James Bloor

Boston University Metropolitan College

MET CS 688

Assignment #6

Fall 1 2021

BU ID: U45448564

**PART 1. (The below tasks add to 65 points.) Twitter social network visualization.**

1. **(4 points) Browse Twitter to find at least four user accounts that you believe will have some common people in their networks. For example, you could choose the U.S. President and Vice President for two of your accounts (but, remember, you need four). These accounts could be, for example, @JoeBiden and @KamalaHarris. You don’t have to choose politicians. You could choose people or organizations on any topic you like.  
     
   Load the list of accounts that each of your (at least) four target accounts are following. If using R,  
     
   library(rtweet)  
   following <- get\_friends(c("JoeBiden", "KamalaHarris", "anotherone", "fourthone")**



1. **(0 points) Eliminate any names in “following” that show up only once. (You may optionally also eliminate any names in “following” that show up twice.) These accounts are being followed by only one of the target accounts (maybe there are some people that only @JoeBiden follows, that is, @KamalaHarris and the others you select are not following). If using R,  
     
   counts <- table(following$user\_id)  
   following.reduced <- following %>% filter(user\_id %in% names(counts[counts > 1]))  
    # you might use > 2 here to reduce the output  
     
   You may also notice rtweet’s get\_friends() function returns Twitter user numbers instead of Twitter screen names. If using R, convert the numbers to screen names by  
     
   following.names <- left\_join(following.reduced, distinct(following.reduced,user\_id) %>%  
    mutate(names = lookup\_users(user\_id)$screen\_name), by = 'user\_id') %>% select(-user\_id)  
   following.matrix <- as.matrix(following.names)**

Text

Description automatically generated

1. **(25 points) Make both a directed and an undirected graph. Color your four (or more) target accounts in a different color from the friends that are found.  
     
   You may find the graph very hard to read as it may have too many vertices. For readability purposes, you could look into using the induced\_subgraph function (R) or .subgraph() method (Python). In either case, you should include all colored vertices and a random selection of the un-colored vertices. In this case, provide a small version of the full graph and a larger, readable version of the subgraph.**

Graphical user interface, text

Description automatically generated

Diagram

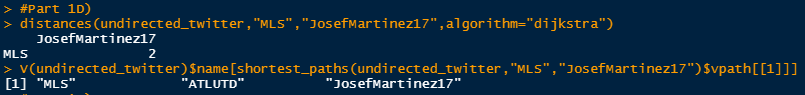
Description automatically generated

Diagram

Description automatically generated

1. **(12 points) Select any two vertices on your *undirected* graph that are more than one hop apart. Write and execute code to apply the Dijkstra algorithm to find the shortest path between them.**





1. **(24 points) Write and execute code to get centrality measurements on your graph. Run this on both your *undirected* and *directed* graphs. You could choose to use the subgraphs instead of the full graphs. If you do, explain the benefits (if any) and the risks (if any) associated with doing so. (Note: if the benefits are not greater than the risks, you should not be running centrality on the subgraphs!)  
     
   Based on what you’ve learned and what you can find out with further research, explain why you see the differences you do in the two results (undirected vs. directed).   
     
   Also look to see whether there are differences in the centrality for the target accounts you selected (in part a) and the centrality for the accounts found by getting the list of accounts the targets are following (in part b). Explain why there are (or are not) differences.  
     
   Please don’t output the full tables of centrality measurements (for most of you this will be too many pages). In the main part of your report you should only share portions of the tables that highlight critical differences. You could share code that would output the full tables in your appendix.**

Text

Description automatically generated

Text

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

In the two outputs we see above, there is difference between the directed and undirected centralities. The reason for this because of the nodes that are reachable from a given point. Centrality is measuring importance. On a directed graph we see often see lowest centralities. This is because an edge in the graphs I made above are only one directed. In the undirected graph we can move in any direction we want which gives us flexibility on moving from node to node. Therefore, we see a different in the centrality measurements above.

Looking at the outputs above there are few items to point out. First is that JosefMartinez17 is the least important of our original twitter accounts. He doesn’t have much overlay with our other three accounts, as we can see in the graphs, and these affects him in our statistics. There are two accounts that are standout among the accounts we have here that are not a part of our four original accounts, the Braves and DEalesATLUTD. This is not a surprise; the Braves are obviously a prominent team here in Atlanta and I chose four related Atlanta United accounts to pull data from. DEalesATLUTD is an account that belongs to the president of Atlanta United to see him in the mix here is expected.

