Text analysis SSI

Gabriella

11/10/2021

# Script

Prior to analysis, text processing was performed, and a cleaning function was created to clean the text (the script is open access and can be found on GitHub [link]). The function created was adapted from code written in the ‘*textclean*’ package [@rinker2018] which provides tools to clean and process text. In this function, numbers were converted to text form, abbreviations were replaced with full words (using the ’*qdap’* abbreviation dictionary [@rinker2018a]). Text was tokenized and analysed using *tidytext* package which provides access to stopword lexicons. Tokenizing the data makes words lower case, strips punctuation and splits it into a one word per row format (or two adjacent words, a bigram). We omitted “spotting” from the lemmatizing because people talked about “blood spotting” and an acne “spot”. Stopwords were removed, and we supplemented the stop word list with our own list of undesirable words. These lists removed commonly used words that did not add meaning to the sentences. We also lemmatized words so that words derived from others, such as “long” and “longer” were grouped by their stem version (long), this was done using a look up dictionary in the lemmatize\_strings function from the *textstem* package. This uses Mechura’s (2016) English lemmatization list available from the lexicon package. We manually create a variable called “too soon to tell”, any responses that included statements such “as too soon to tell”,” haven’t had my period yet since the vaccine” or “not due yet”were counted as “too soon to tell”.

Many of the bigrams were pairs of common words like “the and, ‘to be’, so a second stage of stopword removal was applied. This was done using *tidyr* package separate() function, which splits a column into multiple based on a delimiter. This lets us separate it into two columns, “word1” and “word2”, at which point we can remove cases where either is a stop-word. Once the stop words and undesirable words were removed, bigrams were reunited using tidyr’s unite() function to recombine the columns into one.

We turned the dataset into a wide matrix using *widyr* and looked at relationships among words, by looking at words that co-occur, that are not necessarily directly adjacent, and from this we could calculate the correlation among words (how often they appear together relative to how often they appear separately), which was performed using the the phi coefficient (a common measure for binary correlation).

Finally we used pairwise\_count() from *widyr* to determine how many times pairs of words co occur in a response. These were visualised as a network of co-occuring words using the *ggraph* package.

## Data cleaning

### Pck

#Packages used:  
  
options(repos="https://cran.rstudio.com" )  
  
# install.packages("tidytext")  
# install.packages("dplyr")  
# install.packages("tidyverse")  
# install.packages("tidyr")  
# install.packages("igraph")  
# install.packages("ggraph")  
# install.packages("ggpubr")  
# install.packages("magrittr")  
# install.packages("igraph")  
# install.packages("ggplot2")  
# install.packages("knitr")  
# install.packages("gridExtra")  
# install.packages("textstem")  
# devtools::install\_github('cttobin/ggthemr')  
# install.packages("patchwork")  
#install.packages("xlsx")  
#devtools::install\_github('Mikata-Project/ggthemr')  
#install.packages("devtools")  
  
  
library(kableExtra)  
library(ggpubr)  
library(xlsx)  
library(tidyverse) # Tidy Messy Data  
library(dplyr)   
library(magrittr)  
library(tidytext) # Tidy Messy Data   
library(textstem)  
library(gridExtra)  
library(textclean)  
library(qdap)  
library(tm)  
library(igraph)  
library(ggraph)   
library(widyr)  
library(scales)  
library(patchwork)  
#Manually assign %>% to magrittr because the split chain error keeps coming up   
'%>%' <- magrittr::'%>%'  
library(tidyr)   
library(knitr)  
library(here)  
library(rio)

### data load

#df\_raw <-import(here("data/data\_textanalysis.txt"), header=T)   
df\_raw <- read.table("data/dataextanalysis.text")  
  
