PM 566 Lab 6

AUTHOR
Ziquan 'Harrison' Liu

Required Package

```
library(tidytext)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library(tidyr) # for Q6
```

```
mt_samples <- read_csv("https://raw.githubusercontent.com/USCbiostats/data-science-data/m
```

```
New names:
Rows: 4999 Columns: 6
── Column specification
──────────────────────────────────────────────────────── Delimiter: "," chr
(5): description, medical_specialty, sample_name, transcription, keywords dbl
(1): ...1
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.
• `` -> `...1`
```

```
mt_samples <- mt_samples |>
  select(description, medical_specialty, transcription)
head(mt_samples)
```

```
# A tibble: 6 × 3
  description                                    medical_specialty transcription
  <chr>                                          <chr>             <chr>
1 A 23-year-old white female presents with comp… Allergy / Immuno… "SUBJECTIVE:…
2 Consult for laparoscopic gastric bypass.       Bariatrics        "PAST MEDICA…
```

```
3 Consult for laparoscopic gastric bypass.       Bariatrics        "HISTORY OF …
4 2-D M-Mode. Doppler.                           Cardiovascular /… "2-D M-MODE:…
5 2-D Echocardiogram                             Cardiovascular /… "1.  The lef…
6 Morbid obesity.  Laparoscopic antecolic anteg… Bariatrics        "PREOPERATIV…
```

Question 1: What specialties do we have?

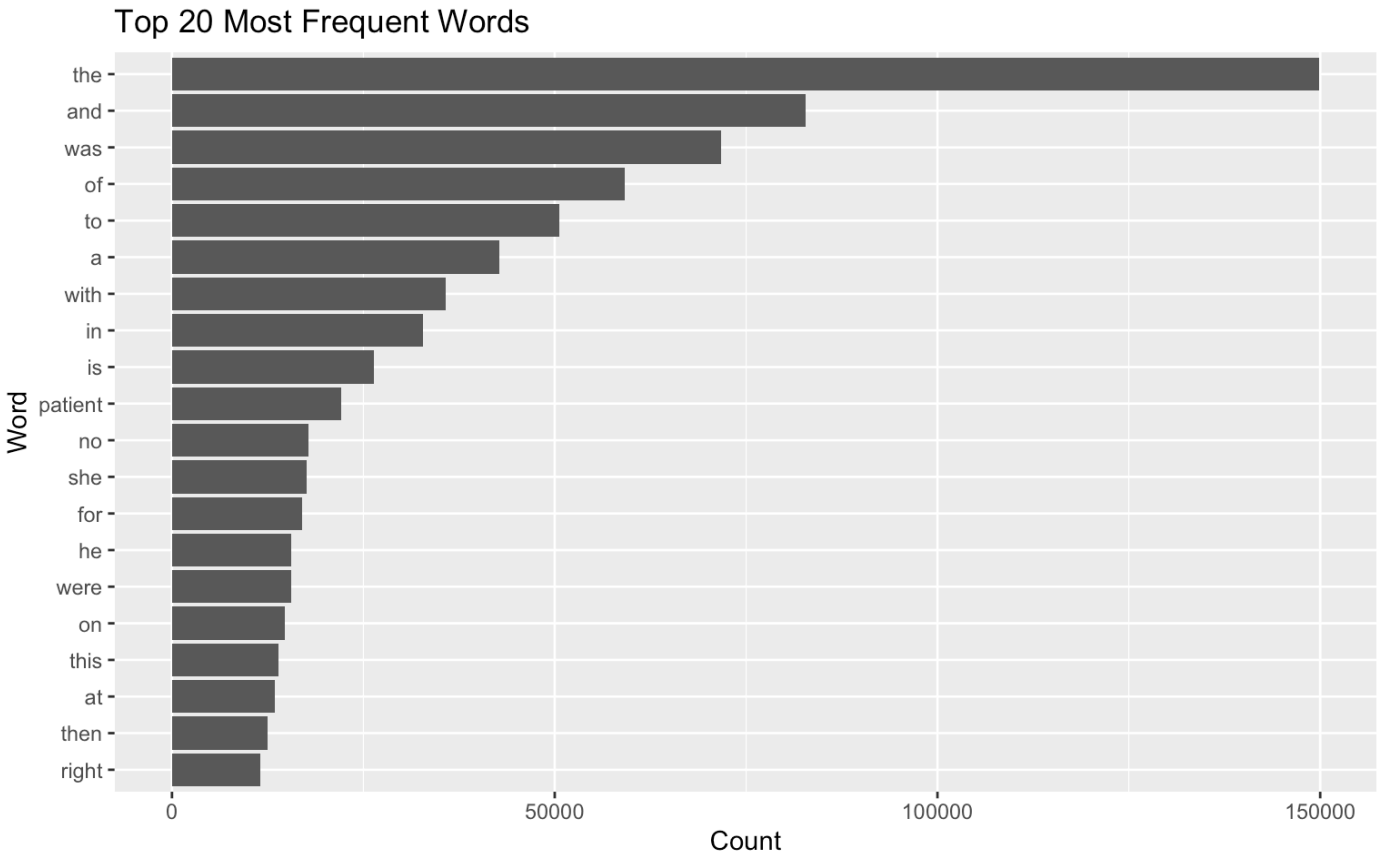
```
mt_samples %>%
  count(medical_specialty, sort = TRUE)
```

```
# A tibble: 40 × 2
   medical_specialty                n
   <chr>                        <int>
 1 Surgery                       1103
 2 Consult - History and Phy.     516
 3 Cardiovascular / Pulmonary     372
 4 Orthopedic                     355
 5 Radiology                      273
 6 General Medicine               259
 7 Gastroenterology               230
 8 Neurology                      223
 9 SOAP / Chart / Progress Notes  166
10 Obstetrics / Gynecology        160
# ℹ 30 more rows
```