**PART 2. (The below tasks add to 35 points.) Bigram text mining analysis. Recall that in our lecture discussion in Module 3, we showed how to extract bigrams using the token="ngrams" and n=2 options in tidytext::unnest\_tokens. Some simple web searches will also show you how to use nltk.bigrams in Python. You will need to be able to extract the bigrams to create this graph.**

1. **(5 points) Acquire a set of text. I recommend a set of tweets on a search of your choice but you may be able to find a worthwhile text data set elsewhere.**

Graphical user interface

Description automatically generated with medium confidence

1. **(0 points) Write the vector of text to a file and upload it with the rest of your code. Provide code in your assignment that reads this file back in. (You should have code that does something like this from Assignment 3 that you can reapply here.)**
   1. **The purpose for this is to allow your facilitator to grade your assignment using the same text that you had (recall that online-sourced data can change after you write your code but before your code is graded).**

Text

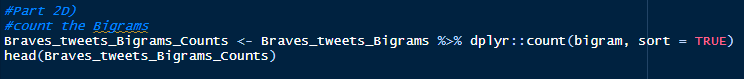
Description automatically generated

1. **(10 points) Create a tidytext or nltk object with bigrams. As you are doing this, you should also preprocess the text. Add a step to the rest of your preprocessing that removes any words that are part of your search phrase.**

Text

Description automatically generated

1. **(5 points) Get a count of how many times each bigram appears in the processed text.**



Text

Description automatically generated

1. **(15 points) Create and plot an undirected weighted graph of the word pairs. For each bigram that you have, you should get an edge that connects the two words. Set the weight for the edge equal to the number of times the bigram appears. Ensure your plot does something to show the weights. Comment on any patterns that appear.**

Text

Description automatically generated

Scatter chart

Description automatically generated with medium confidence

For my twitter search I looked at tweets that included “Braves”. I did filter down the graph to only show bigrams that occurede more than twice, so it was more readable and searchable to find trends in the data. The Atlanta Braves are currenting in the hunt for the MLB World Series and I though that it would be interesting to see what patters we could pull. The largest weight for a bigram we can see is “astro redsox” this is not surprising because when I pulled this data these were currently playing in Game 3 of their series. In these rest of this graph, there are twitter names that I do not recognize but may be players that were participating in the game being played. There is also a graph that sees current teams having playoff struggles. While there is no specific data ono what team is being spoken about, people love to complain about teams on twitter. “red sox” & “blue jay” are teams in the MLB. Eddie Rosario hit the walk off hit in game two to have the Braves win the bottom of the ninth inning, so I am assuming that is why he is being spoke about with sports center. Lastly, “battleatl chopon” are common terms using as hashtags that relate closely to the Braves.

**R CODE:**

> ### --- Bloor\_CS688\_Assignment\_6 --------

> #load libs needed

> library(rtweet)

> library(caret)

> library(tidyverse)

> library(tm)

> library(stringdist)

> library(proxy)

> library(RecordLinkage)

> library(proxy)

> library(wordcloud)

> library(cluster)

> library(stringi)

> library(dendextend)

> library(SnowballC)

> library(textstem)

> library(clusterCrit)

> library("ape")

> library(quanteda)

> library(ggplot2)

> library(plotly)

> library(igraph)

> library(textstem)

> library(dplyr)

> library(tidytext)

>

> #set WD

> setwd("C:\\Users\\James Bloor\\Desktop\\BU\\CS688\\Week 6")

>

> #Part 1A)

> following <- get\_friends(c("ATLUTD", "JosefMartinez17", "MFparkhurst", "atlutdpup"))

1 friend networks collected!

2 friend networks collected!

3 friend networks collected!

4 friend networks collected!

>

> #Part 1B)

> counts <- table(following$user\_id)

> following.reduced <- following %>% filter(user\_id %in% names(counts[counts > 2]))

> table(following.reduced$user\_id)

107146095 1212411541 131540058 1371407737 1374939846

3 3 3 3 3

18091004 197434474 21436663 2177497055 2183443432

3 3 4 3 3

222387125 2253070123 2873124768 2902993470 298172574

3 3 3 4 3

3255162527 326548824 36448925 381182424 41364702

3 3 3 3 3

4906200124 518715968 567708081 601226235 616595504

3 3 3 3 3

62542252 70845171 717741741906898944 874705218533019649 928258614548156417

3 3 3 3 3

> # you might use > 2 here to reduce the output

>

> #You may also notice rtweet's get\_friends() function returns Twitter user numbers

> #instead of Twitter screen names. If using R, convert the numbers to screen names by

>

> following.names <- left\_join(following.reduced, distinct(following.reduced,user\_id) %>%

+ mutate(names = lookup\_users(user\_id)$screen\_name), by = 'user\_id') %>% select(-user\_id)

> following.matrix <- as.matrix(following.names)

> (following.matrix)

user names

[1,] "ATLUTD" "CholMachop"

[2,] "ATLUTD" "laurencewyke"

[3,] "ATLUTD" "Henryapaloo"

[4,] "ATLUTD" "GeorgeC1927"

[5,] "ATLUTD" "AdamJahn14"

