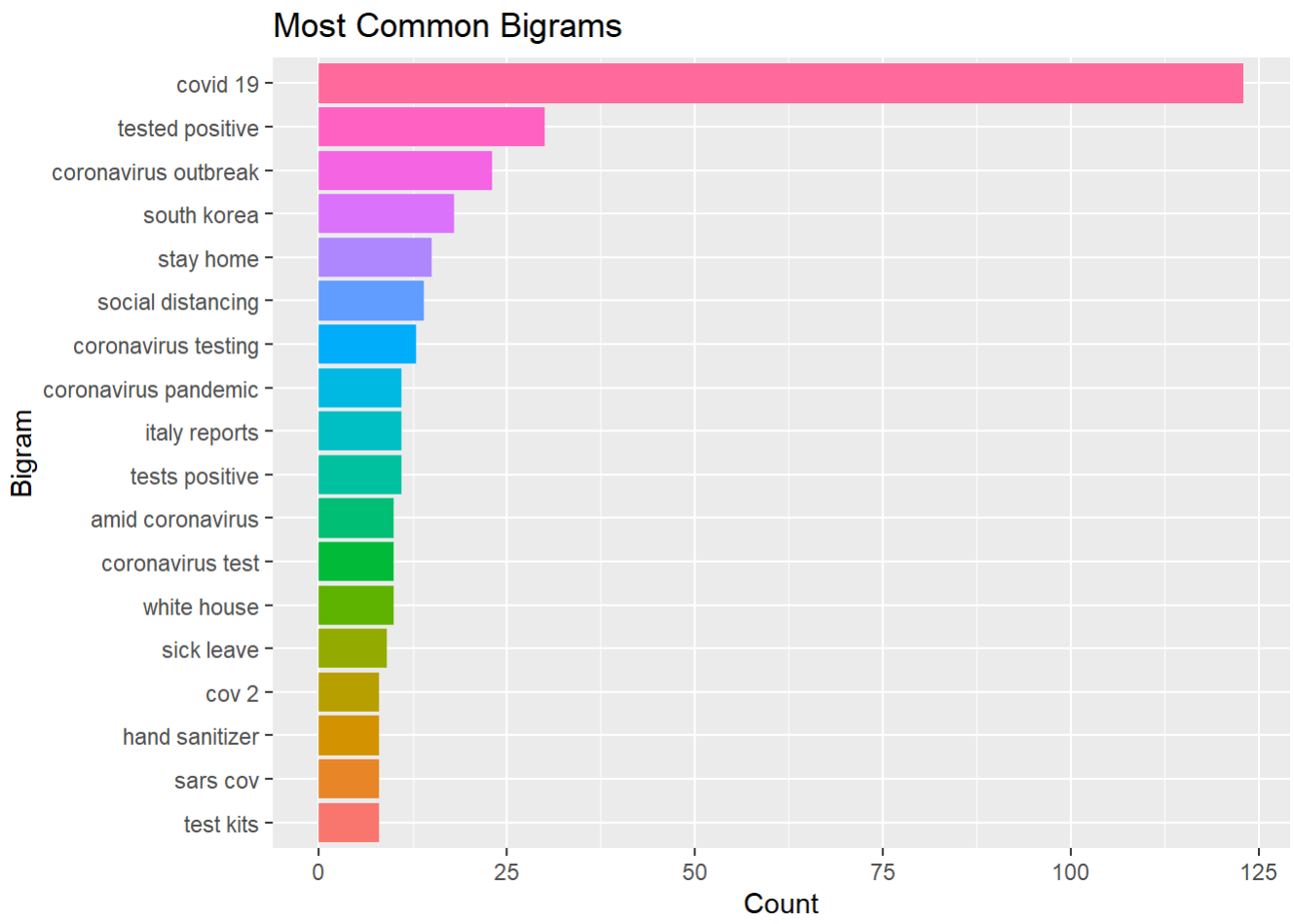
```
count(tokenisedData, word, sort = TRUE) %>% top_n(15, wt = n) %>% mutate(word = factor(word, lev
els = rev(unique(word)))) %>% ggplot(aes(x = word, y = n, fill = word)) + geom_col(show.legend =
FALSE) + coord_flip() + labs(x = "Token", y = "Count", title = "Most Common Tokens")  # we actua
lly get 16 tokens as 'day' and 'hospital' have the same frequency
```



It's not a surprise that *coronavirus* is the most common token. Other expected common terms are *covid*, *positive*, *italy*, and *testing*.