Answers: Based on the resulte above, there are some categories related though overall not so much. all unique specialty, so there is no obvious overlap.

Question 2

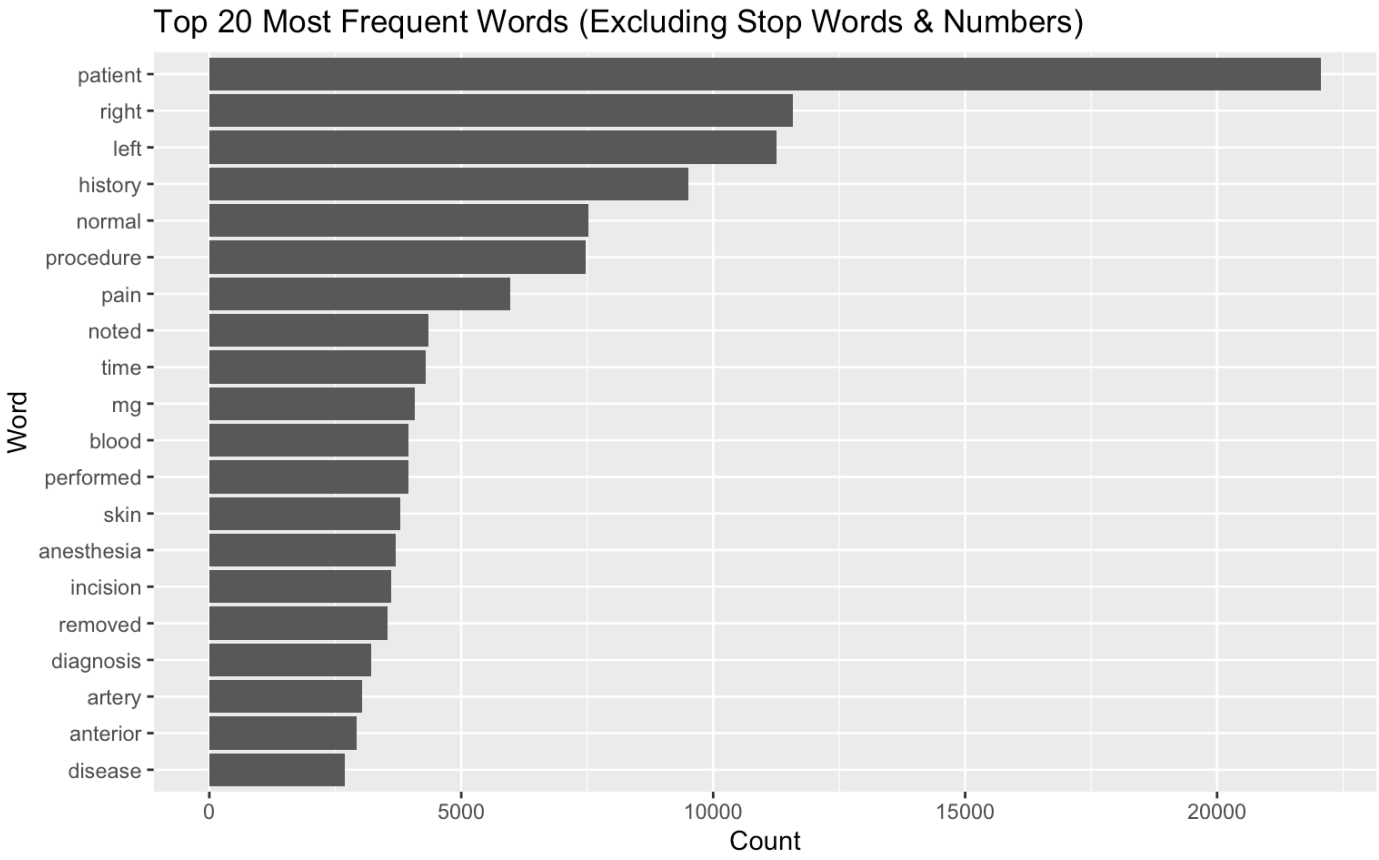
```
# Tokenize the the words in the transcription column
# Count the number of times each token appears
# Visualize the top 20 most frequent words
mt_samples %>%
  unnest_tokens(word, transcription)  %>%
  count(word, sort = TRUE)  %>%
  top_n(20,n)  %>%
  ggplot(aes(x = reorder(word,n), y = n)) +
  geom_col() +
  coord_flip() +
  labs(x="Word", y="Count", title = "Top 20 Most Frequent Words")
```



