FIT3152
Assignment 3

Daniel Nguyen

32471033

Question 1

I have collected 15 news articles about 3 topic areas. Each document covers a different main topic that falls within one of these topic areas. My 3 topics are:

- Climate Change
- UFC (Ultimate Fighting Championship)
- Taiwan / ROC / Chinese Taipei

Sources for each document are contained within the references of this report

Question 2

To convert each news article into a machine readable text document, I copied and pasted the title, body text, and image captions into a .txt file. I made sure to not copy advertisement text, or other promotional text that did not contribute to the article in any meaningful way.

I collected these .txt files in a folder in my working directory, and after processing, wrote the document-term matrix to a .csv file to use in my R code, as demonstrated in the unit lecture material.

Each .txt file is labelled with its main topic area, followed by a number indicating the document number. For example, the first documents of each topic area are labelled as such:

- ClimChange1
- UFC1
- Taiwan1

Question 3

I found that when simply using sparse term removal to reach approximately 20 tokens, that the final list of tokens did not adequately distinguish between documents, so in addition to the base text transformations and sparse term removal, I also used Term Frequency - Inverse Document Frequency (TfIdf) to select terms, which ensured significant terms would be retained in our final token list.

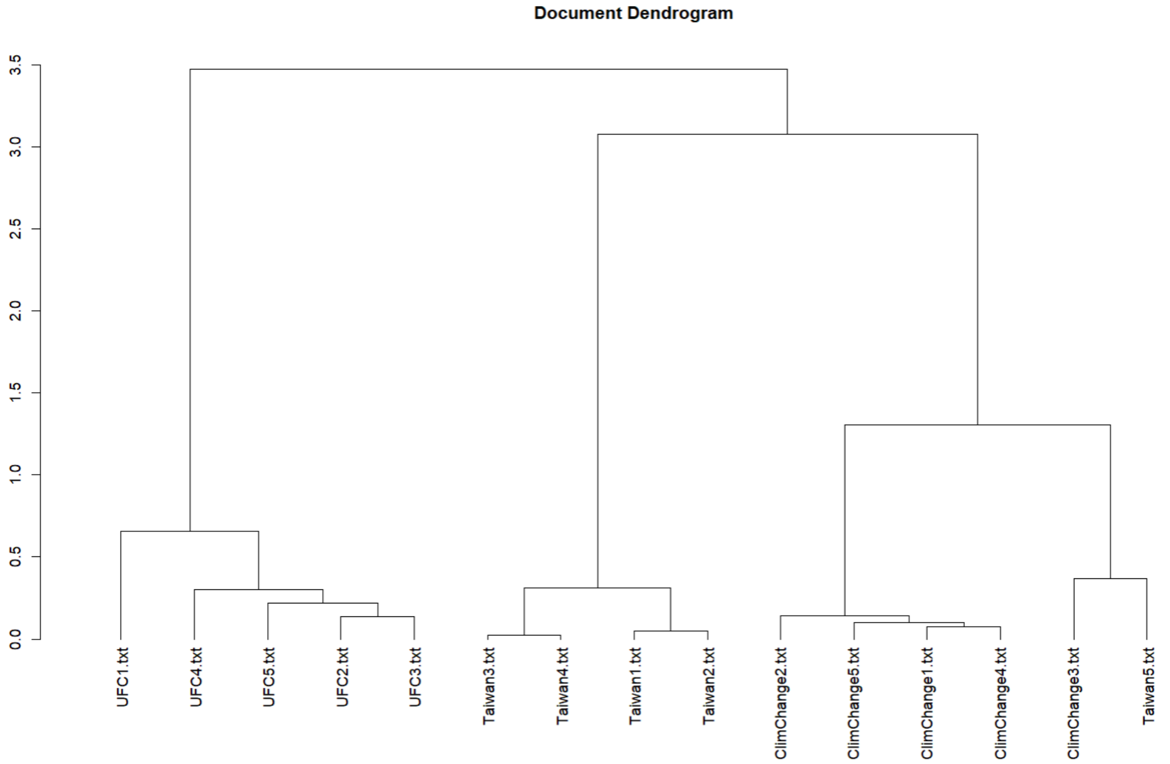
After cleaning the documents, I removed all punctuation, which involved removing all but alphanumeric characters, whitespace, and newline indicators. I then removed numbers, converted all text to lower case, removed stopwords, extra whitespace, and performed stemming on the text. After these steps were taken, there were a total of 2206 unique terms in my document collection.

The function removePunctuation does not properly remove typographical punctuation, so I had to resort to this method to clean the document.