[6,] "ATLUTD" "jackcollison"

[7,] "ATLUTD" "\_moadams"

[8,] "ATLUTD" "Brendan\_Moore24"

[9,] "ATLUTD" "Eric\_Remedi"

[10,] "ATLUTD" "NBCSportsSoccer"

[11,] "ATLUTD" "JillianSakovits"

[12,] "ATLUTD" "atlutd2"

[13,] "ATLUTD" "francoeescobar"

[14,] "ATLUTD" "atlantahumane"

[15,] "ATLUTD" "bguzan"

[16,] "ATLUTD" "\_georgebello"

[17,] "ATLUTD" "kev\_egan"

[18,] "ATLUTD" "JosefMartinez17"

[19,] "ATLUTD" "AntonWalkes"

[20,] "ATLUTD" "\_milesrobinson\_"

[21,] "ATLUTD" "markbloom21"

[22,] "ATLUTD" "mikeyambrose3"

[23,] "ATLUTD" "TitoVillalba15"

[24,] "ATLUTD" "AcademyATLUTD"

[25,] "ATLUTD" "TonyAnnan1"

[26,] "ATLUTD" "Braves"

[27,] "ATLUTD" "BocaBoca3"

[28,] "ATLUTD" "DEalesATLUTD"

[29,] "ATLUTD" "ESPNFC"

[30,] "ATLUTD" "MLS"

[31,] "JosefMartinez17" "markbloom21"

[32,] "JosefMartinez17" "Braves"

[33,] "JosefMartinez17" "DEalesATLUTD"

[34,] "JosefMartinez17" "TitoVillalba15"

[35,] "MFparkhurst" "jackcollison"

[36,] "MFparkhurst" "ESPNFC"

[37,] "MFparkhurst" "AcademyATLUTD"

[38,] "MFparkhurst" "NBCSportsSoccer"

[39,] "MFparkhurst" "CholMachop"

[40,] "MFparkhurst" "Henryapaloo"

[41,] "MFparkhurst" "\_moadams"

[42,] "MFparkhurst" "Braves"

[43,] "MFparkhurst" "kev\_egan"

[44,] "MFparkhurst" "Brendan\_Moore24"

[45,] "MFparkhurst" "JillianSakovits"

[46,] "MFparkhurst" "GeorgeC1927"

[47,] "MFparkhurst" "laurencewyke"

[48,] "MFparkhurst" "\_georgebello"

[49,] "MFparkhurst" "Eric\_Remedi"

[50,] "MFparkhurst" "francoeescobar"

[51,] "MFparkhurst" "atlantahumane"

[52,] "MFparkhurst" "atlutd2"

[53,] "MFparkhurst" "JosefMartinez17"

[54,] "MFparkhurst" "TitoVillalba15"

[55,] "MFparkhurst" "AntonWalkes"

[56,] "MFparkhurst" "TonyAnnan1"

[57,] "MFparkhurst" "markbloom21"

[58,] "MFparkhurst" "\_milesrobinson\_"

[59,] "MFparkhurst" "mikeyambrose3"

[60,] "MFparkhurst" "DEalesATLUTD"

[61,] "MFparkhurst" "AdamJahn14"

[62,] "MFparkhurst" "BocaBoca3"

[63,] "MFparkhurst" "bguzan"

[64,] "MFparkhurst" "MLS"

[65,] "atlutdpup" "GeorgeC1927"

[66,] "atlutdpup" "CholMachop"

[67,] "atlutdpup" "kev\_egan"

[68,] "atlutdpup" "mikeyambrose3"

[69,] "atlutdpup" "laurencewyke"

[70,] "atlutdpup" "JillianSakovits"

[71,] "atlutdpup" "atlantahumane"

[72,] "atlutdpup" "Henryapaloo"

[73,] "atlutdpup" "jackcollison"

[74,] "atlutdpup" "TonyAnnan1"

[75,] "atlutdpup" "Braves"

[76,] "atlutdpup" "AdamJahn14"

[77,] "atlutdpup" "\_moadams"

[78,] "atlutdpup" "BocaBoca3"

[79,] "atlutdpup" "AntonWalkes"

[80,] "atlutdpup" "\_georgebello"

[81,] "atlutdpup" "francoeescobar"

[82,] "atlutdpup" "Brendan\_Moore24"

[83,] "atlutdpup" "ESPNFC"

[84,] "atlutdpup" "Eric\_Remedi"

[85,] "atlutdpup" "NBCSportsSoccer"

[86,] "atlutdpup" "AcademyATLUTD"

[87,] "atlutdpup" "atlutd2"

[88,] "atlutdpup" "\_milesrobinson\_"

[89,] "atlutdpup" "bguzan"

[90,] "atlutdpup" "DEalesATLUTD"

[91,] "atlutdpup" "JosefMartinez17"