# Save raw version   
df\_clean<-df\_raw

### cleaning

#Data frame is df\_clean  
  
# Check text for potential issues  
check\_text(df\_clean$vaccine\_changescycle\_othertxt)  
  
# Function to expand contractions in the text, and replace minor spelling issues  
fix.contractions <- function(doc) {  
 # "won't" is a special case as it does not expand to "wo not"  
 doc <- gsub("it’s", "it has", doc)  
 doc <- gsub("won't", "will not", doc)  
 doc <- gsub("can’t", "can not", doc)  
 doc <- gsub("can't", "can not", doc)  
 doc <- gsub("don't", "do not", doc)  
 doc <- gsub("don’t", "do not", doc)  
 doc <- gsub("'ll", " will", doc)  
 doc <- gsub("it's", " it has", doc)  
 doc <- gsub("'re", " are", doc)  
 doc <- gsub("'ve", " have", doc)  
 doc <- gsub("'m", " am", doc)  
 doc <- gsub("'d", " would", doc)  
 doc <- gsub("i'm", "i am", doc)  
 doc <- gsub("isn't", "is not", doc)  
 doc <- gsub("isn’t", "is not", doc)  
 doc <- gsub("haven't", "have not", doc)  
 doc <- gsub("haven’t", "have not", doc)  
 doc <- gsub("menstral", "menstrual", doc)  
  
 # 's could be 'is' or could be possessive: it has no expansion  
 return(doc)  
}  
  
# Apply the function  
df\_clean$vaccine\_changescycle\_othertxt <- sapply(df\_clean$vaccine\_changescycle\_othertxt, fix.contractions)  
  
# Make cleaning function  
cleaning <- function(x) {  
 x <- replace\_number(x) # Converts numbers to text form  
 x <-replace\_abbreviation(  
 x,  
 abbreviation = qdapDictionaries::abbreviations,  
 replace = NULL,  
 ignore.case = TRUE  
) # QDap'a replace abbreviation   
 #x<-na.omit(x)  
 x <- replace\_html(x) # Removes HTML tags and symbols stick  
 x <- replace\_kern(x) # The replace\_kern function looks for 3 or more consecutive capital letters with spaces in between and removes the spaces (e.g. the B O M B to BOMB)  
 # x <- replace\_incomplete(x) # replaces end of sentances without punctuation with '|'  
 x <- tolower(x) # makes everything lower case  
 x <- stripWhitespace(x) # removes extra white space  
}  
  
## Use the cleaning function   
df\_clean$vaccine\_changescycle\_othertxt<- cleaning(df\_clean$vaccine\_changescycle\_othertxt)

