Lab 08 - Text Mining

William Zhang

```r
knitr::opts_chunk$set(eval = T, include  = T)
```

Learning goals

- Use `unnest_tokens()` and `unnest_ngrams()` to extract tokens and ngrams from text.
- Use dplyr and ggplot2 to analyze text data

Lab description

For this lab we will be working with the medical record transcriptions from https://www.mtsamples.com/. And is loaded and "fairly" cleaned at https://github.com/JSC370/jsc370-2023/blob/main/data/medical_transcriptions/.

Setup packages

You should load in `dplyr`, (or `data.table` if you want to work that way), `ggplot2` and `tidytext`. If you don't already have `tidytext` then you can install with

```r
# install.packages("tidytext")
```

read in Medical Transcriptions

Loading in reference transcription samples from https://www.mtsamples.com/

```r
library(tidytext)
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
data_url <- paste0(
  "https://raw.githubusercontent.com/JSC370/",
  "jsc370-2023/main/data/medical_transcriptions/mtsamples.csv"
  )
mt_samples <- read_csv(data_url)
mt_samples <- mt_samples |>
  select(description, medical_specialty, transcription)
head(mt_samples)
```

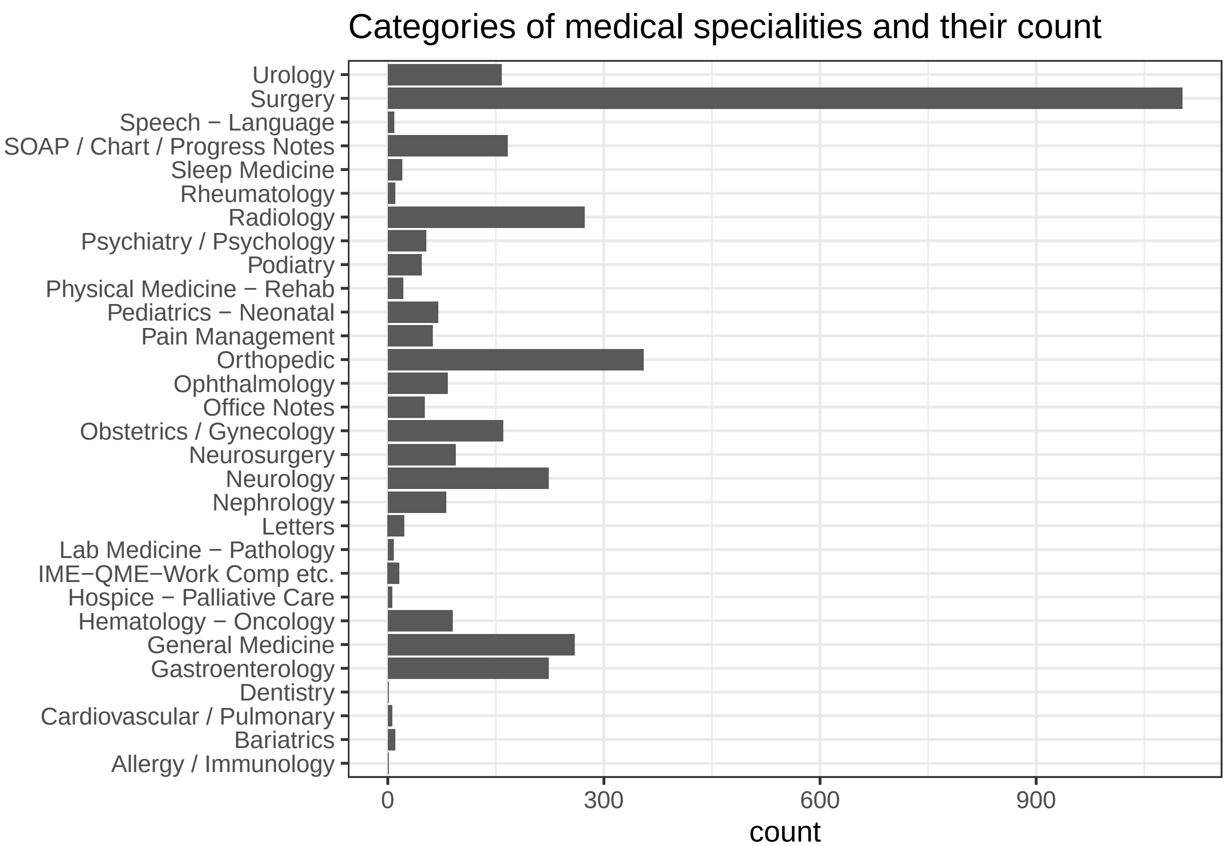
```
## # A tibble: 6 x 3
##   description                                                   medic~1 trans~2
##   <chr>                                                         <chr>   <chr>
## 1 A 23-year-old white female presents with complaint of allergi~ Allerg~ "SUBJE~
## 2 Consult for laparoscopic gastric bypass.                      Bariat~ "PAST ~
## 3 Consult for laparoscopic gastric bypass.                      Bariat~ "HISTO~
```

```
## 4 2-D M-Mode. Doppler.                                    Cardio~ "2-D M~
## 5 2-D Echocardiogram                                      Cardio~ "1.  T~
## 6 Morbid obesity.  Laparoscopic antecolic antegastric Roux-en-Y~ Bariat~ "PREOP~
## # ... with abbreviated variable names 1: medical_specialty, 2: transcription
```

Question 1: What specialties do we have?