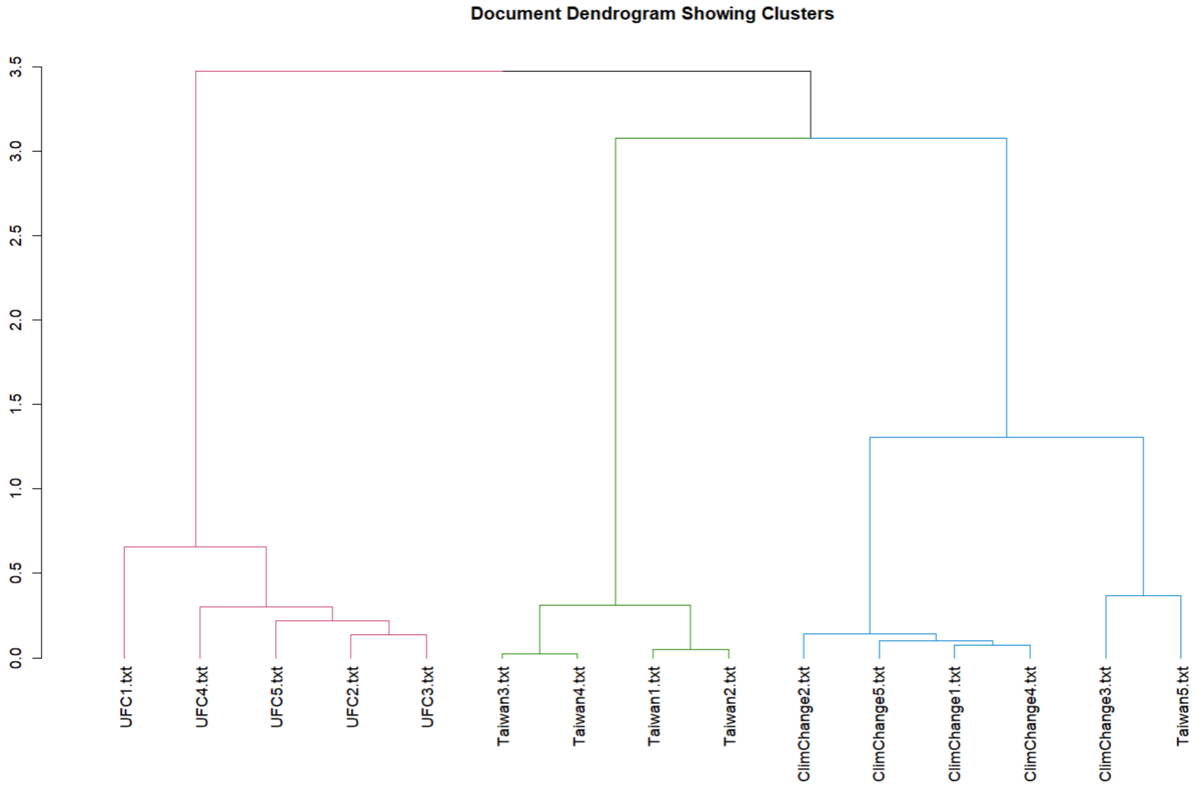
I then removed sparse terms with a cutoff of 0.7, and calculated the TfIdf weights for the remaining 175 terms, and selected the top 20 to create my token list. From this token list I created my document-term matrix, which was written to a .csv file. My DTM is contained in the appendix of this document.

Question 4

From my DTM, I was able to perform hierarchical clustering using cosine distance to identify splits, and create a dendrogram of my documents.



To see how accurate this was in classifying each document, we can colour the branches based on cluster. As there are 3 topic areas, we can choose k = 3 to see our accuracy.



We can see that the document clustering is almost perfect, with only the document Taiwan5.txt falling outside of the cluster that contains the rest of the documents concerning its topic area.

