lab 6

Kendall Dimson

Set-up

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

```
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 x 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?

We can use count() from dplyr to figure out how many different catagories do we have? Are these catagories related? overlapping? evenly distributed?

There are 40 categories of medical specialties. The top five include surgery, cardiovascular/pulmonary, orthopedic, radiology, and general medicine. There is an uneven distribution. There are also overlapping categories: "Consult - History and Phy.," "SOAP / Chart / Progress Notes," Discharge Summary," "Pain Management," "Office Notes," "Letters" outlining administrative data.

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

```
# A tibble: 40 x 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
```

Question 2

Tokenize the the words in the transcription column

Count the number of times each token appears

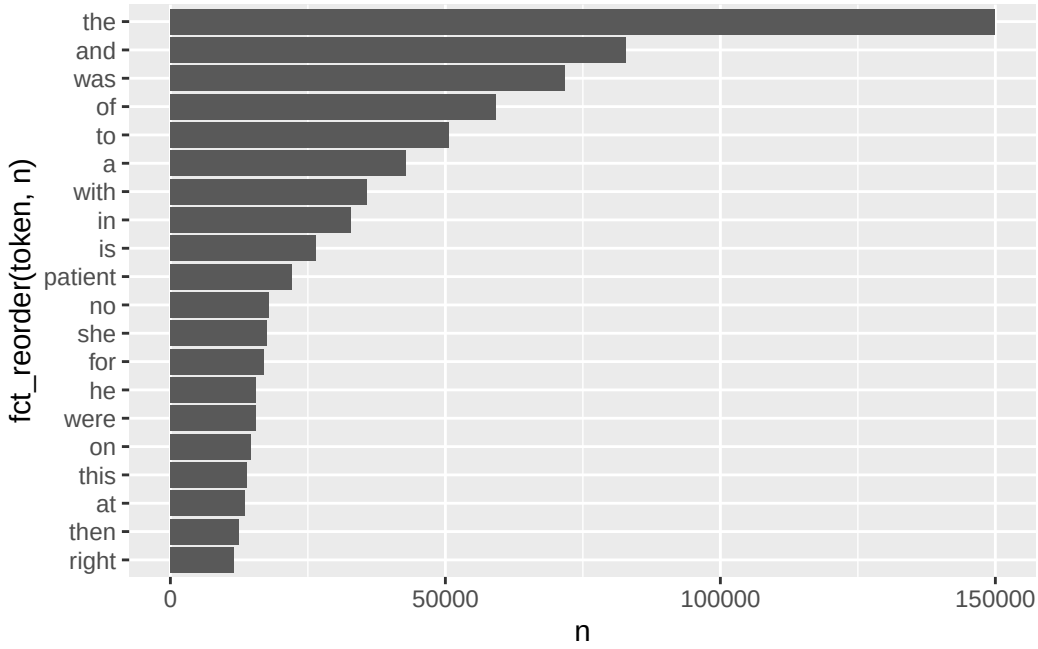
Visualize the top 20 most frequent words

Explain what we see from this result. Does it makes sense? What insights (if any) do we get?

```
mt_samples |>
  unnest_tokens(token, transcription) |>
  count (token, sort=TRUE) |>
  top_n(20,n)
```

```
# A tibble: 20 x 2
   token        n
   <chr>    <int>
 1 the     149888
 2 and      82779
 3 was      71765
 4 of       59205
 5 to       50632
 6 a        42810
 7 with     35815
 8 in       32807
 9 is       26378
10 patient  22065
11 no       17874
12 she      17593
13 for      17049
14 he       15542
15 were     15535
16 on       14694
17 this     13949
18 at       13492
19 then     12430
20 right    11587
```

```
mt_samples |>
  unnest_tokens(token, transcription) |>
  count (token, sort=TRUE) |>
  top_n(20,n)|>
  ggplot(aes(n,fct_reorder(token,n))) + geom_col()
```



The top 20 frequent words such as “the” “and” “was” are very common words we use in our everyday sentences. We would have to remove stop words to get a better picture of the main messages behind the transcription.

Question 3

Redo visualization but remove stopwords before

Bonus points if you remove numbers as well

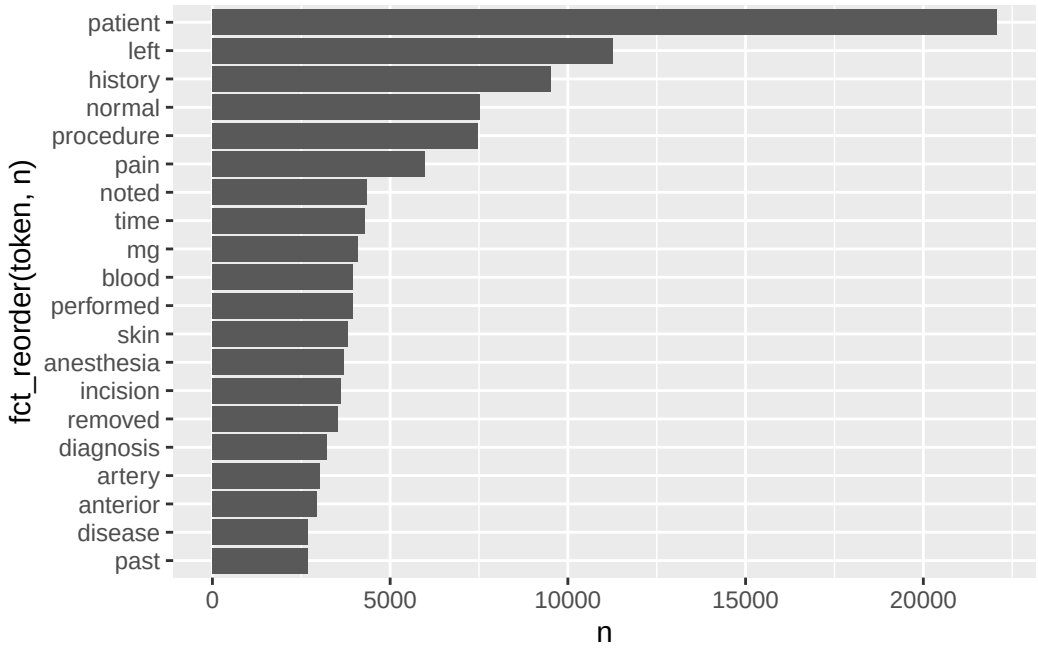
What do we see know that we have removed stop words? Does it give us a better idea of what the text is about?