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

```
mt_samples |> count(sort = TRUE, medical_specialty) %>%
  ggplot(aes(medical_specialty, n)) +
  geom_col() +
  theme_bw() +
  coord_flip() +
  ggtitle("Categories of medical specialities and their count") +
  labs(x = NULL, y = "count")
```



Comment: The distribution of the categories seems uni-modal, this is because "Surgery" has a significant higher count than all other categories. Altough there are also some categories like "Radiology", "Orthopedic" that have a decent amount of appearances, overall they are roughly similar.

Question 2

- Tokenize the the words in the `transcription` column

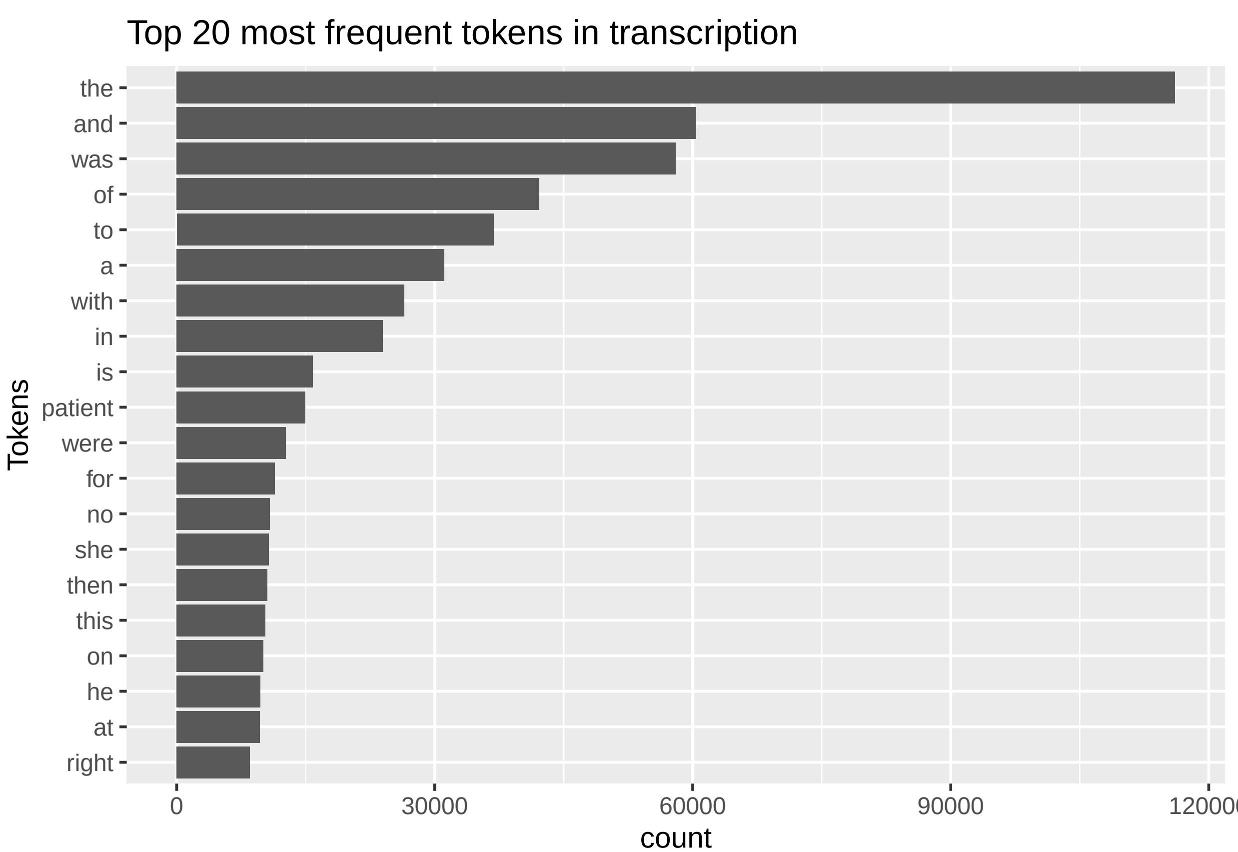
- Count the number of times each token appears
- Visualize the top 20 most frequent words