[92,] "atlutdpup" "MLS"

>

>

> #Part 1C)

> friends <- c("ATLUTD", "JosefMartinez17", "MFparkhurst", "atlutdpup")

> undirected\_twitter <- graph\_from\_data\_frame(following.matrix, directed = F)

> V(undirected\_twitter)[friends]$color = 'red'

> plot(undirected\_twitter)

>

>

> directed\_twitter <- graph\_from\_data\_frame(following.matrix, directed = T)

> V(directed\_twitter)[friends]$color = 'red'

> plot(directed\_twitter)

>

>

> #Part 1D)

> distances(undirected\_twitter,"MLS","JosefMartinez17",algorithm="dijkstra")

JosefMartinez17

MLS 2

> V(undirected\_twitter)$name[shortest\_paths(undirected\_twitter,"MLS","JosefMartinez17")$vpath[[1]]]

[1] "MLS" "ATLUTD" "JosefMartinez17"

> #Part 1E)

> #undirected

> undirected\_centrality <- data.frame(

+ name=V(undirected\_twitter)$name,

+ degree=degree(undirected\_twitter,normalized=T),

+ closeness=closeness(undirected\_twitter,normalized=T),

+ betweenness=betweenness(undirected\_twitter,normalized=T),

+ eigen=eigen\_centrality(undirected\_twitter,scale=F)$vector

+ )

> undirected\_centrality

name degree closeness betweenness eigen

ATLUTD ATLUTD 0.93750 0.9411765 0.3062836022 0.4115981

JosefMartinez17 JosefMartinez17 0.21875 0.5614035 0.0081077189 0.1817970

MFparkhurst MFparkhurst 0.93750 0.9411765 0.3062836022 0.4115981

atlutdpup atlutdpup 0.87500 0.8888889 0.2525201613 0.3891329

CholMachop CholMachop 0.09375 0.5245902 0.0002112135 0.1281680

laurencewyke laurencewyke 0.09375 0.5245902 0.0002112135 0.1281680

Henryapaloo Henryapaloo 0.09375 0.5245902 0.0002112135 0.1281680

GeorgeC1927 GeorgeC1927 0.09375 0.5245902 0.0002112135 0.1281680

AdamJahn14 AdamJahn14 0.09375 0.5245902 0.0002112135 0.1281680

jackcollison jackcollison 0.09375 0.5245902 0.0002112135 0.1281680

\_moadams \_moadams 0.09375 0.5245902 0.0002112135 0.1281680

Brendan\_Moore24 Brendan\_Moore24 0.09375 0.5245902 0.0002112135 0.1281680

Eric\_Remedi Eric\_Remedi 0.09375 0.5245902 0.0002112135 0.1281680

NBCSportsSoccer NBCSportsSoccer 0.09375 0.5245902 0.0002112135 0.1281680

JillianSakovits JillianSakovits 0.09375 0.5245902 0.0002112135 0.1281680

atlutd2 atlutd2 0.09375 0.5245902 0.0002112135 0.1281680

francoeescobar francoeescobar 0.09375 0.5245902 0.0002112135 0.1281680

atlantahumane atlantahumane 0.09375 0.5245902 0.0002112135 0.1281680

bguzan bguzan 0.09375 0.5245902 0.0002112135 0.1281680

\_georgebello \_georgebello 0.09375 0.5245902 0.0002112135 0.1281680

kev\_egan kev\_egan 0.09375 0.5245902 0.0002112135 0.1281680

AntonWalkes AntonWalkes 0.09375 0.5245902 0.0002112135 0.1281680

\_milesrobinson\_ \_milesrobinson\_ 0.09375 0.5245902 0.0002112135 0.1281680

markbloom21 markbloom21 0.09375 0.5245902 0.0000672043 0.1062483

mikeyambrose3 mikeyambrose3 0.09375 0.5245902 0.0002112135 0.1281680

TitoVillalba15 TitoVillalba15 0.09375 0.5245902 0.0000672043 0.1062483

AcademyATLUTD AcademyATLUTD 0.09375 0.5245902 0.0002112135 0.1281680

TonyAnnan1 TonyAnnan1 0.09375 0.5245902 0.0002112135 0.1281680

Braves Braves 0.12500 0.5333333 0.0002112135 0.1473876

BocaBoca3 BocaBoca3 0.09375 0.5245902 0.0002112135 0.1281680

DEalesATLUTD DEalesATLUTD 0.12500 0.5333333 0.0002112135 0.1473876

ESPNFC ESPNFC 0.09375 0.5245902 0.0002112135 0.1281680

MLS MLS 0.09375 0.5245902 0.0002112135 0.1281680

>

> #directed

> directed\_centrality <- data.frame(

+ name=V(directed\_twitter)$name,

+ degree=degree(directed\_twitter,normalized=T),

+ closeness=closeness(directed\_twitter,normalized=T),

+ betweenness=betweenness(directed\_twitter,normalized=T),

+ eigen=eigen\_centrality(directed\_twitter,scale=F)$vector

+ )