## Calculate and remove data including “Too soon to tell”

# Create too soon to tell variable, done three times as it doesn't work if its too long  
df\_clean<-df\_clean %>%  
 mutate(  
 toosooner = if\_else(  
 stringr::str\_detect(`vaccine\_changescycle\_othertxt`, "(jab this week|can not be sure yet, it's only been a few weeks|didn't have menstruation yet after being vaccinated.|waiting for period since vaccine given mid cycle|only had the vaccine two days ago|can not be sure yet|not menstruated since my vaccine yet|not enough data|no period since my vaccination|can not say yet|can't say|can’t say|my vaccine was yesterday!|not sure just had the vaccine|i haven't been due my period since the vaccination|i have not been due my period since the vaccination|should be end of month|i am yet to have my monthly bleed|haven't had a new cycle since the vaccine|no, but only had one full cycle since the jab|not had a cycle since vaccination|not had a cycle yet|unsure at present time|too soon to know|the vaccination was yesterday, it's a little too soon to tell|i have not had a full cycle since having the vaccine|(^not had a period yet$)|waiting to come on in the next few days|(^too soon to tell (had the vaccine two days ago)$)|i have only had one,too soon to say|(^too soon to say$)|i have not had a period since as it isn't due yet|(^not had one yet$)|(^i have not had one since (due this week), so can not report if any changes will occur$)|(^i have not menstruated since vaccine as only been three weeks.$)|not applicable, was only vaccinated two days ago|(^too soon to tell$)|my jab was two days ago so i cannot say|too early to tell either way|not had one since vaccine. due in a couple of days|to recent to comment|vaccinated twenty threerd feb two thousand twenty one too early to say if theres been an impact|n/a|was only vaccinated two days ago|way too soon to tell|waiting for bleeding now|vaccinated three days ago so do not know|a bit soon to tell, but i am not expecting any changes!|vaccine occurred on ten/three/twenty one. period is not due for another week, so not sure as yet if my cycle has been interrupted by the vaccine yet.|vaccine too recent to see effect|vaccine was only two days ago (on the last day of my last period)|unsure|too soon|too recent to say|too recent to tell|too recently vaccinated to tell (next week)|too short a time period to notice any effect|(^too early to tell$)|(^too soon to answer$)|too early to say|(^too early to state$)|too early to really comment|to early to say|to early to tell, was starting new cycle day of shot|to early to tell.|to early to tell. although it was thirty five days since last period which is substantially longer than normal|to short of a time period to comment|to soon to say|am waiting for next period, too early to tell if disruptive at this stage.|not had a chance as i had my vaccination the day my period started.|do not know yet, only just had it.|its only been a week, not due a period yet|it's too early to say.|literally got it this morning|still waiting to see|recently vaccinated|my menstrual cycle is not due for another sixdays.|possibly too early to know|only received twond vaccination last week|only recently had the vaccine so it|only vaccinated on one/four/twenty one, too soon to answer.|(^too soon to tell.$)|only vaccinated on seventeenth march|only vaccinated two days|had vaccine yesterday and period is due now.|have not had a period yet, was only vaccinated last week, my periodo is due in two weeks.|have not menstruated since being vaccinated yet|i am due to start my onest menstrual cycle since being vaccinated this week,|i do not really know yet|have not yet had a period since the vaccine|have not yet had one since vaccination. due any day.|have not had since as not yet due|said no but only vaccinated a day|my onest period following having had the vaccine is not yet due|i got my first shot just after my last period, so i have not had another period yet.|however i have not noticed any changes in relation to my cycle.|only got the vaccine yesterday|only got vaccinated today.|have only just been vaccinated and have not had a period yet|have only just had injection three days ago and period is due tomorrow so i have not had chance to see if it has changed|have yet to have a period as only had vaccine this week.|have not had a period yet since my vaccine|have not had a period yet. only had the jab two weeks.  
 |have not had my next cycle yet, so unsure.|have not had one when accidentally forgot pill until now.|have not yet had a period since the vaccine.|have not bleed since my vaccine|i got the vaccine yesterday and i'm due to start my period this week so can not say for now|no but i am only a few weeks post vaccination|not applicable|not due on my period yet.|not due period for a time since vaccination|not due period yet since vaccination|not due yet so cannot comment|not enough time since jab to know|not enough time to have had a period yet|not enough time to tell|not enough time to tell, i had my first vaccine on twenty six/three/twenty one.|not had a menstrual cycle as of yet as i have just come off the day i had my vaccination.|not had a menstrual cycle or bled since having the vaccine yet|not had a period|not had a period since|not had a period since having vaccine|not had a period since my covid vaccination which was very recently|not had a period since.|not had a period yet|not had a period yet since jab|it's only been six days since vaccine so i have no idea.|it's been about a month after, so tough to say. i feel like i am about to get my period though|it's been irregular before so basically hadn't had a proper period in a while.  
 |it's not been long enough to check - i need to keep monitoring|jab was only five days ago!|just had it recently so can not tell any difference yet.|just had the jag|just had the vaccine|literally got it this morning (twenty three/three/two thousand twenty one)|my menstrual cycle isn't due for another sixdays.|my vaccination was this morning, but i am hoping it makes no difference!|not had a period yet.|not had chance as i had my vaccination the day my period started. and it only finished yesterday.  
 |not had my period yet|not had one yet as due shortly|not had one yet since|it has not been long enough to know yet|it has only been two weeks since my vaccine so cannot say|it has only been two.five weeks since i was vaccinated so it is too early to say.|it hasn't been long enough to tell a difference|it hasn't been long enough to say|it is too early to tell as i only had the vaccine three days ago.|it is too early to tell. my cooke was taken out on the same day as i had my jab. i had my period that day and have not stopped bleeding yet.|it is too soon to know at the moment.|it was only a week ago|it was only two days ago|it was this afternoon!|it's been less than a month so i have not yet had a period following my first jab.|it's been less than one cycle so i can not say yet really.|it's not been long enough to judge|not had one yet since vaccine so can not comment|not had period as yet|not had period since jab as finished day i had jab|not had period yet since vaccine but am due any day now|i was menstruating when i got my vaccine (this week)|i was vaccinated one week ago and my period isn't due for another week.|i was vaccinated two days before starting my period and have not yet seen any changes|i would say the vaccine was too recent to notice any changes.|^i've not had a period since$|i've not had a period since my vaccine|i've not had a period yet|i've not had a period yet since vaccination as i'm not due yet|about to have my onest post-vaccine period|i've not had my period yet|i've not yet received today's date nineteen/three/two thousand twenty one and receiving on twenty eight/three/two thousand twenty one and currently on period so have to wait next month (april) to find out|not had vaccine long enough to say|not known yet|not long enough since injection to know|due my first period anytime now which will be the first one since the vaccine.|not reached next expected period since first dose|i was menstruating when i got my vaccine)"),"unsure",NA\_character\_))  
   