	Cluster 1	Cluster 2	Cluster 3
ClimChange	5		
Taiwan	1	4	
UFC			5

Correctly classified documents = 14
Total documents = 15

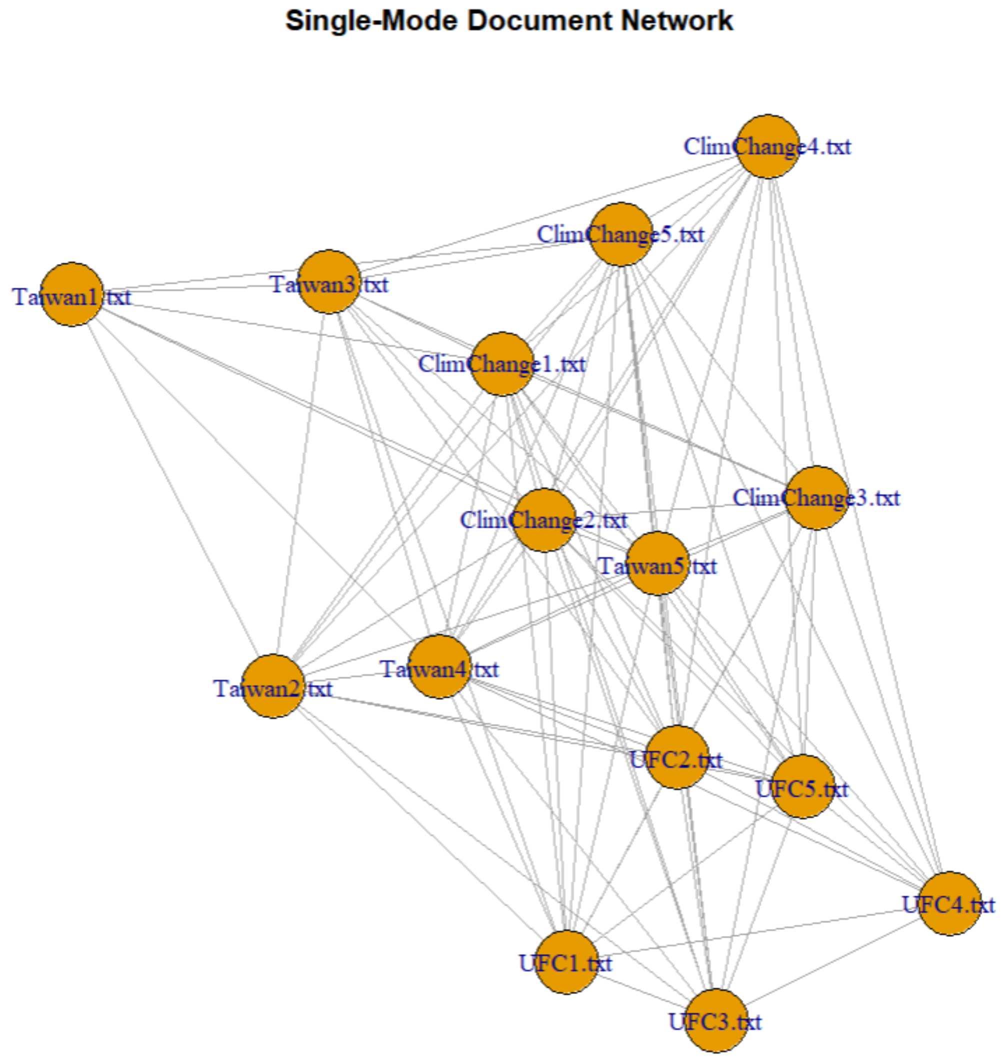
Accuracy = 14 / 15
= 93%

Overall, the clustering was very accurate, with the exception of the document Taiwan5.txt. Observing the contents of this news article, while the other Taiwan documents concern themselves with the geopolitical struggles of the region, Taiwan5.txt is about the rising cost of living crisis in the area, which is markedly different from the other news articles. After taking an outside, human look at the articles, it is understandable that this document was misclassified during the clustering.

Word stems that were prevalent in the other 4 documents, such as parti, and forc were not seen in this document, while other word stems such as chang, find, and dont were more used in this document, as well as the ClimChange documents, while they were rarely used, if at all, in the other documents from the same topic area.

Question 5

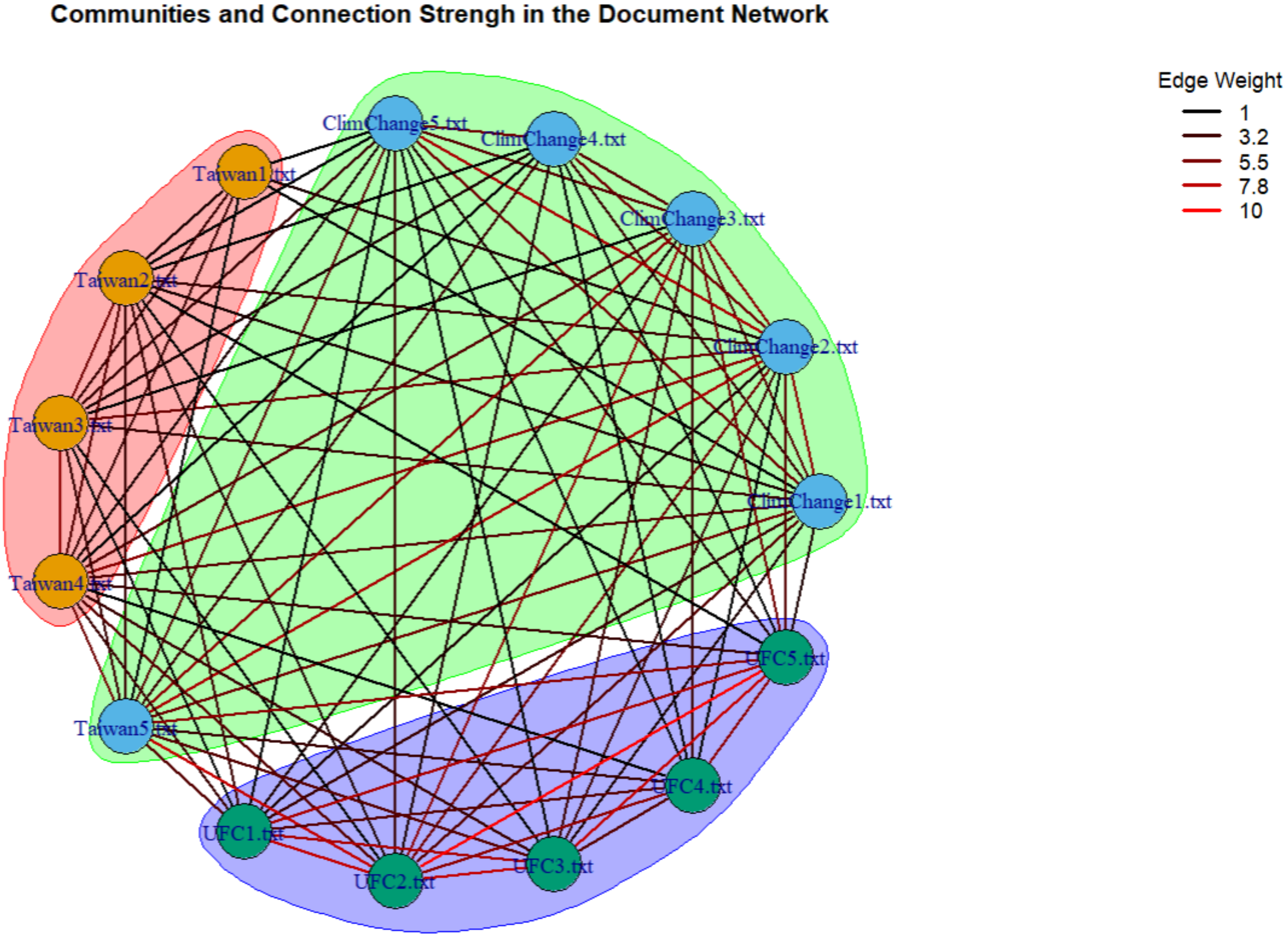
Creating a basic single-mode network based on unit lecture material, we are left with the following graph.