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

```r
mt_samples %>% unnest_tokens(token, transcription) %>% count(token, sort=TRUE)
```

```
## # A tibble: 22,830 x 2
##    token        n
##    <chr>    <int>
##  1 the     116095
##  2 and      60381
##  3 was      58047
##  4 of       42147
##  5 to       36842
##  6 a        31120
##  7 with     26462
##  8 in       23955
##  9 is       15842
## 10 patient  14971
## # ... with 22,820 more rows
```

```r
library(forcats)
mt_samples %>% unnest_tokens(token, transcription) %>% count(token) %>%  top_n(20, n) %>%
  ggplot(aes(n, fct_reorder(token, n))) +
  geom_col() +
  labs(y = "Tokens", x = "count") +
  ggtitle("Top 20 most frequent tokens in transcription")
```



Comment: We see that the most frequent tokens are stop words like "the", "and", "is", "was", and such. This is reasonable since usually in most texts, the stopwords make up for a large proportion of the text, which means we can't really gain much insight about the actual content of the dataset.

Question 3

- Redo visualization for the top 20 most frequent words after removing stop words
- 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?

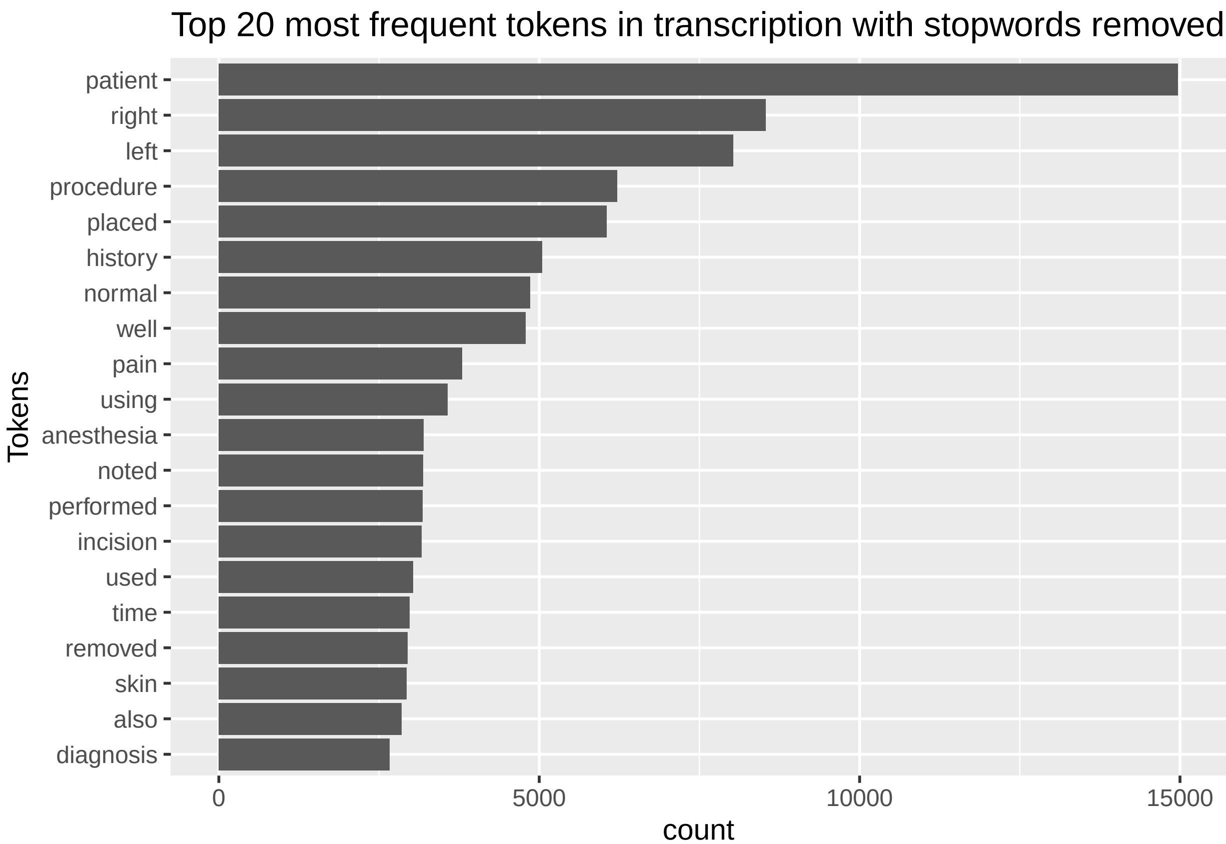
```
library(stopwords)
head(stopwords("english"))
```

```
## [1] "i"      "me"     "my"     "myself" "we"     "our"
```

```
length(stopwords("english"))
```

```
## [1] 175
```

```
mt_samples %>% unnest_tokens(token, transcription) %>%
  filter(!token %in% stopwords("english")) %>%
  filter(stringr::str_detect(token, "[^0-9].")) %>%  # remove numbers
  count(token, sort = TRUE) %>%
  top_n(20, n) %>%
  ggplot(aes(n, fct_reorder(token, n))) +
  geom_col() +
  labs(y = "Tokens", x = "count") +
  ggtitle("Top 20 most frequent tokens in transcription with stopwords removed")
```