Answers: It does mot make sense since top words were stop words like “the” , “and”, and “was”, etc are not insight words, just people will generally use such words in communication.

Question 3

```
mt_samples %>%
  unnest_tokens(word, transcription)  %>%
  anti_join(stop_words %>% filter(!word %in% c("right"))
            , by = "word") %>%
  filter(!grepl("^[0-9]+$", word)) %>%
  count(word, sort = TRUE)  %>%
  top_n(20,n)  %>%
  ggplot(aes(x = reorder(word,n), y = n)) +
  geom_col() +
  coord_flip() +
  labs(x="Word", y="Count", title = "Top 20 Most Frequent Words (Excluding Stop Words & N
```



Answers: After stop words were removed, the interesting insight was showed from the output. The patient rank top1 frequent ussing in medical.That is helpful of us to understand what this dataset talked about. (the original code part: However, the word “right” can be use as both stop words and words meaning as direction. Here, the coding let “right” being removed, which may increase concern like “right” was utilized as medical term such as “right ventricle”.)

Question 4

```
# Bi-grams
mt_samples %>%
  unnest_tokens(bigram, transcription, token = "ngrams", n=2)  %>%
  count(bigram, sort = TRUE)  %>%
  top_n(20,n)
```

```
# A tibble: 20 × 2
   bigram          n
   <chr>       <int>
 1 the patient 20307
```

```
 2 of the      19062
 3 in the      12790
 4 to the      12374
 5 was then     6956
 6 and the      6350
 7 patient was  6293
 8 the right    5509
 9 on the       5241
10 the left     4860
11 with a       4857
12 history of   4537
13 to be        4345
14 is a         4014
15 with the     4002
16 there is     3950
17 at the       3657
18 there was    3334
19 patient is   3332
20 was placed   3328
```

```
# Tri-grams
mt_samples %>%
  unnest_tokens(bigram, transcription, token = "ngrams", n=3)  %>%
  count(bigram, sort = TRUE)  %>%
  top_n(20,n)
```

```
# A tibble: 22 × 2
   bigram                n
   <chr>             <int>
 1 the patient was    6104
 2 the patient is     3075
 3 as well as         2243
 4 there is no        1678
 5 the operating room 1532
 6 patient is a       1491
 7 prepped and draped 1490
 8 was used to        1480
 9 and draped in      1372
10 at this time       1333
# ℹ 12 more rows
```

Answers:In the Bi-grams, the most common phrase are “the patient” and “of the” which are not really insightful. However, for the Tri-grams more information were provided, such as “the operating romm”, which let us suspect that one of the main point of this dataset was relevant to the surgery.