   
 df\_clean<-df\_clean %>%  
 mutate(  
 toosoon = if\_else(  
 stringr::str\_detect(`vaccine\_changescycle\_othertxt`, "(not soon enough to tell yet as it was less than a week ago.|not sure as been irregular in last year so not sure when it's coming|not sure as only had two periods since vaccine was given|not sure as since covid not sure had a proper cycle|not yet applicable|not yet as i had my vaccine after my last period|not yet menstruated since first jab.|not yet menstruated since vaccine.|have not had a period since the jab, which was only four days ago|have not had a period yet after vaccine. next one should be april twelve.|have not had a period yet since vaccination (not due for another three weeks)|have not had one yet|i had my vaccination today therefore too soon to tell|(^i have not come on my period$)|i have just had the vaccine this week so too soon to say.|i have not had enough to judge|i have not had my period yet after vaccine|i have not yet had a cycle since the vaccine|i have not yet had a period due since the vaccine.|i have only had four periods before my vaccine and one since so i can not compare.|n/a as i have not had a period since receiving the vaccine a few days ago.|i have only had one cycle since vaccination, so i can not tell yet.|i have only had one cycle since vaccine|i have only just had it, so cannot tell yet|i have only received the vaccine a few days ago.|i have yet to have a period|i have yet to have a period as i was vaccinated very recently|i have not had a period for fifty four days but my vaccine was one week ago.|i have not had a period since my twond vaccine|i have not had a period yet since having my vaccine. it is due at the end of march|i have not had one since (due this week), so can not report if any changes will occur|i have not been due once since. only vaccinated twenty sixth march|i have not had a period since my first vaccination|i have not had a period yet|i have not had a period yet after receiving the vaccine.|i havent had a period since having first vaccine|i havnt had a period since my first vaccine.|i just got vaccinated a week ago so i couldn't observe any changes|i only gave birth thirteen days ago. no period just yet.|i only got it two days ago|i only got vaccinated two weeks ago so no changes as yet|i only had it yesterday so too soon to say|i only had the vaccine one days ago and my period started two days after making me a week early|i only received it a few days ago so have not had time to notice|i only received my first dose of the vaccine yesterday!|have not had period since yet|only had injection one week ago so not due to have a period yet|only had it three days ago )"),"unsure",NA\_character\_))  
   
