Widyr

## R Markdown

1.Load the packages tidytext and tidyverse

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.0 v dplyr 1.0.5  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(tidytext)

2.Read-in the following datasets in R (the dataset is available on Canvas).

#read in the data  
tweets <- read\_csv("Twitter Data-Corona.csv")

## Warning: Missing column names filled in: 'X1' [1]

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## X1 = col\_double(),  
## nonrts.created\_at = col\_character(),  
## nonrts.place.country = col\_character(),  
## nonrts.lang = col\_character(),  
## nonrts.place.country.1 = col\_character(),  
## nonrts.text = col\_character(),  
## nonrts.user.id\_str = col\_double(),  
## nonrts.user.screen\_name = col\_character(),  
## nonrts.user.verified = col\_logical(),  
## nonrts.user.friends\_count = col\_double(),  
## nonrts.user.followers\_count = col\_double(),  
## nonrts.reply\_count = col\_double(),  
## nonrts.retweet\_count = col\_double(),  
## nonrts.favorite\_count = col\_double()  
## )

tweets <- tweets %>%  
 filter(nonrts.lang=="en")%>%  
 select (created\_at=nonrts.created\_at,text=nonrts.text)  
tweets

## # A tibble: 964 x 2  
## created\_at text   
## <chr> <chr>   
## 1 Fri Mar 27 00:48:56 +00~ "At least he\x92s being honest <U+0001F937><U+0001F~  
## 2 Fri Mar 27 00:48:56 +00~ "@SEPTA YALL GOT THE CORONA?"   
## 3 Fri Mar 27 00:48:56 +00~ "Greedy ass America smh this should be the least of~  
## 4 Fri Mar 27 00:48:56 +00~ "@RealCandaceO You silly bitch, those are all estim~  
## 5 Fri Mar 27 00:48:56 +00~ "Sharp rise in number of calls to ChildLine over co~  
## 6 Fri Mar 27 00:48:57 +00~ "i do not fucking care, corona bitch, i need sex."   
## 7 Fri Mar 27 00:48:57 +00~ "@callme\_anqiiex3 bra look at this <U+0001F602><U+~  
## 8 Fri Mar 27 00:48:57 +00~ "@LWN\_ @RepDanCrenshaw Using deflection as a means ~  
## 9 Fri Mar 27 00:48:57 +00~ "Coronavirus | G20 commits $5 trillion amid COVID-1~  
## 10 Fri Mar 27 00:48:57 +00~ "Make sure to check our website for news and update~  
## # ... with 954 more rows

1. Create a tidytext dataset of Tweets - Tokenize by unigrams

tidy <- tweets %>%  
 unnest\_tokens("word", text)

1. Pre-Process Text by Removing numbers, whitespaces, special characters, undesirable words, and stop words

undesirable\_words <-c("https","t.co","amp","rt","tco","coronavirus","corona","covid","virus")  
  
tidy\_filtered <- tidy %>%  
 anti\_join(stop\_words)%>% #remove the stop words using the lexicon called stop\_word  
 filter (!word %in% undesirable\_words) %>% #remove the undesirable words using dplyr's filter() function   
 filter(nchar(word) > 3) %>% #remove words with less than 3 characters using dplyr's filter() verb  
 filter(   
 !str\_detect(word, "^\\b\\d+\\b"), #remove numbers  
 !str\_detect(word, "\\s+"), #remove white spaces  
 !str\_detect(word, "[^a-zA-Z]")) #remove special characters

## Joining, by = "word"

1. Count the most frequently used words.

tidy\_counts <- tidy\_filtered %>%  
 count(word, sort = TRUE)  
tidy\_counts

## # A tibble: 2,966 x 2  
## word n  
## <chr> <int>  
## 1 world 47  
## 2 trump 46  
## 3 people 44  
## 4 china 39  
## 5 pandemic 26  
## 6 news 24  
## 7 confirmed 23  
## 8 care 17  
## 9 country 17  
## 10 bill 16  
## # ... with 2,956 more rows

For all the talk of a world pandemic, “Trump” shows up the second most times, indicating once again that tweets can be more of an American-centric pop culture measuring device rather than a source for facts. “World”, “people”, “China” and “pandemic” are other popular terms.