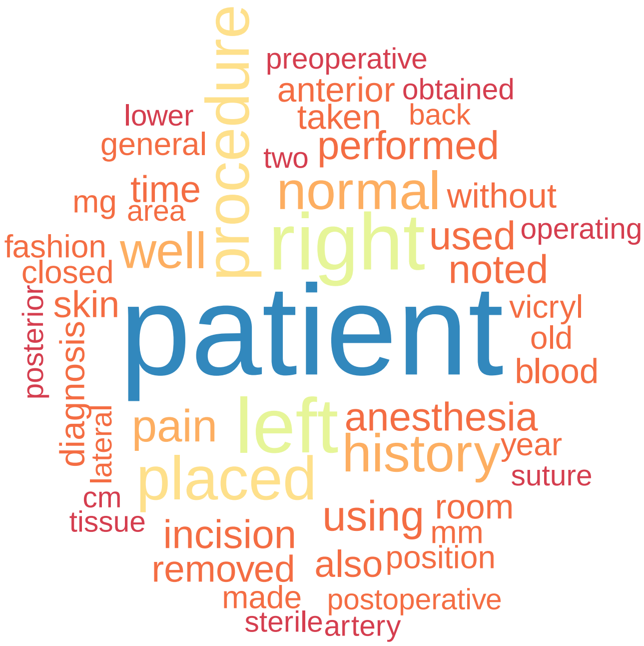
Comment: After removing the stopwords and the numbers, we see that now the most frequent tokens are quite representive of words that typically appear in medical transcripts, like "patients", "procedure", "history", and such.

Another method for visualizing word counts is using a word cloud via `wordcloud::wordcloud()`. Create a world cloud for the top 50 most frequent words after removing stop words (and numbers).

```
library(wordcloud)
```

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

```
pal <- brewer.pal(8, "Spectral")
mt_samples %>% unnest_tokens(token, transcription) %>%
  filter(!token %in% stopwords("english")) %>%
  filter(stringr::str_detect(token, "[^0-9].")) %>%
  count(token, sort = TRUE) %>%
  top_n(50, n) %>%
  with(wordcloud(token, n, random.order = FALSE, max.words = 100, colors=pal))
```