 df\_clean<-df\_clean %>%  
 mutate(  
 toosoonest = if\_else(  
 stringr::str\_detect(`vaccine\_changescycle\_othertxt`, "( only had one since vaccine|only had the jab twenty fourhrs ago so too early to tell|only had the vaccination on twenty six/three and period started today twenty eight/three|only had the vaccine yesterday, too early to tell|only just had jab on the saturday nineteenth|do not know yet, i have just started the first period today since vaccine two weeks ago.|do not know|dont get yet my period|hard to say as i had my vaccine on my last day of bleeding so i have not had a period or completed a cycle since my vaccine as it was less than two weeks ago.|have not had since as not yet due|only just had my jab today.|(^i was vaccinated one week ago and my period is not due for another week.$)|(^it has only been six days since vaccine so i have no idea.$)|(^it has been less than a month so i have not yet had a period following my first jab.$)|(^  
it has not been long enough to check - i need to keep monitoring$)|(^it has not been long enough to judge$)|(^only given yesterday$)|(^not had chance as i had my vaccination the day my period started. and it only finished yesterday.$)|(^it hasn’t been long enough to say$)|(^only had it three days ago$)|(^too early to tell as my periods are very irregular anyway$)|(^only one cycle so far$)|(^too early$)|only just had the vaccine so too soon to tell|only just had vaccination - period is due this weekend.|only just had vaccine and period started two days later|only just had vaccine so too early to tell.|(^too early to tell.$)|only just received vaccine and not due a period yet so cannot comment|its too soon to notice|cannot comment due to stopping progesterone pill since having vaccine.|i just got vaccinated a week ago|(^too recent to know$)|too recent to know. my last period ended the day after my vaccination.|haven't had a period yet|can not be sure yet, it has only been a few weeks|i havnt had a period since my first vaccine|haven't had one since|hasn’t been long enough to know|have only just had injection three days ago and period is due tomorrow so i haven't had chance to see if it has changed|have yet to have a period as only had vaccine this week|it's too soon to say|only had my vaccination yesterday!|i only had it today so not sure|(^only had one period$)|been long enough since i had the jabs.|been long enough to know|only just had vaccine| only just had vaccination|have only just been vaccinated|had chance to see if it has changed|cannot comment|its too early to say|only been one week. too early to tell| haven't had a period yet should be end of month|havnt had one yet|cannot answer this question fairly|so it's difficult to know|vaccine today|vaccinated yesterday|i have not had one since the vaccine|(^can not tell yet$)|(^n/a$)|(^cant say yet$)|(^the vaccination was yesterday, it has a little too soon to tell$)|(^too early to say$)|(^not had a period yet$)|(^am waiting for next period, too early to tell if disruptive at this stage.$)|(^too early to say$)|(^have not had a period since the jab, which was only four days ago$)|(^have not had a new cycle since the vaccine$)|(^to recent to comment$)|(^have not had a period yet since vaccination (not due for another three weeks)$)|(^not had my period yet$)|(^can not answer - jab this week$)|(^can not be sure yet, it has only been a few weeks$)|(^can not say yet$)|(^can not tell, it hasn’t been long enough since i had the jabs.$)|(^do not know yet, i have just started the first period today since vaccine two weeks ago.$)|(^  
do not know yet, only just had it.$)|(^had vaccine today$)|(^vaccinated twenty threerd feb two thousand twenty one too early to say if theres been an impact$)|(^my vaccination was this morning, but i am hoping it makes no difference!$)|(^vaccination was yesterday, so it has a bit soon to tell, but i am not expecting any changes!$)|(^too soon to know$))"),"unsure",NA\_character\_))  
  
#Combine all the too soon to tells into 1 variable   
df\_clean<-select(df\_clean, vaccine\_changescycle\_othertxt, toosoon,toosooner,toosoonest,everything())   
table(df\_clean$toosoon) #39 unsure  
table(df\_clean$toosooner) #249 unsure  
table(df\_clean$toosoonest) #83 unsure  
  
#Make a variable that puts the text box in, but removing the 'unsure' bits and replacing with NA.   
df\_clean$trimother<-df\_clean$vaccine\_changescycle\_othertxt  
df\_clean<-select(df\_clean, vaccine\_changescycle\_othertxt, toosooner,toosoonest,toosoon,trimother, everything())   
  
df\_clean$trimother[!is.na(df\_clean$toosooner)] <- NA  
df\_clean$trimother[!is.na(df\_clean$toosoon)] <- NA  
df\_clean$trimother[!is.na(df\_clean$toosoonest)] <- NA  
  
df\_clean<-select(df\_clean, vaccine\_changescycle\_othertxt, toosoon,toosooner,toosoonest,everything())   
df\_clean <- subset(df\_clean, select = -c(toosoonest,toosoon,toosooner) )  
  
#Make a column of other for the new variable to make it easy   
df\_listoftoosoon<-select(df\_clean, trimother,id)   
df\_listoftoosoon$toosoon<-df\_listoftoosoon$trimother  
 df\_listoftoosoon<-df\_listoftoosoon %>% mutate(toosoon = if\_else(is.na(toosoon), 0, 1)) #1 response, 0 unsure  
df\_listoftoosoon$toosoon <- na\_if(df\_listoftoosoon$toosoon, "1")  
df\_listoftoosoon$toosoon[df\_listoftoosoon$toosoon =="0"]<-"unsure"  
  
#Export too soon as csv to check it works  
#write.xlsx(df\_listoftoosoon, "~/OneDrive - Nexus365/DPhil/PeriodsCOVID/data/IDtoosoon.xlsx", sheetName = "toosoon",row.names = TRUE, append = FALSE)