```
mt_samples |>
  unnest_tokens(token, transcription) |>
  anti_join(stop_words, by = c("token" = "word")) |>
  filter(!str_detect(token, "^[0-9]+$")) |>
  count(token, sort = TRUE)
```

```
# A tibble: 22,348 x 2
   token         n
   <chr>     <int>
 1 patient   22065
```

```
 2 left       11258
 3 history     9509
 4 normal      7526
 5 procedure   7463
 6 pain        5976
 7 noted       4348
 8 time        4287
 9 mg          4087
10 blood       3956
# i 22,338 more rows
```

```
mt_samples |>
 unnest_tokens(token, transcription) |>
  anti_join(stop_words, by = c("token" = "word")) |>
  filter(!str_detect(token, "^[0-9]+$")) |>
  count(token, sort = TRUE) |>
    top_n(20,n)|>
  ggplot(aes(n,fct_reorder(token,n))) + geom_col()
```



After the removal of stop words (and numbers), it gives a much better visualization that the transcription is about medical examinations. Of the top words, the top three now are "patient," "left," and "history" which makes much more sense.

Question 4

Repeat question 2, but this time tokenize into bi-grams. how does the result change if you look at tri-grams?

```
#bigrams

mt_samples |>
  unnest_ngrams(ngram, transcription, n = 2) |>
   count(ngram, sort = TRUE)
```

```
# A tibble: 301,419 x 2
   ngram           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
# i 301,409 more rows
```

```
#trigrams
mt_samples |>
  unnest_ngrams(ngram, transcription, n = 3) |>
   count(ngram, sort = TRUE)
```

```
# A tibble: 655,442 x 2
   ngram                  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 655,432 more rows
```

The significance of bigrams and trigrams seem similar. In this lab, I will answer question 5 with bigrams.

Question 5

Using the results you got from questions 4. Pick a word and count the words that appears after and before it.

```
mt_samples |>
  unnest_ngrams(ngram, transcription, n =2) |>
    separate(ngram, into = c("word1", "word2"), sep = " ") |>
  select(word1, word2) |>
  filter(word2 == "pain") |>
   count(word1, sort = TRUE)
```

```
# A tibble: 326 x 2
   word1         n
   <chr>     <int>
 1 chest       707
 2 back        561
 3 abdominal   461
 4 the         361
 5 neck        218
 6 for         178
 7 of          163
 8 and         130
 9 his         127
10 her         115
# i 316 more rows
```

```
mt_samples |>
  unnest_ngrams(ngram, transcription, n =2) |>
    separate(ngram, into = c("word1", "word2"), sep = " ") |>
  select(word1, word2) |>
  filter(word1 == "pain") |>
   count(word2, sort = TRUE)
```

```
# A tibble: 537 x 2
   word2           n
   <chr>       <int>
 1 and           500
 2 in            323
 3 or            255
 4 with          231
 5 is            199
 6 he            191
 7 the           185
 8 she           165
 9 no            147
10 medications   103
# i 527 more rows
```

Question 6

Which words are most used in each of the specialties. you can use group_by() and top_n() from dplyr to have the calculations be done within each specialty. Remember to remove stopwords. How about the most 5 used words?

```
specialties <- mt_samples |>
  unnest_tokens(token, transcription) |>
  anti_join(stop_words, by = c("token" = "word")) |>
  filter(!str_detect(token, "^[0-9]+$")) |>
  group_by(medical_specialty) |>
  count(token, sort = TRUE) |>
  top_n(5,n) |>
  ungroup()

specialties
```

```
# A tibble: 210 x 3
   medical_specialty          token          n
   <chr>                      <chr>      <int>
 1 Surgery                    patient     4855
 2 Surgery                    left        3263
 3 Surgery                    procedure   3243
 4 Consult - History and Phy. patient     3046
 5 Consult - History and Phy. history     2820
 6 Orthopedic                 patient     1711
```

```
 7 Surgery                    anesthesia  1687
 8 Surgery                    incision    1641
 9 Cardiovascular / Pulmonary left        1550
10 Cardiovascular / Pulmonary patient     1516
# i 200 more rows
```

The five most used words are “patient,” “left” “procedure” “patient” and “history” represented in the medical specialties of Surgery and “Consult-History and Phy.” as in the general medical consult notes.