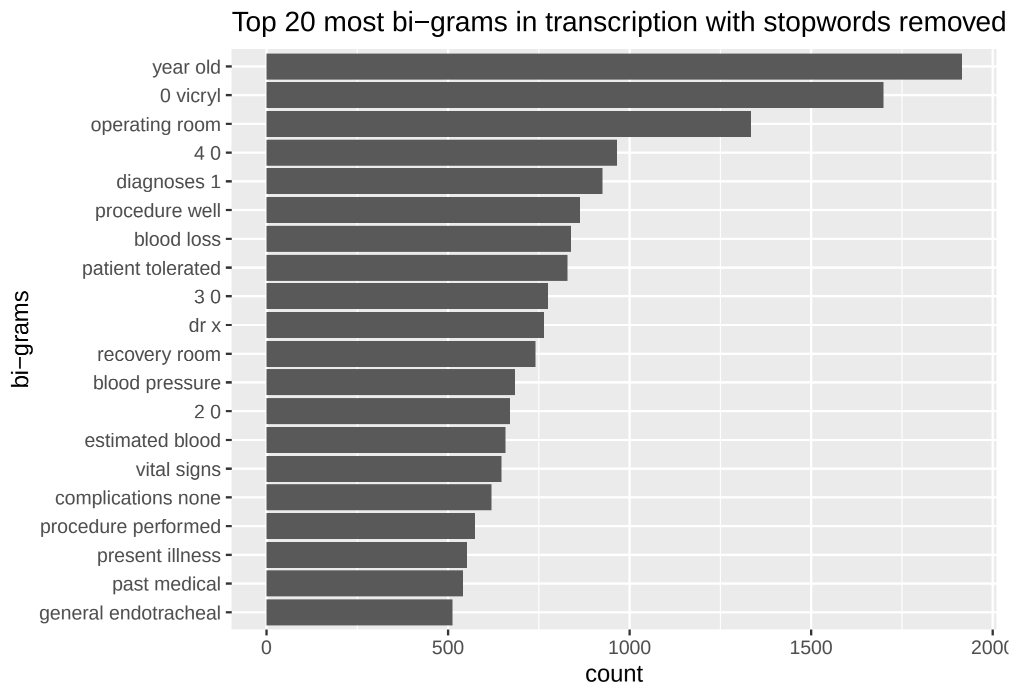
Question 4

Repeat question 3, but this time tokenize into bi-grams. How does the result change if you look at tri-grams? (You don't need to create the word clouds.)

```
sw_start <- paste0("^", paste(stopwords("english"), collapse = "|^"), collapse = " ")
sw_end <- paste0(" ", paste(stopwords("english"), collapse = "$|"), collapse = "$")

bigram <- mt_samples %>% select(transcription) %>%
  unnest_tokens(ngram, token = "ngrams", transcription, n = 2) %>%
  filter(!grepl(sw_start, ngram, ignore.case =TRUE)) %>%
    filter(!grepl(sw_end, ngram, ignore.case =TRUE)) %>%
  count(ngram, sort = TRUE)
```

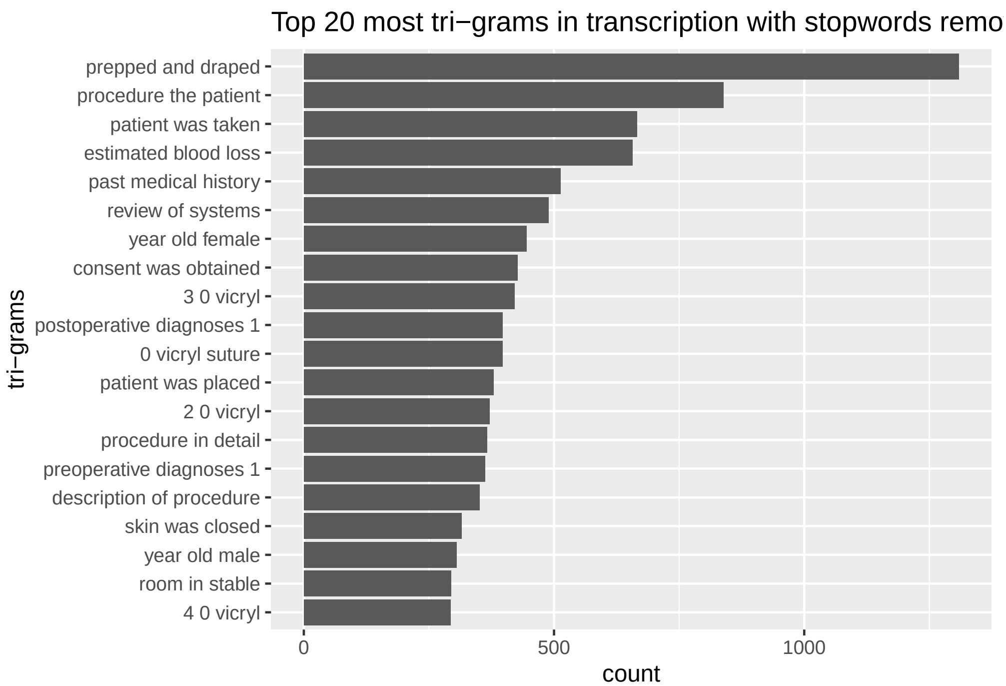
```
# bar plots
bigram %>% top_n(20, n) %>%
  ggplot(aes(n, fct_reorder(ngram, n))) +
  geom_col() +
  labs(y = "bi-grams", x = "count") +
  ggtitle("Top 20 most bi-grams in transcription with stopwords removed")
```



Tri-gram:

```
trigram <- mt_samples %>% select(transcription) %>%
  unnest_tokens(ngram, token = "ngrams", transcription, n = 3) %>%
  filter(!grepl(sw_start, ngram, ignore.case =TRUE)) %>%
  filter(!grepl(sw_end, ngram, ignore.case =TRUE)) %>%
  count(ngram, sort = TRUE)
```

```
trigram %>% top_n(20, n) %>%
  ggplot(aes(n, fct_reorder(ngram, n))) +
  geom_col() +
  labs(y = "tri-grams", x = "count") +
  ggtitle("Top 20 most tri-grams in transcription with stopwords removed")
```



Comment: For bigram, phrases that often come in two like “year old”, “operating room”, “blood pressure” appears more often. Similarly, for trigram, phrases that often in three words like “prepped and draped”, “past medical history”, “estimated blood loss” appears more often.