Warning message:

In closeness(directed\_twitter, normalized = T) :

At centrality.c:2874 :closeness centrality is not well-defined for disconnected graphs

> directed\_centrality

name degree closeness betweenness eigen

ATLUTD ATLUTD 0.93750 0.33333333 0.000000000 0.4115981

JosefMartinez17 JosefMartinez17 0.21875 0.03448276 0.002016129 0.1817970

MFparkhurst MFparkhurst 0.93750 0.33333333 0.000000000 0.4115981

atlutdpup atlutdpup 0.87500 0.32653061 0.000000000 0.3891329

CholMachop CholMachop 0.09375 0.03030303 0.000000000 0.1281680

laurencewyke laurencewyke 0.09375 0.03030303 0.000000000 0.1281680

Henryapaloo Henryapaloo 0.09375 0.03030303 0.000000000 0.1281680

GeorgeC1927 GeorgeC1927 0.09375 0.03030303 0.000000000 0.1281680

AdamJahn14 AdamJahn14 0.09375 0.03030303 0.000000000 0.1281680

jackcollison jackcollison 0.09375 0.03030303 0.000000000 0.1281680

\_moadams \_moadams 0.09375 0.03030303 0.000000000 0.1281680

Brendan\_Moore24 Brendan\_Moore24 0.09375 0.03030303 0.000000000 0.1281680

Eric\_Remedi Eric\_Remedi 0.09375 0.03030303 0.000000000 0.1281680

NBCSportsSoccer NBCSportsSoccer 0.09375 0.03030303 0.000000000 0.1281680

JillianSakovits JillianSakovits 0.09375 0.03030303 0.000000000 0.1281680

atlutd2 atlutd2 0.09375 0.03030303 0.000000000 0.1281680

francoeescobar francoeescobar 0.09375 0.03030303 0.000000000 0.1281680

atlantahumane atlantahumane 0.09375 0.03030303 0.000000000 0.1281680

bguzan bguzan 0.09375 0.03030303 0.000000000 0.1281680

\_georgebello \_georgebello 0.09375 0.03030303 0.000000000 0.1281680

kev\_egan kev\_egan 0.09375 0.03030303 0.000000000 0.1281680

AntonWalkes AntonWalkes 0.09375 0.03030303 0.000000000 0.1281680

\_milesrobinson\_ \_milesrobinson\_ 0.09375 0.03030303 0.000000000 0.1281680

markbloom21 markbloom21 0.09375 0.03030303 0.000000000 0.1062483

mikeyambrose3 mikeyambrose3 0.09375 0.03030303 0.000000000 0.1281680

TitoVillalba15 TitoVillalba15 0.09375 0.03030303 0.000000000 0.1062483

AcademyATLUTD AcademyATLUTD 0.09375 0.03030303 0.000000000 0.1281680

TonyAnnan1 TonyAnnan1 0.09375 0.03030303 0.000000000 0.1281680

Braves Braves 0.12500 0.03030303 0.000000000 0.1473876

BocaBoca3 BocaBoca3 0.09375 0.03030303 0.000000000 0.1281680

DEalesATLUTD DEalesATLUTD 0.12500 0.03030303 0.000000000 0.1473876

ESPNFC ESPNFC 0.09375 0.03030303 0.000000000 0.1281680

MLS MLS 0.09375 0.03030303 0.000000000 0.1281680

> #need to answer rest of question in Word Doc

>

> #Part 2A)

> Braves\_tweets <- search\_tweets("Braves", n =50, include\_rts = FALSE,lang = "en")

> Braves\_tweets <- (Braves\_tweets$text)

>

> #Part 2B)

> write.csv(Braves\_tweets, file="Braves\_tweets.csv")

> Braves\_tweets <- read.csv("Braves\_tweets.csv",header=T)

> Braves\_tweets <- Braves\_tweets[c(-1)]

> #Part 2C)

> Braves\_tweets <- as.data.frame(Braves\_tweets)

>

> #Pre-process titles we have

> Braves\_tweetsScorp <- VCorpus(VectorSource(Braves\_tweets))

>

>

> #lowercase

> Braves\_tweetsScorp <- tm\_map(Braves\_tweetsScorp, content\_transformer(tolower))

> head(Braves\_tweetsScorp[[1]]$content)

[1] "@jolly\_olive john smoltz might finally get over 1996 and 1999 if the braves win it this year."