1. Load the package widyr.

library(widyr)

## Warning: package 'widyr' was built under R version 4.0.5

1. Use pairwise\_count() from the widyr package to count how many times each pair of words occurs together in the tweets. Explain the results.{it may take a few minutes to run this chunck of code}

word\_pairs <- tidy\_filtered %>%  
 pairwise\_count(word, created\_at, sort = TRUE)

## Warning: `distinct\_()` was deprecated in dplyr 0.7.0.  
## Please use `distinct()` instead.  
## See vignette('programming') for more help

## Warning: `tbl\_df()` was deprecated in dplyr 1.0.0.  
## Please use `tibble::as\_tibble()` instead.

word\_pairs

## # A tibble: 181,422 x 3  
## item1 item2 n  
## <chr> <chr> <dbl>  
## 1 world confirmed 13  
## 2 people trump 13  
## 3 confirmed world 13  
## 4 trump people 13  
## 5 china world 12  
## 6 world china 12  
## 7 trump pandemic 10  
## 8 people pandemic 10  
## 9 pandemic trump 10  
## 10 pandemic people 10  
## # ... with 181,412 more rows

The words that are commonly found together are combinations of “world”, “confirmed”, “trump”, “people”, “china” and “pandemic” . The first 14 rows are some combination of that.

1. Use pairwise\_cor() from the widyr package to calculate the correlation among words in the tweets (filter for at least relatively common words first n() >= 10).Explain the results.{it may take a few minutes to run this chunck of code}

word\_cors <- tidy\_filtered %>%  
 group\_by(word) %>%  
 filter(n() >= 10) %>%  
 pairwise\_cor(word, created\_at, sort = TRUE)  
  
word\_cors

## # A tibble: 1,482 x 3  
## item1 item2 correlation  
## <chr> <chr> <dbl>  
## 1 home stay 0.695  
## 2 stay home 0.695  
## 3 positive tested 0.539  
## 4 tested positive 0.539  
## 5 testing kids 0.490  
## 6 kids testing 0.490  
## 7 italy china 0.423  
## 8 china italy 0.423  
## 9 kids care 0.369  
## 10 care kids 0.369  
## # ... with 1,472 more rows

“pairwise\_cor” goes through our data and determines the number of times each word occurs with every other word. The higher the number, the higher the proportion that they occur together. “Stay” and “home” were found together 69% of the time, with “tested” and “positive” next at 54%. It looks like people are interested in their children since “kids/testing” came up at 495 and “Kids/care” came in at 37%

1. load the packages igraph and ggraph

library (igraph)

## Warning: package 'igraph' was built under R version 4.0.5

##   
## Attaching package: 'igraph'

## The following objects are masked from 'package:dplyr':  
##   
## as\_data\_frame, groups, union

## The following objects are masked from 'package:purrr':  
##   
## compose, simplify

## The following object is masked from 'package:tidyr':  
##   
## crossing

## The following object is masked from 'package:tibble':  
##   
## as\_data\_frame

## The following objects are masked from 'package:stats':  
##   
## decompose, spectrum

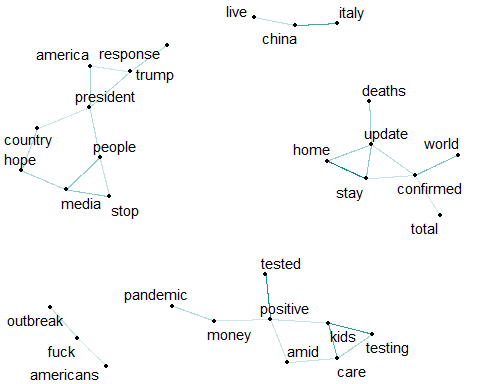
## The following object is masked from 'package:base':  
##   
## union

library (ggraph)

## Warning: package 'ggraph' was built under R version 4.0.5

1. Plot networks of these correlations among words (use the ggraph package and the layout=“fr”")).Explain the results.

set.seed(1234)  
   
word\_cors %>%  
 filter(correlation > .20) %>%  
 graph\_from\_data\_frame() %>%  
 ggraph(layout = "fr") +  
 geom\_edge\_link(aes(edge\_alpha = correlation), edge\_colour = "cyan4", show.legend = FALSE) +  
 geom\_node\_point( size = 1) +  
 geom\_node\_text(aes(label = name), repel = TRUE) +  
 theme\_void()



There is definately some word clustering going on. The largest cluster seems to be concerned with America and their response, with an obligatory mention of the media. A smaller cluster has some detectable negative feelings toward Americans. Another cluster connects testing, kids, money, and care together, dealing with the actual realities of the pandemic. The final cluster pertains to general information about staying home and measures about hte pandemic.

1. Reflect on this assignment and provide a summary of your findings.

When using our toolbox to analyze the tweets one word at a time, we can see a clustering of thoughts and ideas about the pandemic according to the tweets provided. The tweets do seem more skewed toward American politics and our social condition, which is sad because the pandemic is actually about diseases and not making political points or getting liked for your comments.