Finding the 15 most common bigrams:

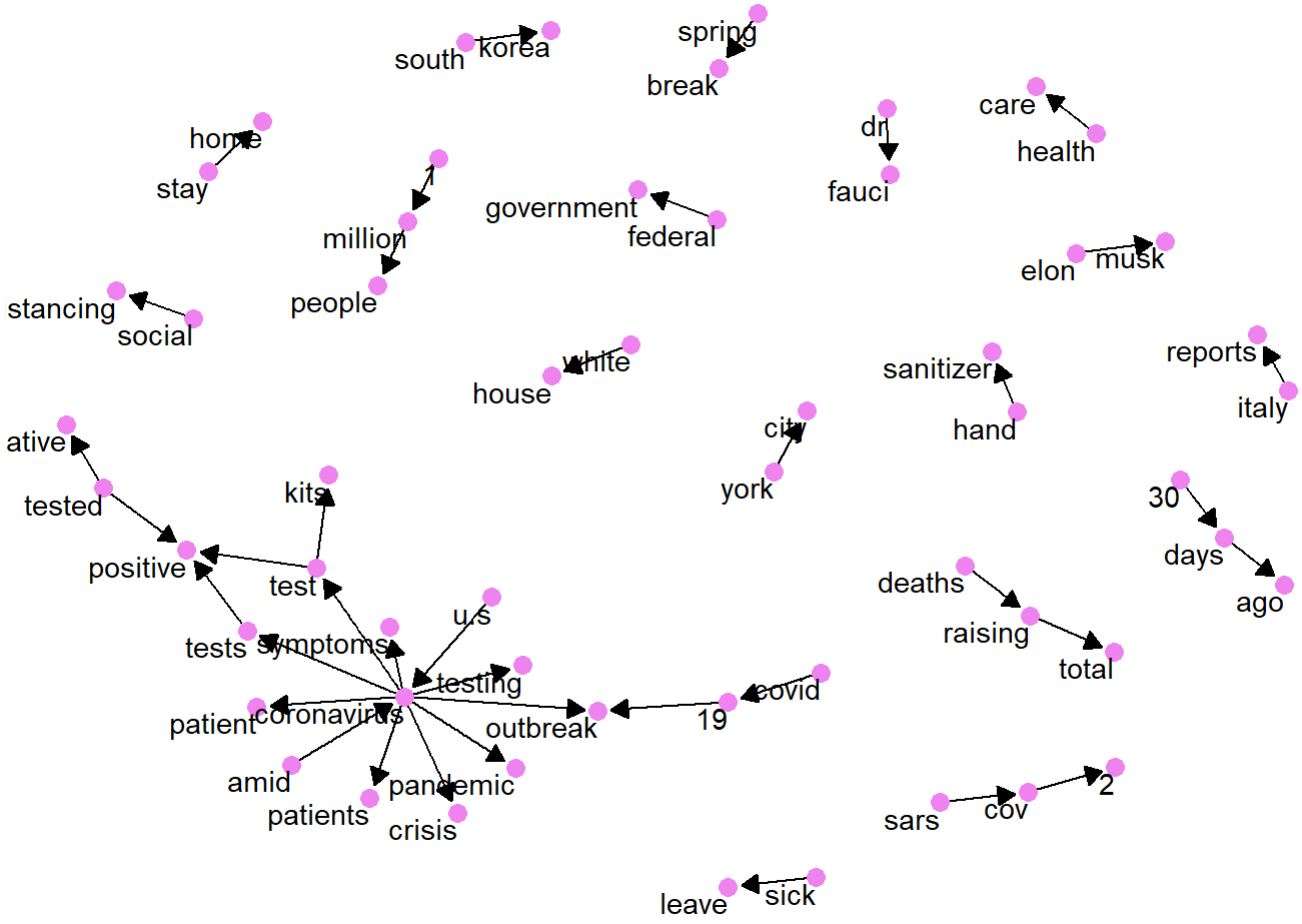
```
count(bigramData, bigram, sort = TRUE) %>% top_n(15, wt = n) %>% mutate(bigram = factor(bigram,
 levels = rev(unique(bigram)))) %>% ggplot(aes(x = bigram, y = n, fill = bigram)) + geom_col(sho
w.legend = FALSE) + coord_flip() + labs(x = "Bigram", y = "Count", title = "Most Common Bigrams"
)  # we actually get 18 tokens as the last 4 bigrams have the same frequency
```



Just as with tokens, there's nothing surprsing amongst the most common tokens. We can see that bigrams such as *social distancing* and *stay home* are frequently used as they have become the main way to prevent the spread of the virus, leading to more and more people using these terms.

Bigram relationship graph:

```
graph <- bigramData %>% separate(bigram, c("w1", "w2"), sep = " ") %>% count(w1, w2, sort = TRUE
) %>% filter(n >= 5) %>% graph_from_data_frame()
arr <- grid::arrow(type = "closed", length = unit(0.1, "inches"))
ggraph(graph, layout = "fr") + geom_edge_link(show.legend = FALSE, arrow = arr, end_cap = circle
(.07, 'inches')) + geom_node_point(color = "#ee82ee", size = 3) + geom_node_text(aes(label = nam
e), vjust = 1.2, hjust = 1) + theme_void()
```


south
korea
spring
break
home
stay
million
people
government
federal
dr
fauci
care
health
elon
musk
stancing
social
house
white
sanitizer
hand
reports
italy
ative
tested
positive
kits
test
city
york
30
days
ago
deaths
raising
total
tests
symptoms
u.s
testing
patient
coronavirus
outbreak
covid
19
amid
patients
pandemic
crisis
sars
cov
2
leave
sick