[2] "@jjw4jld then they'd be the nl toronto blue jays ...\na+ offense with average (at best) pitching. \n\ngotta say, though ... the pitching has\nbeen excellent (so far <u+0001f91e>) against\nthe brewers &amp; the dodgers!\n\nblue jays play in the toughest division ...\nthe braves play in the weakest division. <u+0001f937><u+200d><u+2642><u+fe0f>"

[3] "@dalepla42824047 @frankqu55904875 @zwarrenz @realtoddkalas @astros @redsox @braves both teams were 3 games above .500 against nl playoff teams."

[4] "@dalepla42824047 @frankqu55904875 @zwarrenz @realtoddkalas @astros @redsox @braves and, unbalanced sched had braves playing al east (4 teams &gt; 90 wins) while la played al west (1 team over 90). braves 11 games against playoff teams in al, la 4.\ndodgers built a massive record off rockies and dbacks. (as well as pirates, mets, &amp; nats post trades)"

[5] "@dalepla42824047 @frankqu55904875 @zwarrenz @realtoddkalas @astros @redsox @braves not really\nwhen the unbalanced schedule is taken into effect &amp; records after trade deadline, the current teams are pretty equal.\nafter deadline was a 5 gm adv for la. imo that is most telling &amp; shows 18 game differential doesn’t show how close these two current teams are."

[6] "@robykenobi @braves when he comes back, and hopefully soroka comes back, its over for the nl east bitches"

> #filter out braves

> Braves\_tweetsScorp <- tm\_map(Braves\_tweetsScorp, removeWords, "braves")

> head(Braves\_tweetsScorp[[1]]$content)

[1] "@jolly\_olive john smoltz might finally get over 1996 and 1999 if the win it this year."

[2] "@jjw4jld then they'd be the nl toronto blue jays ...\na+ offense with average (at best) pitching. \n\ngotta say, though ... the pitching has\nbeen excellent (so far <u+0001f91e>) against\nthe brewers &amp; the dodgers!\n\nblue jays play in the toughest division ...\nthe play in the weakest division. <u+0001f937><u+200d><u+2642><u+fe0f>"

[3] "@dalepla42824047 @frankqu55904875 @zwarrenz @realtoddkalas @astros @redsox @ both teams were 3 games above .500 against nl playoff teams."

[4] "@dalepla42824047 @frankqu55904875 @zwarrenz @realtoddkalas @astros @redsox @ and, unbalanced sched had playing al east (4 teams &gt; 90 wins) while la played al west (1 team over 90). 11 games against playoff teams in al, la 4.\ndodgers built a massive record off rockies and dbacks. (as well as pirates, mets, &amp; nats post trades)"

[5] "@dalepla42824047 @frankqu55904875 @zwarrenz @realtoddkalas @astros @redsox @ not really\nwhen the unbalanced schedule is taken into effect &amp; records after trade deadline, the current teams are pretty equal.\nafter deadline was a 5 gm adv for la. imo that is most telling &amp; shows 18 game differential doesn’t show how close these two current teams are."

[6] "@robykenobi @ when he comes back, and hopefully soroka comes back, its over for the nl east bitches"

> #remove punctuation

> Braves\_tweetsScorp <- tm\_map( Braves\_tweetsScorp, removePunctuation )

> Braves\_tweetsScorp <- tm\_map( Braves\_tweetsScorp, content\_transformer(function(x) gsub("'","",x)) )

> head(Braves\_tweetsScorp[[1]]$content)

[1] "jollyolive john smoltz might finally get over 1996 and 1999 if the win it this year"

[2] "jjw4jld then theyd be the nl toronto blue jays \na offense with average at best pitching \n\ngotta say though the pitching has\nbeen excellent so far u0001f91e against\nthe brewers amp the dodgers\n\nblue jays play in the toughest division \nthe play in the weakest division u0001f937u200du2642ufe0f"

[3] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox both teams were 3 games above 500 against nl playoff teams"

[4] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox and unbalanced sched had playing al east 4 teams gt 90 wins while la played al west 1 team over 90 11 games against playoff teams in al la 4\ndodgers built a massive record off rockies and dbacks as well as pirates mets amp nats post trades"

[5] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox not really\nwhen the unbalanced schedule is taken into effect amp records after trade deadline the current teams are pretty equal\nafter deadline was a 5 gm adv for la imo that is most telling amp shows 18 game differential doesn’t show how close these two current teams are"

[6] "robykenobi when he comes back and hopefully soroka comes back its over for the nl east bitches"

> # get rid of http links

> Braves\_tweetsScorp <- tm\_map( Braves\_tweetsScorp, content\_transformer(function(x) gsub("http\\S+","",x)) )

> head(Braves\_tweetsScorp[[1]]$content)

[1] "jollyolive john smoltz might finally get over 1996 and 1999 if the win it this year"