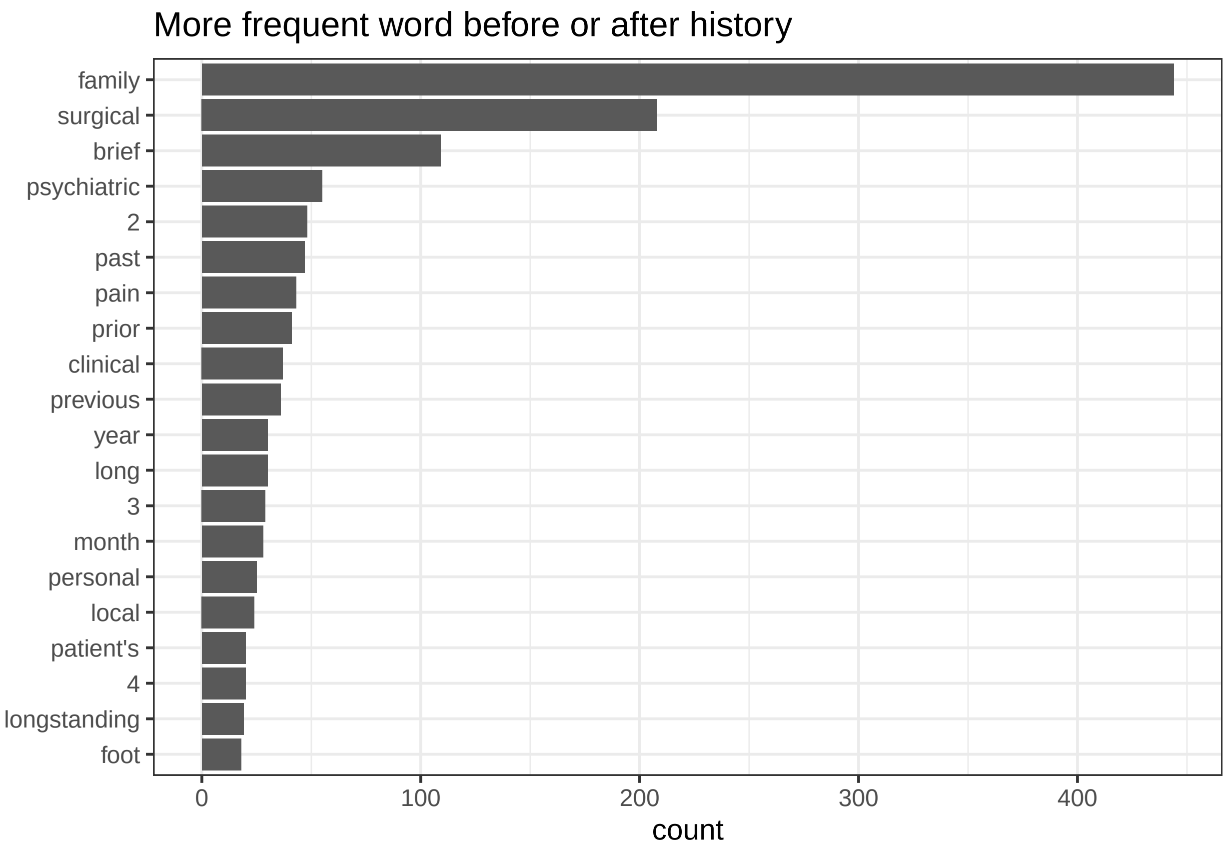
Question 5

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

```
library(stringr)

# We pick the word "history"

bigram %>% filter(str_detect(ngram, regex("\\shistory$|^history\\s"))) %>%
  mutate(word = str_remove(ngram, "history"),
         word = str_remove_all(word, " ")) %>%
  slice_max(n, n = 20) %>%
  ggplot(aes(reorder(word, n), n)) +
  theme_bw() +
  labs(y = "count", x = NULL) +
  ggtitle("More frequent word before or after history") +
  geom_bar(stat = "identity") +
  coord_flip()
```



Comment: With no surprise, "family" pairs very often with history, then follows with "surgical". These are all common phrases we would see in medical transcripts.

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 stop words. How about the most 5 used words?

```
mt_samples %>% unnest_tokens(token, transcription) %>%
  filter(!token %in% stopwords("english")) %>%
  filter(stringr::str_detect(token, "[^0-9].")) %>%
  group_by(medical_specialty) %>%
  count(token, sort = TRUE) %>%
  top_n(5, n)
```

```
## # A tibble: 171 x 3
## # Groups:   medical_specialty [30]
##    medical_specialty token         n
##    <chr>             <chr>     <int>
##  1 Surgery           patient    4855
##  2 Surgery           left       3263
##  3 Surgery           right      3261
##  4 Surgery           procedure  3243
##  5 Surgery           placed     3025
##  6 Orthopedic        patient    1711
```

```
##  7 General Medicine  patient    1356
##  8 Orthopedic        right      1172
##  9 General Medicine  history    1027
## 10 Orthopedic        left        998
## # ... with 161 more rows
```

Comment: For almost all specialties, "patient" is the word that appears most often, then follows by "left" and "right". "History" and "normal" are also common occurence words.

