Module 6 - Extra Credit Assignment

## R Markdown

1.Load the packages tidytext and tidyverse

library(tidyverse)  
library(tidytext)

1. 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. Describe the variables in the data set and number of observations.

There are two variables. One is the date/timestamp of when the comment was made. The second is the comment itself, filled with insightful tidbits. After filtering to just use English, there are 964 observations.

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

bigrams <- tweets %>%  
 unnest\_tokens(bigram, text, token = "ngrams", n = 2)  
bigrams

## # A tibble: 15,017 x 2  
## created\_at bigram   
## <chr> <chr>   
## 1 Fri Mar 27 00:48:56 +0000 2020 at least   
## 2 Fri Mar 27 00:48:56 +0000 2020 least he   
## 3 Fri Mar 27 00:48:56 +0000 2020 he s   
## 4 Fri Mar 27 00:48:56 +0000 2020 s being   
## 5 Fri Mar 27 00:48:56 +0000 2020 being honest  
## 6 Fri Mar 27 00:48:56 +0000 2020 honest u   
## 7 Fri Mar 27 00:48:56 +0000 2020 u 0001f937   
## 8 Fri Mar 27 00:48:56 +0000 2020 0001f937 u   
## 9 Fri Mar 27 00:48:56 +0000 2020 u 0001f3fd   
## 10 Fri Mar 27 00:48:56 +0000 2020 0001f3fd u   
## # ... with 15,007 more rows

1. Filter stop-words, undesirable words, and words less than 3 charcters.

undesirable\_words <-c("https","t.co","amp","rt","tco","coronavirus","corona","covid","virus")  
   
bigrams\_separated <- bigrams %>%  
 separate(bigram, c("word1", "word2"), sep = " ")   
  
  
bigrams\_filtered <- bigrams\_separated %>%  
 filter(!word1 %in% stop\_words$word,  
 !word2 %in% stop\_words$word )%>%   
 filter(!word1 %in% undesirable\_words,  
 !word2 %in% undesirable\_words)%>%   
 filter(!nchar(word1) < 3,  
 !nchar(word2) < 3)

Count the most common bigrams.

There aren’t a lot of big hitters, but many of the bigrams are two-word combos you would expect to see together. “Tested positive”, “begin testing”, “donald Trump”, and “task force” are you’d expect to see together. Being the father of teenagers, I know what “animal crossing” is, and “Mark Blum” was one of the famous early deaths due to coronavirus.

bigram\_counts <- bigrams\_filtered %>%   
 count(word1, word2, sort = TRUE)  
  
bigram\_counts

## # A tibble: 2,059 x 3  
## word1 word2 n  
## <chr> <chr> <int>  
## 1 tested positive 5  
## 2 animal crossing 4  
## 3 begin testing 4  
## 4 business insider 4  
## 5 china reports 4  
## 6 donald trump 4  
## 7 mark blum 4  
## 8 president donald 4  
## 9 release urgent 4  
## 10 task force 4  
## # ... with 2,049 more rows

1. Load the package igraph:

library(igraph)

1. Use the output from step 5 and build a network of common bigrams [filter for only relatively common combinations (based on n - (n>2) – use lines instead of directed arrows between nodes (graph\_from\_data\_frame (directed = FALSE))].

# filter for only relatively common combinations  
   
set.seed(1234)  
bigram\_graph <- bigram\_counts %>%  
 filter(n >= 2) %>%  
 graph\_from\_data\_frame(directed = FALSE)

1. Load the package ggraph

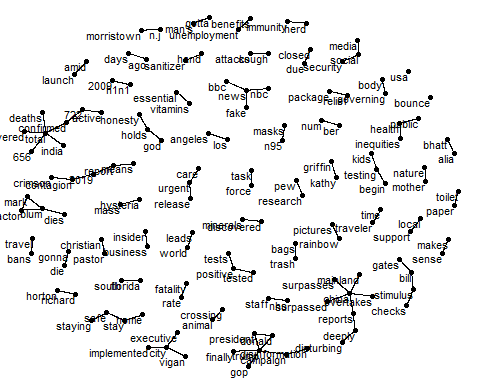
library(ggraph)

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

1. Visualize the graph - Use the Fruchterman-Reingold to visualize the nodes and ties (“fr”). Explain the results.

set.seed(2017)  
  
ggraph(bigram\_graph, layout = "fr") +  
 geom\_edge\_link() +  
 geom\_node\_point() +  
 geom\_node\_text(aes(label = name), size=3, point.padding=unit(0.2,"lines"), vjust = 1, hjust = 1)+  
 theme\_void()

## Warning: Ignoring unknown parameters: point.padding



1. Provide a summary of your findings.

The contents of the tweets appear to be all over the place. There are few networks with more than 4 nodes. Looking at the networks, they do look like the contents of Twitter, with “Donald Trump” lumped in with “disinformation” and non-sequitors like “Kathy Griffin” and “rainbow pictures” making an appearance. There is not a lot of structure here, which is kind of the way social discourse goes these days.