Question 5

```
mt_samples %>%
  unnest_tokens(bigram, transcription, token = "ngrams", n=2)  %>%
  separate(bigram,c("word1","word2"), sep="") %>%
  filter(word1 =="patients" | word2 == "operating") %>%
  count(word1,word2, sort = TRUE)
```

```
Warning: Expected 2 pieces. Additional pieces discarded in 2398597 rows [1, 2, 3, 4, 5,
6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

# A tibble: 0 × 3
# ℹ 3 variables: word1 <chr>, word2 <chr>, n <int>
```

Question 6

```
mt_samples %>%
  unnest_tokens(word, transcription)  %>%
  anti_join(stop_words %>% filter(!word %in% c("right"))
            , by = "word") %>%
  filter(!grepl("^[0-9]+$", word)) %>%
  group_by(medical_specialty) %>%
  count(word, sort = TRUE)  %>%
  top_n(5,n)  %>%
  arrange(medical_specialty, desc(n))
```

```
# A tibble: 208 × 3
# Groups:   medical_specialty [40]
   medical_specialty    word          n
   <chr>                <chr>     <int>
 1 Allergy / Immunology history      38
 2 Allergy / Immunology noted        23
 3 Allergy / Immunology patient      22
 4 Allergy / Immunology allergies    21
 5 Allergy / Immunology nasal        13
 6 Allergy / Immunology past         13
 7 Autopsy              right       108
 8 Autopsy              left         83
 9 Autopsy              inch         59
10 Autopsy              neck         55
# ℹ 198 more rows
```

Answers: The result above,the table showed the 5 most-used words for each specialty.

Question 7

```
# Seventh-grams
mt_samples %>%
  unnest_tokens(bigram, transcription, token = "ngrams", n=7)  %>%
  count(bigram, sort = TRUE)  %>%
  top_n(20,n)
```

```
# A tibble: 20 × 2
   bigram                                            n
   <chr>                                         <int>
 1 history of present illness the patient is       423
 2 patient was taken to the operating room         418
 3 the patient was taken to the operating          417
 4 prepped and draped in the usual sterile         388
 5 was prepped and draped in the usual             386
 6 of present illness the patient is a             349
 7 the patient tolerated the procedure well and    342
 8 the patient was brought to the operating        293
 9 patient was brought to the operating room       285
10 and draped in the usual sterile fashion         280
11 procedure the patient was taken to the          216
12 history of present illness this is a            208
13 the patient was prepped and draped in           199
14 to the recovery room in stable condition        199
15 patient tolerated the procedure well and was    197
16 was taken to the operating room and             180
17 procedure the patient was brought to the        175
18 prepped and draped in the usual fashion         169
19 was brought to the operating room and           152
20 patient was prepped and draped in the           135
```

Answers: Based on th result above, when we increase bigrams into 7, the infomration from the dataset became more clear since we can find the rubric of surgery were confirmed based on the description multimple times

---

```
# Victor's example code
top_specialties <- mt_samples %>%
  count(medical_specialty, sort = TRUE)  %>%
  top_n(4,n)  %>%
  pull(medical_specialty)

mt_samples %>%
  filter(medical_specialty %in% top_specialties) %>%
  unnest_tokens(word, transcription) %>%
  anti_join(stop_words %>% filter(!word %in% c("right"))
```

```
          , by = "word") %>%
  filter(!grepl("^[0-9]+$", word)) %>%
  group_by(medical_specialty, word) %>%
  summarise(n=n(),.groups = "drop") %>%
  group_by(medical_specialty) %>%
  top_n(10,n) %>%
  ungroup() %>%
  mutate(word = reorder_within(word, n, medical_specialty)) %>%
  ggplot(aes(x = n, y = word, fill = medical_specialty)) +
  geom_col() +
  scale_y_reordered()+
  facet_wrap(~medical_specialty, scales = "free_y")+
  labs(x="Count", y="Word", title = "Top 10 Words by Specialty") +
  theme(legend.position = "none")
```