The densely connected graph shows that the documents share many of the significant terms with each other. However, from a basic look, we can identify that, visually, there are clear communities in line with the results of our earlier hierarchical clustering. Additionally, there are some documents that share relatively more terms with other documents, such as ClimChange2.txt and Taiwan5.txt, located in the centre of the graph. However, this is just a visual analysis. We can calculate the centrality of each node by counting how connected it is. A table of the documents degree of centrality is contained in the appendix.

When looking at the centrality measures of the graph, we can see that ClimChange2, ClimChange5, Taiwan4 and Taiwan5 are the most connected documents, as each shares at least 1 term with all 14 other documents.

We can highlight the communities in the document, and the strength of each connection in the following graph:

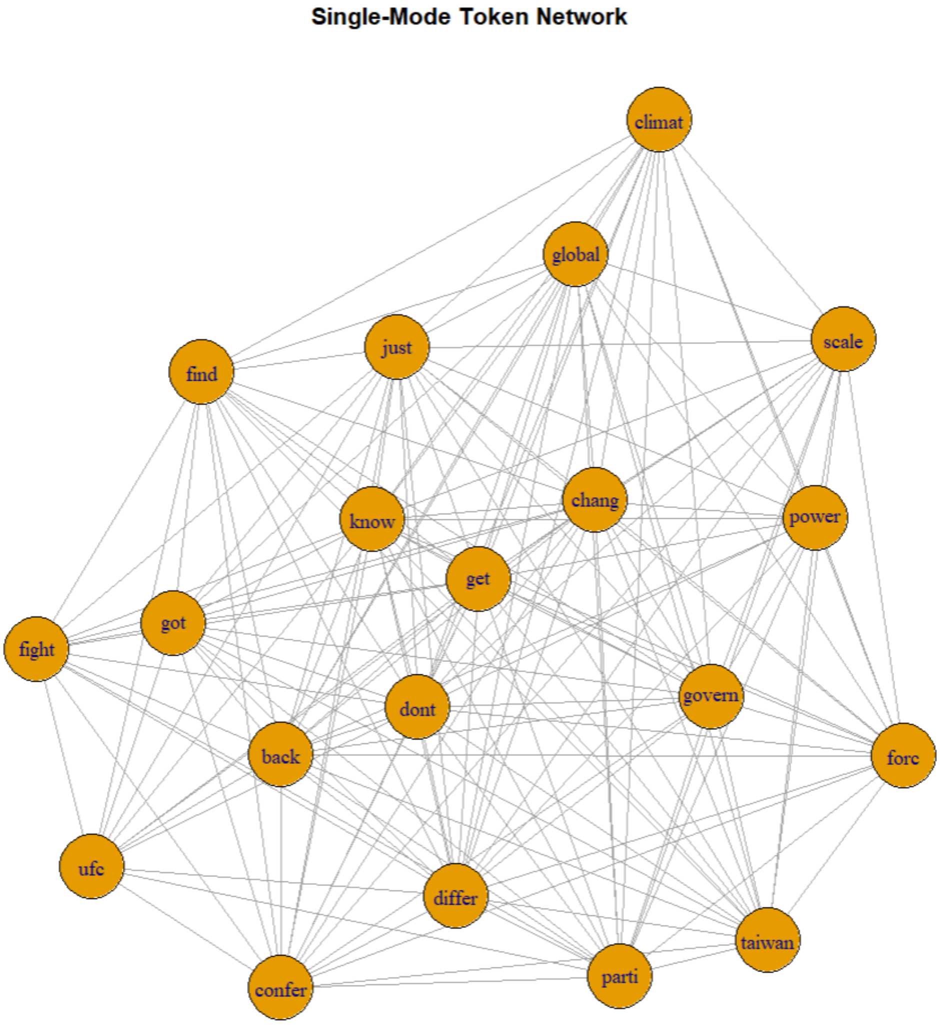


To create this graph, I used spinglass clustering from the igraph package to identify communities, and gave edges a colour, from black to red based on edge weight, which increased with the number of shared terms. Additionally, I used a circle layout for the graph.

From the graph, we can see the same clusters that we identified in our dendrogram, in addition to the strength of connectivity between documents. In particular, we can see that the UFC documents share many terms with each other, as indicated by the colour of the edges, as well as a relative weakness in connectivity within the taiwan cluster, which has more lines that are closer to black, which indicate that fewer terms are shared between documents.