[2] "jjw4jld then theyd be the nl toronto blue jays \na offense with average at best pitching \n\ngotta say though the pitching has\nbeen excellent so far u0001f91e against\nthe brewers amp the dodgers\n\nblue jays play in the toughest division \nthe play in the weakest division u0001f937u200du2642ufe0f"

[3] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox both teams were 3 games above 500 against nl playoff teams"

[4] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox and unbalanced sched had playing al east 4 teams gt 90 wins while la played al west 1 team over 90 11 games against playoff teams in al la 4\ndodgers built a massive record off rockies and dbacks as well as pirates mets amp nats post trades"

[5] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox not really\nwhen the unbalanced schedule is taken into effect amp records after trade deadline the current teams are pretty equal\nafter deadline was a 5 gm adv for la imo that is most telling amp shows 18 game differential doesn’t show how close these two current teams are"

[6] "robykenobi when he comes back and hopefully soroka comes back its over for the nl east bitches"

> ## remove stopwords

> Braves\_tweetsScorp <- tm\_map(Braves\_tweetsScorp, removeWords, stopwords("english"))

> head(Braves\_tweetsScorp[[1]]$content)

[1] "jollyolive john smoltz might finally get 1996 1999 win year"

[2] "jjw4jld theyd nl toronto blue jays \n offense average best pitching \n\ngotta say though pitching \n excellent far u0001f91e \n brewers amp dodgers\n\nblue jays play toughest division \n play weakest division u0001f937u200du2642ufe0f"

[3] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox teams 3 games 500 nl playoff teams"

[4] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox unbalanced sched playing al east 4 teams gt 90 wins la played al west 1 team 90 11 games playoff teams al la 4\ndodgers built massive record rockies dbacks well pirates mets amp nats post trades"

[5] "dalepla42824047 frankqu55904875 zwarrenz realtoddkalas astros redsox really\n unbalanced schedule taken effect amp records trade deadline current teams pretty equal\n deadline 5 gm adv la imo telling amp shows 18 game differential doesn’t show close two current teams "

[6] "robykenobi comes back hopefully soroka comes back nl east bitches"

> # Stemming

> Braves\_tweetsScorp <- tm\_map(Braves\_tweetsScorp, stemDocument)

> head(Braves\_tweetsScorp[[1]]$content)

[1] "jollyol john smoltz might final get 1996 1999 win year"

[2] "jjw4jld theyd nl toronto blue jay offens averag best pitch gotta say though pitch excel far u0001f91 brewer amp dodger blue jay play toughest divis play weakest divis u0001f937u200du2642ufe0f"

[3] "dalepla42824047 frankqu55904875 zwarrenz realtoddkala astro redsox team 3 game 500 nl playoff team"

[4] "dalepla42824047 frankqu55904875 zwarrenz realtoddkala astro redsox unbalanc sched play al east 4 team gt 90 win la play al west 1 team 90 11 game playoff team al la 4 dodger built massiv record rocki dback well pirat met amp nat post trade"

[5] "dalepla42824047 frankqu55904875 zwarrenz realtoddkala astro redsox realli unbalanc schedul taken effect amp record trade deadlin current team pretti equal deadlin 5 gm adv la imo tell amp show 18 game differenti doesn’t show close two current team"

[6] "robykenobi come back hope soroka come back nl east bitch"

> # Lemmatization

> Braves\_tweetsScorp <- tm\_map(Braves\_tweetsScorp, content\_transformer(lemmatize\_strings))

> head(Braves\_tweetsScorp[[1]]$content)

[1] "jollyol john smoltz may final get 1996 1999 win year"

[2] "jjw4jld theyd nl toronto blue jay offens averag good pitch gotta say though pitch excel far u0001f91 brewer amp dodger blue jay play tough divis play weak divis u0001f937u200du2642ufe0f"

[3] "dalepla42824047 frankqu55904875 zwarrenz realtoddkala astro redsox team 3 game 500 nl playoff team"

[4] "dalepla42824047 frankqu55904875 zwarrenz realtoddkala astro redsox unbalanc sched play al east 4 team gt 90 win la play al west 1 team 90 11 game playoff team al la 4 dodger build massiv record rocki dback good pirat meet amp nat post trade"

[5] "dalepla42824047 frankqu55904875 zwarrenz realtoddkala astro redsox realli unbalanc schedul take effect amp record trade deadlin current team pretti equal deadlin 5 gm adv la imo tell amp show 18 game differenti doesn ’ t show close two current team"

[6] "robykenobi come back hope soroka come back nl east bitch"

> #remove numbers

> Braves\_tweetsScorp <- tm\_map(Braves\_tweetsScorp,removeNumbers)

> head(Braves\_tweetsScorp[[1]]$content)

[1] "jollyol john smoltz may final get win year"

[2] "jjwjld theyd nl toronto blue jay offens averag good pitch gotta say though pitch excel far uf brewer amp dodger blue jay play tough divis play weak divis ufuduufef"