## 1. Commonality of words

### Unigrams

#Set colours for graphs   
vaccinetype <- c("#002147", "#8B3A3A")  
contraception<- c("#D02090", "#9AC0CD", "#458B00", "#436EEE")  
covidtype<-c("#4876FF", "#4D4D4D", "#7A378B")  
reprostate<- c("#698B69", "#BC8F8F", "#B0C4DE", "#FFD700")  
  
  
  
#tokenise words  
df\_token <- df\_clean %>%   
unnest\_tokens(word, trimother)  
  
# Lemmatize the words in new column  
#Because the lemmatising dictionary doesn't have "painful", I'm changing all painful to pain.   
fix.lemmas <- function(doc) {  
 # painful to pain  
 doc <- gsub("painful", "pain", doc)  
 return(doc)  
}  
  
#run that   
df\_token$word<-sapply(df\_token$word,fix.lemmas)  
#lemmatize words but omitting spotting   
df\_token$lemword<-lemmatize\_strings(df\_token$word, dictionary = lexicon::hash\_lemmas[!token%in%c("shot", "spotting"), ])  
df\_token$lemword<-as.character(df\_token$lemword)  
  
#Create list of undesirable words  
undersirablewords <- tibble(word = c("since","days","yet","one","two","tell","soon","due","had","first","day","week","twenty","time","three","weeks","jab","say","just","last","know","it's","started","ago","got","five","months","thousand","now","enough","haven't","next","eight","four","six","i've","today","also","came","march","much","thirty","year","recent","can","following","january","april","month","sure","yesterday","however","still","think","yes","period","NA","around","vaccinated","vaccine","vaccination","covid","twond","perhaps",NA,"i'm","cycle","periods","cycles","menstrual","february","pill","coil","contraceptive","seem","sunday","may","seven","taking","twelve","september","level","become","n","august","even","cannot","injection","ten","2","pre","nineteen","come","19","2","hadn ' t","notice","can't","though","fifty","although","happen","go","far","december","expect","etc","usual","usually","receive","onest","answer","can’t","cant", "hasn't","hasn ’ t","it ’ s","i ’ be", "i ’ have","haven ’ t","hadn ’ t","anyway","aren ’ t","ûª","ûªt","ûªve","ûªm","û","ûïmore","ûïperiod"))  
  
  
#Removing stop words   
df\_token<-df\_token %>%  
 anti\_join(undersirablewords, by=c("lemword"="word"))  
  
df\_token<-df\_token %>%  
 anti\_join(get\_stopwords(), by=c("lemword"="word"))  
  
#Count  
library(plyr)  
df\_token %>%  
 dplyr::count(word, sort = TRUE)  
  
## Most common single words   
uni\_common<- df\_token %>%  
 dplyr::count(word, sort = TRUE)  
  
  
## Most common single words LEM  
uni\_commonLEM<- df\_token %>%  
 dplyr::count(lemword, sort = TRUE)  
  
#Visualise top words  
graph\_topuniLEM<-df\_token %>%  
 # remove stop words  
 dplyr::count(lemword, sort = TRUE) %>%  
 slice\_max(n, n = 20) %>%  
 mutate(lemword = reorder(lemword, n)) %>%  
 # put `n` on the x-axis and `lemword` on the y-axis  
 ggplot(aes(n, lemword)) +  
 geom\_col(fill="#7AC5CD")+  
 labs(title="",x="Number of Times Used", y="Unigram")+  
 theme(axis.text=element\_text(size=12))+theme\_bw()

### Bigrams

############### Bigrams ###########  
  
# Bigram undesirable words  
bigramundesirable <- tibble(word = c(NA,"ago","vaccine","covid","thousand","ten","twond","twenty","threerd","onest","april","coil","fourteen","menstrual","2","19","january","eighteen"))  
  
# Lemmatize text  
df\_clean$lemword<-lemmatize\_strings(df\_clean$trimother, dictionary = lexicon::hash\_lemmas[!token%in%c("shot", "spotting"), ])  
  
# Tokenizing by adjacent words, bigrams.  
df\_bigram <- df\_clean %>%  
 unnest\_tokens(bigram, lemword, token = "ngrams", n = 2)   
  
#Seperate bigrams   
bigrams\_separated <- df\_bigram %>%  
 separate(bigram, c("word1", "word2"), sep = " ")  
  