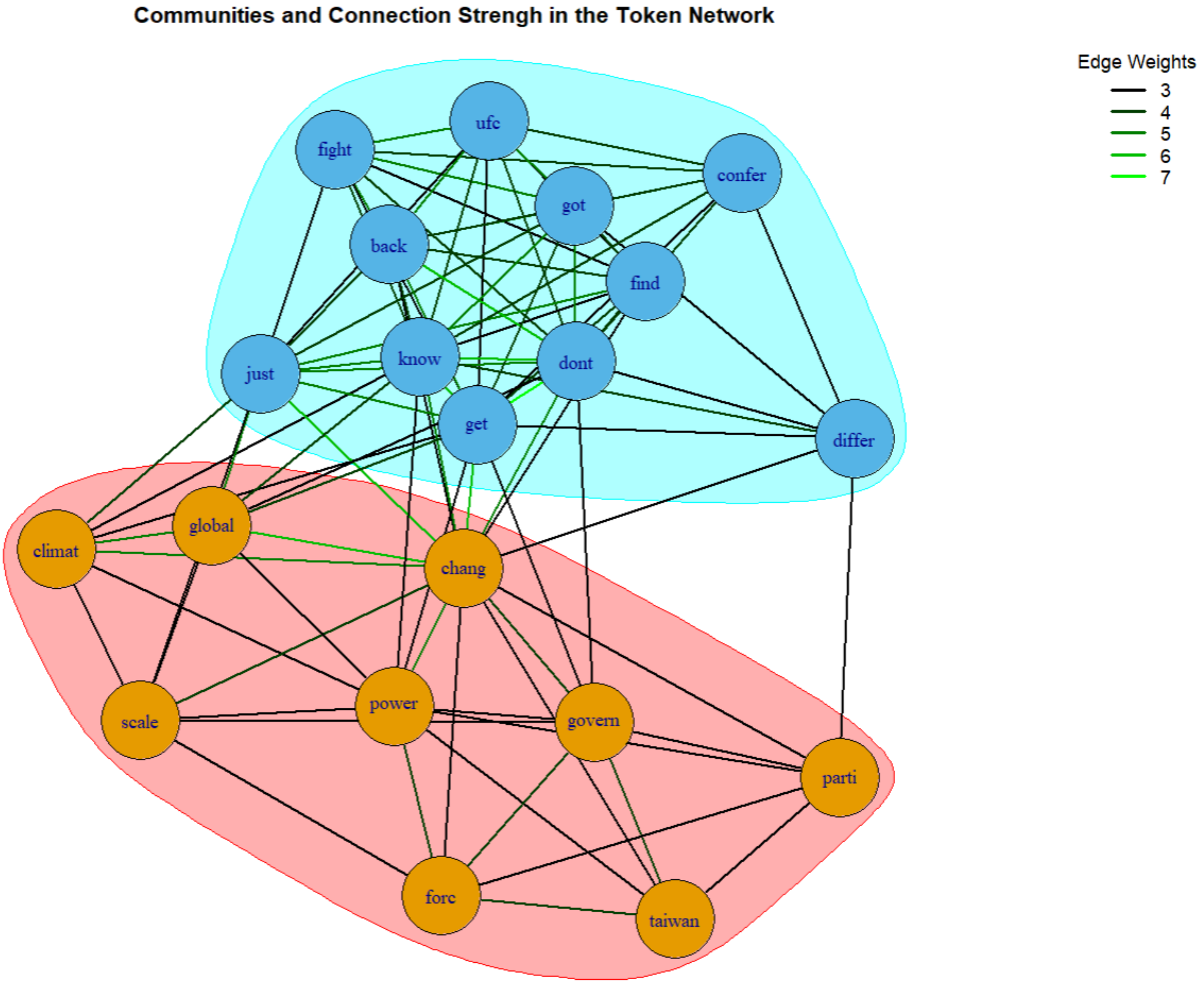
Question 6

A simple token network based on shared documents is shown below:



The network shows a node for each token in our DTM, and the links between the nodes indicate that at least 1 document uses both of those terms. We can see some nodes with a high degree of centrality, such as the stem get, which indicates that there are many other terms in the network that are used along with get in our documents.

Compared to our document network, it is more difficult to visually distinguish communities in our token network, as the graph has a higher density of connections. However, we can use various algorithms to find communities and highlight connection strength (higher edge weight indicates more documents share terms) between terms in our DTM:

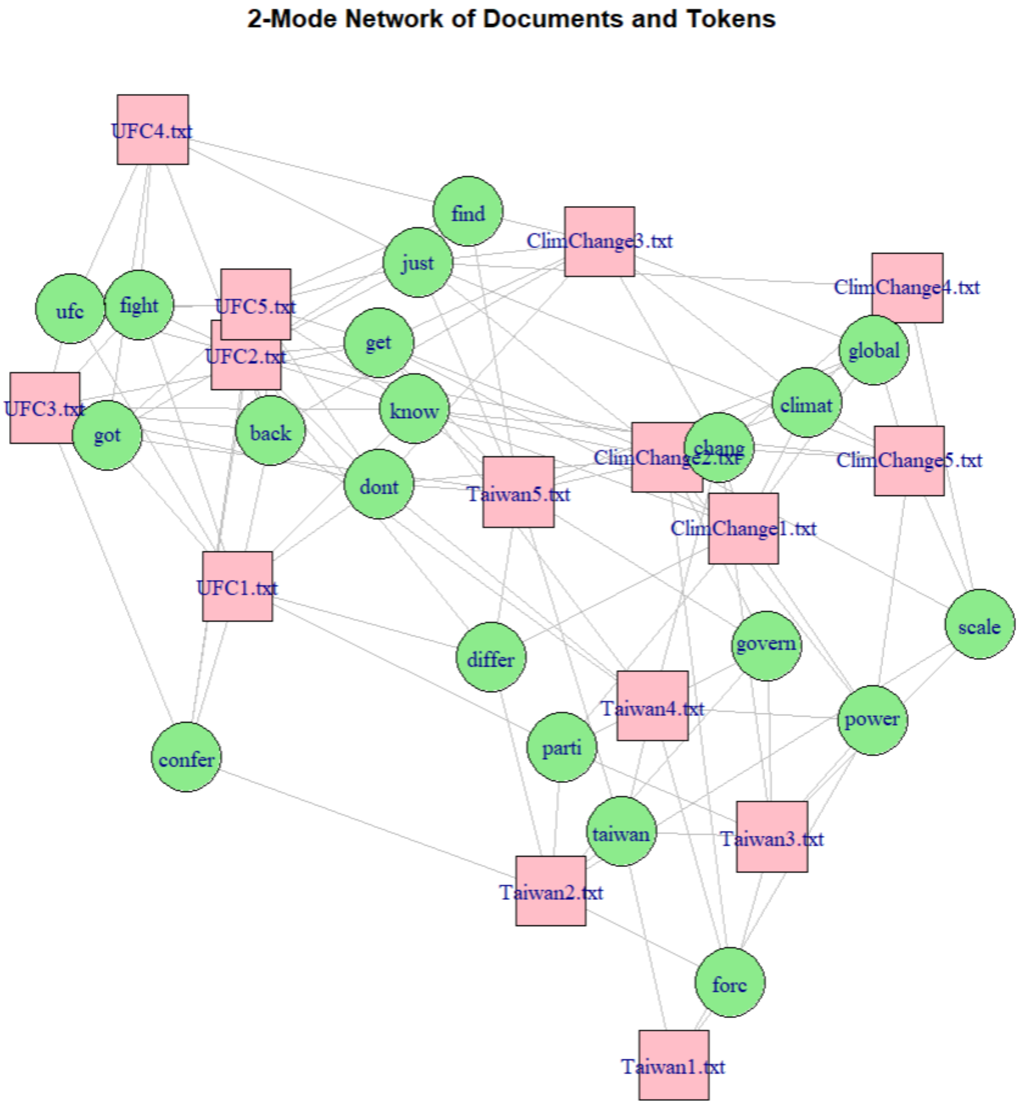


Since the graph is so dense, I have chosen to remove edges that are not significant, by taking the mean edge weight, and only taking those that are above the mean. This preserves edges that show significant links, while making the graph easier to visually parse.

Using igraph’s cluster_optimal function, 2 groups have been identified, showing that there are 2 groups of words that are more often used together in a document. Edges have also been coloured based on their weight, which highlights links that show 2 words have been often used together in a document. For example, we can see that the link between stems climat and chang shows a relatively higher weight than much of the rest of the edges in the network. This is to be expected, as one of our topic areas was indeed “climate change.”

Question 7

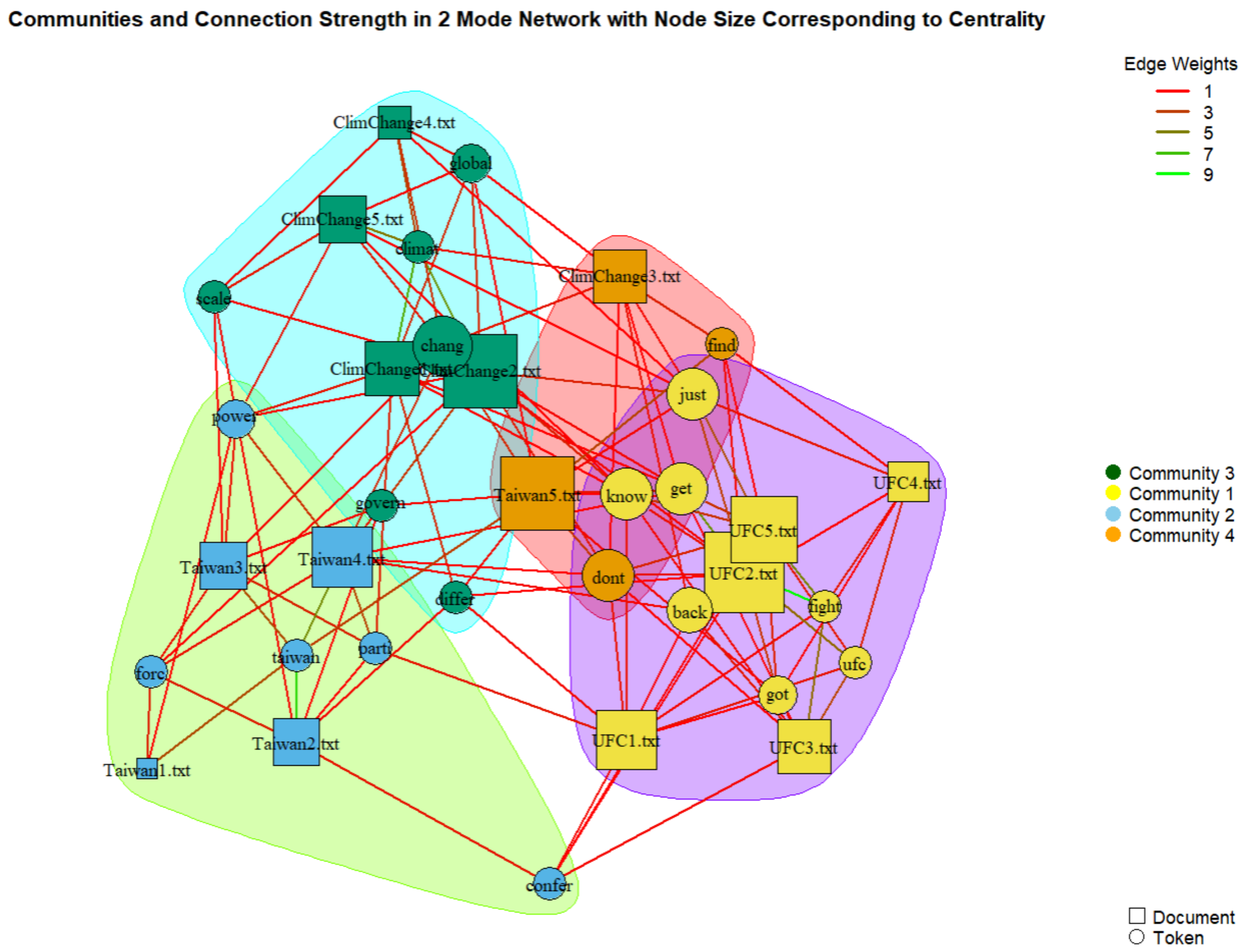
To create a 2-mode network of our documents and tokens, we can create an edge list, which details links between documents nodes and token nodes based on the number of times a token is used in each document, we then remove edges with 0 weight (the token is never used in the document), and from this data we can create a simple 2-mode network graph.