[3] "dalepla frankqu zwarrenz realtoddkala astro redsox team game nl playoff team"

[4] "dalepla frankqu zwarrenz realtoddkala astro redsox unbalanc sched play al east team gt win la play al west team game playoff team al la dodger build massiv record rocki dback good pirat meet amp nat post trade"

[5] "dalepla frankqu zwarrenz realtoddkala astro redsox realli unbalanc schedul take effect amp record trade deadlin current team pretti equal deadlin gm adv la imo tell amp show game differenti doesn ’ t show close two current team"

[6] "robykenobi come back hope soroka come back nl east bitch"

> #remove special characters

> Braves\_tweetsScorp <- tm\_map( Braves\_tweetsScorp, content\_transformer(function(x) gsub("[^[:alnum:]]", " ", x)) )

> head(Braves\_tweetsScorp[[1]]$content)

[1] "jollyol john smoltz may final get win year"

[2] "jjwjld theyd nl toronto blue jay offens averag good pitch gotta say though pitch excel far uf brewer amp dodger blue jay play tough divis play weak divis ufuduufef"

[3] "dalepla frankqu zwarrenz realtoddkala astro redsox team game nl playoff team"

[4] "dalepla frankqu zwarrenz realtoddkala astro redsox unbalanc sched play al east team gt win la play al west team game playoff team al la dodger build massiv record rocki dback good pirat meet amp nat post trade"

[5] "dalepla frankqu zwarrenz realtoddkala astro redsox realli unbalanc schedul take effect amp record trade deadlin current team pretti equal deadlin gm adv la imo tell amp show game differenti doesn t show close two current team"

[6] "robykenobi come back hope soroka come back nl east bitch"

>

> #make data df

> Braves\_tweets\_dataframe <- data.frame(text=unlist(sapply(Braves\_tweetsScorp, `[`, "content")),

+ stringsAsFactors=F)

> #make bigram

> Braves\_tweets\_Bigrams <- Braves\_tweets\_dataframe %>% unnest\_tokens(bigram, text, token = "ngrams", n = 2)

>

>

> #Part 2D)

> #count the Bigrams

> Braves\_tweets\_Bigrams\_Counts <- Braves\_tweets\_Bigrams %>% dplyr::count(bigram, sort = TRUE)

> head(Braves\_tweets\_Bigrams\_Counts)

bigram n

1 astro redsox 5

2 dalepla frankqu 3

3 frankqu zwarrenz 3

4 realtoddkala astro 3

5 zwarrenz realtoddkala 3

6 battleatl chopon 2

>

> #Part 2E)

> #make a df so we can plot our words

> Braves\_tweets\_Bigrams = Braves\_tweets\_Bigrams %>%

+ separate(bigram, c("Word1", "Word2"), sep = " ")

> head(Braves\_tweets\_Bigrams)

Word1 Word2

1.content1...1 jollyol john

1.content1...2 john smoltz

1.content1...3 smoltz may

1.content1...4 may final

1.content1...5 final get

1.content1...6 get win

>

> Braves\_tweets\_Bigrams <- Braves\_tweets\_Bigrams %>%

+ filter(!Word1 %in% stop\_words$word) %>%

+ filter(!Word2 %in% stop\_words$word)

> head(Braves\_tweets\_Bigrams)

Word1 Word2

1.content1...1 jollyol john

1.content1...2 john smoltz

1.content2...3 jjwjld theyd

1.content2...4 theyd nl

1.content2...5 nl toronto

1.content2...6 toronto blue

>

> #time to plot our data and make sure that we make the counts visible in one way

> Braves\_tweets\_Bigrams <- Braves\_tweets\_Bigrams %>%

+ count(Word1, Word2, sort = TRUE)

> head(Braves\_tweets\_Bigrams)

Word1 Word2 n

1 astro redsox 5

2 dalepla frankqu 3

3 frankqu zwarrenz 3

4 realtoddkala astro 3

5 zwarrenz realtoddkala 3

6 battleatl chopon 2

> #Want to filter down the data so we can see trends and not have too mnay nodes

> Braves\_tweets\_Bigrams <- filter(Braves\_tweets\_Bigrams, Braves\_tweets\_Bigrams$n > 1)

> Braves\_tweets\_Bigrams\_Words <- Braves\_tweets\_Bigrams[c(1,2)]

> Braves\_tweets\_Bigrams\_Words <- as.matrix(Braves\_tweets\_Bigrams\_Words)

> #Braves\_tweets\_Bigrams <- as.matrix(Braves\_tweets\_Bigrams)

> G<- graph\_from\_edgelist(Braves\_tweets\_Bigrams\_Words, directed = F)

> E(G)$weight <- Braves\_tweets\_Bigrams[,3]

> plot(G,edge.width = E(G)$weight)