#Remove stop words  
bigrams\_filtered <- bigrams\_separated %>%  
 filter(!word1 %in% stop\_words$word) %>%  
 filter(!word2 %in% stop\_words$word) %>%  
 filter(!word1 %in% bigramundesirable$word) %>%  
 filter(!word2 %in% bigramundesirable$word)  
  
  
# Reunite bigrams  
bigrams\_united <- bigrams\_filtered %>%  
 unite(bigram, word1, word2, sep = " ")  
  
# A few don't make sense when put back together, so remove them:   
bigrampairremove <- tibble(word = c("abdominal abdominal"))  
bigrams\_united <- bigrams\_united %>%  
 filter(!bigram %in% bigrampairremove$word)   
  
# Visualise top bigram  
graph\_topbiLEM<-bigrams\_united %>%  
 # remove stop words  
 dplyr::count(bigram, sort = TRUE) %>%  
 slice\_max(n, n = 20) %>%  
 mutate(bigram = reorder(bigram, n)) %>%  
 # put `n` on the x-axis and `lemword` on the y-axis  
 ggplot(aes(n, bigram)) +  
 geom\_col(fill="#7AC5CD")+  
 labs(title="",x="Number of Times Used", y="Bigram")+  
 theme(axis.text=element\_text(size=12))+theme\_bw()

### Figure

graph\_topuniandbi<-grid.arrange (graph\_topuniLEM,graph\_topbiLEM, ncol=2)  
ggsave(graph\_topuniandbi,filename="unibigram.pdf", height=5, width=10)

## 2. Co-occurence

# Creating Co occurence dataset  
df\_wide\_words <-df\_clean %>%   
 mutate(section = id) %>%  
 filter(section > 0) %>%  
 unnest\_tokens(lemword, vaccine\_changescycle\_othertxt) %>%  
 filter(!lemword %in% undersirablewords$word) %>%  
 filter(!lemword %in% stop\_words$word)  
  
#lemmatise all words except spotting   
df\_wide\_words$lemword<-lemmatize\_strings(df\_wide\_words$lemword, dictionary =   
lexicon::hash\_lemmas[!token%in%c("shot", "spotting"), ])  
#Removing stop words in lemmatised  
df\_wide\_words<-df\_wide\_words %>%  
 anti\_join(undersirablewords, by=c("lemword"="word"))  
  
df\_wide\_words<-df\_wide\_words %>%  
 anti\_join(get\_stopwords(), by=c("lemword"="word"))  
  
  
# count words co-occuring within sections  
library(widyr)  
word\_pairs <- df\_wide\_words %>%  
 pairwise\_count(lemword, section, sort = TRUE)  
  
# filter out uncommon words  
word\_cors <- df\_wide\_words %>%  
 group\_by(lemword) %>%  
 filter(n() >= 10) %>%  
 pairwise\_cor(lemword, section, sort = TRUE)  
  
  
#Now could find the words most correlated with a word like “irregular” using a filter operation.  
word\_cors %>%   
 filter(item1 == "cramp")->justtotry  
  
#Now choose particular interesting words and find the other words most associated with them.   
 # filter(item1 %in% c("stop","light","change","late","spotting")) %>%  
  
# Lets filter for words that are correlated over and below 0.05 and -0.05  
filteredword\_cors<-word\_cors  
  
filteredword\_cors<-filteredword\_cors %>%   
 filter (correlation > 0.10 |  
 correlation < -0.10)  
  
  
graph\_correlated\_wordsfiltered<-filteredword\_cors %>%  
 filter(item1 %in% c  
 ("bleed", "heavy","light","early",  
 "late","spotting","change","irregular",  
 "cramp","clotting","normal")) %>%  
 group\_by(item1) %>%  
 slice\_max(order\_by=correlation, n = 5) %>%  
 ungroup() %>%  
 mutate(item2 = reorder(item2, correlation)) %>%  
 ggplot(aes(item2, correlation)) +  
 geom\_bar(stat = "identity") +  
 facet\_wrap(~ item1, scales = "free\_y",ncol = 2) +  
 coord\_flip()+  
 labs(y="Phi Coefficient", x="Correlated Words", title = "Highest Correlated Words")

#### Correlation matrix

# Filter uncommon words, but less  
word\_cors <- df\_wide\_words %>%  
 group\_by(lemword) %>%  
 filter(n() >= 10) %>%  
 pairwise\_cor(lemword, section, sort = TRUE)  
  