The graph shows us the links between tokens and documents, where a link between a token and a document means that the token is present in the document. Some tokens, such as the stem get, are very well connected, meaning they are used in many documents, while some have only a few edges, such as the stem confer, which has 4, meaning that the token is used in only 4 documents.

We can identify some groups of documents and words, the most pronounced being the tokens ufc and fight, which are surrounded by the document nodes for the UFC news articles, indicating that they are used in those articles, but rarely if at all in the other ones. Other groups can also be seen, such as around the Taiwan document nodes, with tokens such as taiwan and forc being locally central.

To better illustrate node centrality, communities in the network and connection strength, we can improve this graph:



We can see that 3 main communities have been identified, with a smaller community of only 2 documents and 2 tokens. This indicates that the documents, and the words used in those documents may be different from the rest.

The node sizes indicate the degree of each node, where larger ones are have more connections to nodes around them. Documents such as Taiwan5.txt contain many of the tokens in our DTM, and tokens such as chang or dont are used in many of the documents in our DTM.

Finally, edge weights are also highlighted, where green edges indicate that there is a strong link between the respective word and document, such as UFC2.txt and fight, which has a very high edge weight, indicating that the token fight is used many times in that document.

Question 8

Overall, I was able to use clustering and network analysis to split the documents into groups that were relatively accurate to the original topic areas. Documents like Taiwan5.txt and ClimChange2.txt were quite important, as they shared many tokens with the other documents in the corpus, while documents like Taiwan1.txt had few links to the other components in the network. I found that clustering worked well for identifying groups in the in the data, however network analysis was much more helpful in identifying important relationships, as the network graph type is much better suited for this than clustering, or a dendrogram. I was able to identify both tokens and documents that were central in their networks, as well as only locally central, such as the token fight.

Better feature selection would greatly help in creating a better DTM, where improvements in my initial DTM construction would allow more relevant tokens to be used. Such improvements could include better stop word removal, and a more thorough selection of tokens using TFIDF, or even a manual selection of terms.

Appendix

DTM (split into 2 for formatting)

	taiwan	climat	fight	ufc	parti	chang	forc	find	just	power
ClimChange1.txt	0	23	0	0	5	12	0	0	0	4
ClimChange2.txt	0	20	0	0	0	12	1	0	5	2
ClimChange3.txt	0	3	0	0	0	4	0	5	2	0
ClimChange4.txt	0	10	0	0	0	7	0	0	1	0
ClimChange5.txt	0	17	0	0	0	4	0	0	1	3
Taiwan1.txt	11	0	0	0	0	0	5	0	0	1
Taiwan2.txt	29	0	0	0	1	0	5	0	0	0
Taiwan3.txt	11	0	0	0	6	2	3	0	0	3
Taiwan4.txt	19	0	0	0	10	6	3	0	0	7
Taiwan5.txt	9	0	0	0	0	6	0	10	2	0
UFC1.txt	0	0	2	4	4	0	0	0	0	0
UFC2.txt	0	0	34	15	0	2	0	2	9	0
UFC3.txt	0	0	15	10	0	0	0	0	0	0
UFC4.txt	0	0	4	5	0	0	0	2	4	0
UFC5.txt	0	0	18	4	0	0	0	1	11	0

	find	just	power	back	got	know	dont	global	differ	govern	confer	get	scale
ClimChange1.txt	0	0	4	0	0	1	0	6	6	0	0	1	0
ClimChange2.txt	0	5	2	0	0	5	4	6	0	8	0	2	2
ClimChange3.txt	5	2	0	3	0	0	1	1	0	0	0	1	0
ClimChange4.txt	0	1	0	0	0	0	0	3	0	0	0	0	2
ClimChange5.txt	0	1	3	0	0	1	0	2	0	0	0	0	4
Taiwan1.txt	0	0	1	0	0	0	0	0	0	0	0	0	0
Taiwan2.txt	0	0	0	0	0	0	0	0	2	1	4	0	2
Taiwan3.txt	0	0	3	0	0	0	0	0	0	2	0	0	1
Taiwan4.txt	0	0	7	2	0	0	1	0	0	3	0	1	0
Taiwan5.txt	10	2	0	0	2	1	8	1	2	2	0	4	0
UFC1.txt	0	0	0	1	1	3	3	0	2	0	2	0	0
UFC2.txt	2	9	0	20	8	3	3	0	1	0	1	20	0
UFC3.txt	0	0	0	4	3	2	1	0	0	0	2	1	0
UFC4.txt	2	4	0	1	1	0	0	0	0	0	0	0	0
UFC5.txt	1	11	0	2	5	8	6	0	0	0	1	8	0

Document Centrality Measure (Degree)

Document	Degree
ClimChang1	13
ClimChang2	14
ClimChang3	12
ClimChang4	11
ClimChang5	14
Taiwan1	7
Taiwan2	12
Taiwan3	11
Taiwan4	14
Taiwan5	14
UFC1	12
UFC2	13
UFC3	11
UFC4	10
UFC5	12