Question 7 - extra

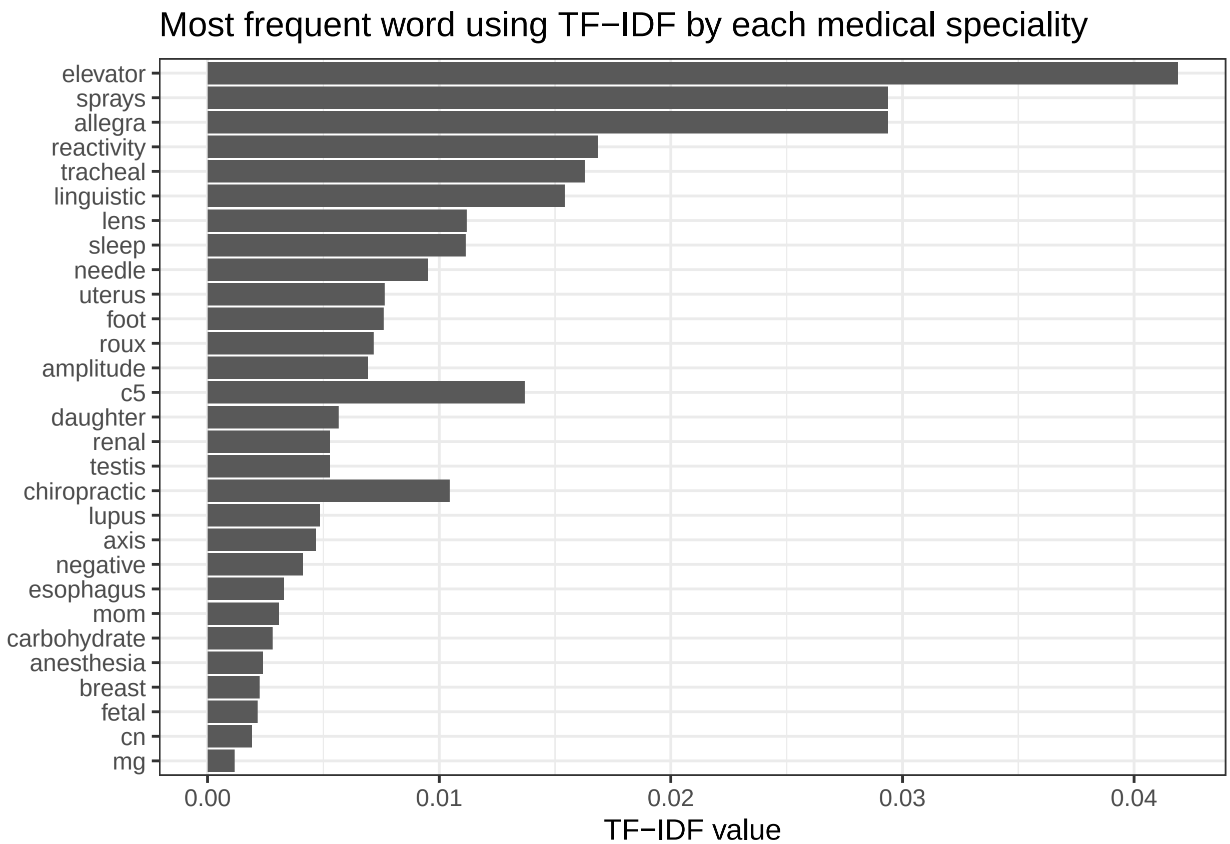
Find your own insight in the data:

Ideas:

- Use TF-IDF to see if certain words are used more in some specialties then others. Compare the list of words compared to the list from Question 6.

```
tf_idf_byspeciality <- mt_samples %>%
  unnest_tokens(word, transcription) %>%
  filter(
    !word %in% stopwords("english")
    ) %>%
  count(word, medical_specialty) %>%
  bind_tf_idf(word, medical_specialty, n)

tf_idf_byspeciality %>%
  group_by(medical_specialty) %>%
  slice_max(tf_idf, n = 1) %>%
  ggplot(aes(reorder(word, tf_idf), tf_idf)) +
  theme_bw() +
  labs(y = "TF-IDF value", x = NULL) +
  ggtitle("Most frequent word using TF-IDF by each medical speciality") +
  geom_bar(stat = "identity") +
  coord_flip()
```