#Make two datasets to use for this:   
# trying to make a count column to have numbered nodes  
counts <- dplyr::count(df\_token, lemword, sort = TRUE)  
   
# Join counts df to my word\_cors df  
attempteddf<-left\_join(word\_cors, counts,by=c(item1="lemword"))  
  
# edge data  
  
edge\_data <- attempteddf %>%   
# filter(correlation > 0.1 |correlation < -0.1) %>%  
 select(item1, item2, correlation)  
  
# from edgelist to matrix  
mygraph <- graph.data.frame(edge\_data)  
M<-get.adjacency(mygraph, sparse = FALSE, attr='correlation')  
  
  
# Plot figure  
library(corrplot)  
  
pdf(file="coorplot.pdf",width=10,height=10)  
  
corrplot(M, type = 'lower', order = 'hclust', tl.col = 'black', cl.ratio = 0.3, tl.srt = 10, col = brewer.pal(n = 10, name = 'Spectral'))   
dev.off()  
  
  
  
corrplot(M, type = 'lower', order = 'hclust', tl.col = 'black', cl.ratio = 0.3, tl.srt = 10, col = brewer.pal(n = 10, name = 'Spectral'))

#### Network

options(ggrepel.max.overlaps = 100) # so that all labels are represented  
  
#Again filter uncommon words, but less  
word\_cors <- df\_wide\_words %>%  
 group\_by(lemword) %>%  
 filter(n() >= 5) %>%  
 pairwise\_cor(lemword, section, sort = TRUE)  
  
## add colours for correlation  
word\_cors$group<-ifelse(word\_cors$correlation>0.2,"high","low")  
  
  
# Now look at clusters   
set.seed(2016)  
  
#Make two datasets to use for this:   
# trying to make a count column to have numbered nodes  
counts <- dplyr::count(df\_token, lemword, sort = TRUE)  
   
# Join counts df to my word\_cors df  
attempteddf<-left\_join(word\_cors, counts,by=c(item1="lemword"))  
  
# Create a modification where you can create the two dataframes for edge and vertex information separately, which should give you the control you might need.  
#Make edges  
edge\_data <- attempteddf %>%   
 filter(correlation > 0.1 ) %>%  
 select(item1, item2, correlation,group)  
  
# Make nodes  
node\_data <- attempteddf %>%   
 select(item1, n) %>%  
 # Using distinct to keep only the first row when word1 is duplicated to avoid error below  
 distinct(item1, .keep\_all = TRUE)  
#Make graph  
g <- graph\_from\_data\_frame(edge\_data, vertices = node\_data)  
# Error in graph\_from\_data\_frame(edge\_data, directed = TRUE, vertices = node\_data) :   
  
# Duplicate vertex names  
g\_low<-g %>%  
 ggraph(layout = "fr") +  
 geom\_edge\_link(aes(edge\_alpha = correlation, edge\_width = correlation)) +  
 # Now the number column is accessible!  
 geom\_node\_point(aes(size = n))+  
 geom\_node\_text(aes(label = name, size=15), repel = TRUE)+  
 # point.padding = unit(0.4, "lines"))+  
 scale\_size(range=c(4,12),guide = "none")  
  
  
  
edge\_data <- attempteddf %>%   
 filter(correlation > 0.1) %>%  
 select(item1, item2, correlation)  
node\_data <- attempteddf %>%   
 select(item1, n) %>%  
 # Using distinct to keep only the first row when word1 is duplicated to avoid error below  
 distinct(item1, .keep\_all = TRUE)  
  
g <- graph\_from\_data\_frame(edge\_data, vertices = node\_data)   
  
g1<-g %>%  
 ggraph(layout = "fr") +  
 geom\_edge\_link(aes(edge\_alpha = correlation, edge\_width = correlation), edge\_colour = "lightgrey") +  
 # Now the number column is accessible!  
 geom\_node\_point(aes(size = n))+  
 geom\_node\_text(aes(label = name), repel = TRUE)+  
 # point.padding = unit(0.4, "lines"))+  
 scale\_size(range=c(5,8),guide = "none")+  
 theme\_graph()  
  
  
# high res plot  
pdf("socialnetwork.pdf")  
g1 # Make plot  
dev.off()  
  
  
g1