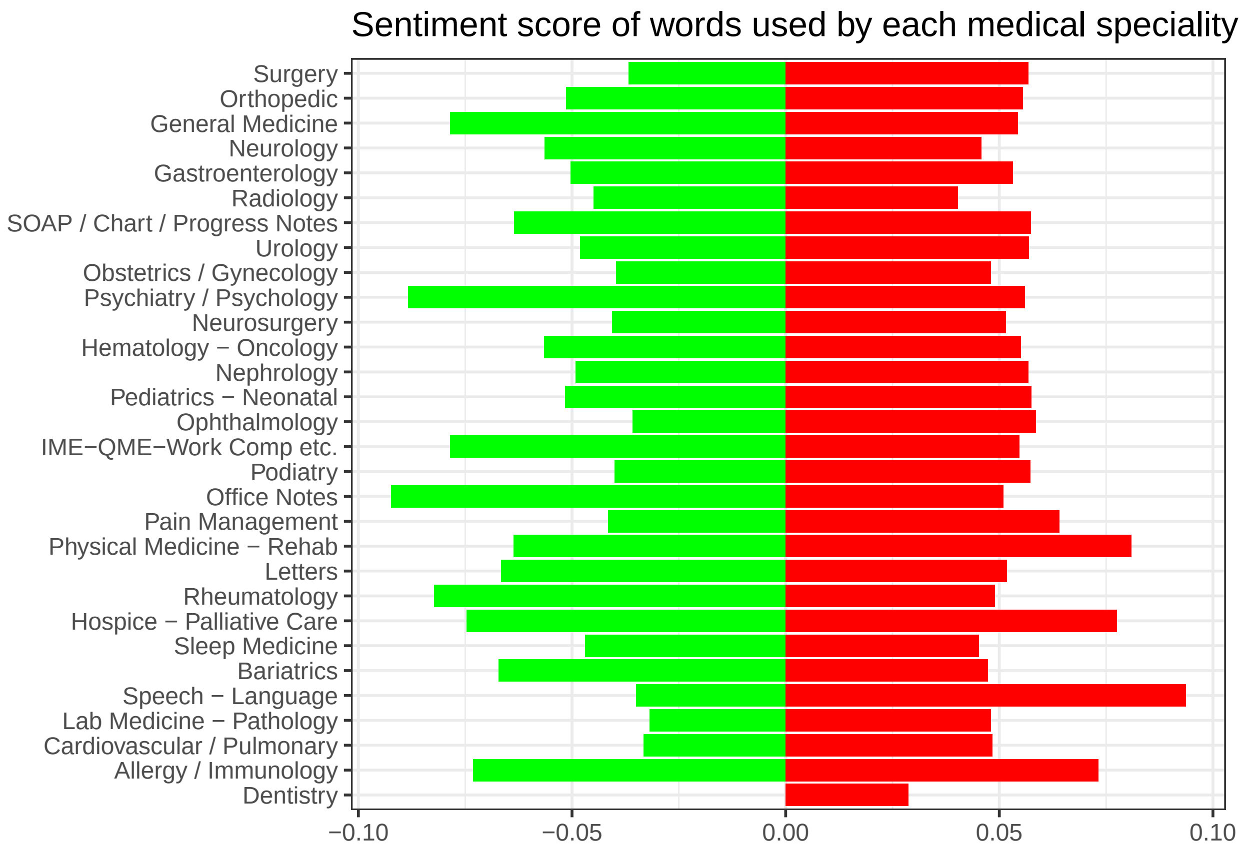
- Sentiment analysis to see if certain specialties are more optimistic than others. How would you define "optimistic"?

```
sentiment_list <- get_sentiments("bing")

sentiments_in_med <- tf_idf_byspeciality %>%
  left_join(sentiment_list, by = "word")

sentiments_in_med_by_sp <- sentiments_in_med %>%
  group_by(medical_specialty) %>%
  summarise(
    n_positive = sum(ifelse(sentiment == "positive", n, 0), na.rm = TRUE),
    n_negative = sum(ifelse(sentiment == "negative", n, 0), na.rm = TRUE),
    n = sum(n)
  )

sentiments_in_med_by_sp %>%
  ggplot(aes(reorder(medical_specialty, n_negative + n_positive))) +
  theme_bw() +
  geom_col(aes(y = -n_negative / n ), fill = "green") +
  geom_col(aes(y = n_positive / n ), fill = "red") +
  labs(x = NULL, y = NULL) +
  ggtitle("Sentiment score of words used by each medical speciality") +
  coord_flip()
```



- Find which specialty writes the longest sentences.

```
mt_samples %>% group_by(medical_specialty) %>%
  summarise(mean_length = mean(nchar(transcription), na.rm = TRUE)) %>%
  arrange(desc(mean_length))
```

```
## # A tibble: 30 x 2
##    medical_specialty         mean_length
##    <chr>                           <dbl>
##  1 IME-QME-Work Comp etc.          6733.
##  2 Psychiatry / Psychology         5125.
##  3 Hospice - Palliative Care       4238.
##  4 Neurosurgery                    3724.
##  5 Orthopedic                      3639.
##  6 Neurology                       3247.
##  7 Surgery                         3174.
##  8 Rheumatology                    3132.
##  9 Obstetrics / Gynecology         3090.
## 10 Letters                         3075.
## # ... with 20 more rows
```

Comment: It seems like that this "IME-QME-Work Comp etc." specialty writes the longest transcriptions with a mean character length of 6733, followed by the "Psychiatry / Psychology" specialty with a mean character length of 5124.

Deliverables

1. Questions 1-7 answered, pdf or html output uploaded to Quercus