References

China declares it is ready to “forcefully” stop Taiwan independence. (2024, June 2). ABC News. https://www.abc.net.au/news/2024-06-02/china-defence-chief-says-beijing-will-forcefully-stop-taiwan/103924938

China warns on Taiwan, South China Sea at Shangri-La forum. (2024, June 2). Voice of America. https://www.voanews.com/a/china-warns-on-taiwan-south-china-sea-at-shangri-la-forum/7639355.html

Dana White likens Conor McGregor to Jon Jones amid late night partying videos ahead of UFC 303. (2024, June 2). Bloodyelbow.com. https://bloodyelbow.com/2024/06/02/dana-white-likens-conor-mcgregor-to-jon-jones-amid-late-night-partying-videos-ahead-of-ufc-303/

Evans, S. (2024, June 4). Reality check: the Reform UK party’s claims on the climate crisis examined. The Guardian. https://www.theguardian.com/environment/article/2024/may/31/factcheck-no-richard-tice-volcanoes-are-not-to-blame-for-climate-change

Extreme heat: Climate change’s silent killer. (2024, June 2). Voice of America. https://www.voanews.com/a/extreme-heat-climate-change-s-silent-killer/7639605.html

Gergis, J. (2024, June 2). Are the climate wars really over, or has a new era of greenwashing just begun? The Guardian. https://www.theguardian.com/environment/article/2024/jun/03/are-the-climate-wars-really-over-or-has-a-new-era-of-greenwashing-just-begun

How fossils of giant ringtail possums, marsupial lions and monster koalas could hold clues to surviving climate change. (2024, June 1). ABC News. https://www.abc.net.au/news/2024-06-01/central-queensland-fossil-discoveries-scientists-climate-change/103919978

In expensive Taiwan, some young people are giving up on real estate to join the “moonlight clan” instead. (2024, May 27). ABC News. https://www.abc.net.au/news/2024-05-28/taipei-real-estate-expensive-young-people-give-up-on-owning-home/103897520

“Kick him back down to the minor leagues”... Dana White slams UFC 302 judge after controversial scoring. (2024, June 2). Bloodyelbow.com. https://bloodyelbow.com/2024/06/02/kick-him-back-down-to-the-minor-leagues-dana-white-slams-ufc-302-judge-after-controversial-scoring/

Martin, D. (2024, June 2). Dustin Poirier explains what went wrong against Islam Makhachev and where he goes from here. MMA Fighting. https://www.mmafighting.com/2024/6/2/24169756/dustin-poirier-explains-what-went-wrong-against-islam-makhachev-and-where-he-goes-from-here

Meshew, J. (2024, June 2). Khabib Nurmagomedov: Islam Makhachev “grew a lot” in UFC 302 win over Dustin Poirier. MMA Fighting. https://www.mmafighting.com/2024/6/2/24169744/khabib-nurmagomedov-islam-makhachev-grew-a-lot-in-ufc-302-win-over-dustin-poirier

“No discussion, no democracy”: Why Taiwan’s new parliamentary reforms are leading to scenes like this. (2024, May 31). ABC News. https://www.abc.net.au/news/2024-06-01/why-have-taiwan-s-politicians-been-brawling-in-parliament/103915644

Ognyanova, K. (n.d.). Katya Ognyanova: Rutgers Prof., Network Researcher, Data Scientist. Katya Ognyanova. Retrieved June 6, 2024, from https://kateto.net/

Taiwan’s parliament passes bill pushing pro-China changes. (n.d.). Al Jazeera. Retrieved June 6, 2024, from https://www.aljazeera.com/news/2024/5/28/taiwans-parliament-passes-bill-pushing-pro-china-changes

UFC 302 Wrap: Makhachev submits Poirier to defend belt; “criminal” scorecard on co-main event. (2024, June 2). Fox Sports. https://www.foxsports.com.au/ufc/ufc-302-islam-makhachev-vs-dustin-poirier-live-updates-start-time-full-fight-card-results-blog-stream-sean-strickland-vs-paulo-costa/news-story/09f898d5a756a697f1c575c6c5d0fd34

Wegener, J., Mehdi Ghissassi, & Maher, H. (2024, May 29). How AI and climate change innovations can be used in the real world. World Economic Forum. https://www.weforum.org/agenda/2024/05/ai-lift-climate-research-out-lab-and-real-world/

R Code

```
library(slam)
library(tm)
library(SnowballC)
library(igraph)
library(dendextend)
library(igraphdata)

rm(list = ls())

#create corpus
cname = file.path(".", "Documents")
docs = Corpus(DirSource((cname)))
summary(docs)

#text transformations
toJIT <- content_transformer(function(x, pattern) gsub(pattern, "JIT", x))
docs <- tm_map(docs, toJIT, "Just-In-Time")

toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "-")

#tokenisation and stemming
#removePunctuation does not work well with typographical punctuation marks
#so i am using this to remove all non alphanumeric characters.
#i have kept newline char to preserve spaces between words on different lines
#https://stackoverflow.com/questions/30994194/quotes-and-hyphens-not-removed-by-tm-package-functions-while-cleaning-corpus
removeSpecialChars <- function(x) gsub("[^a-zA-Z0-9 \n]","",x)
docs <- tm_map(docs, removeSpecialChars)

docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, stripWhitespace)
docs <- tm_map(docs, stemDocument, language = "english")

#print ClimChange1
writeLines(as.character(docs[[1]]))

#create dtm
dtm <- DocumentTermMatrix(docs)

#print number of terms
freq <- colSums(as.matrix(dtm))
length(freq)
ord = order(freq)

#show most/least used terms
freq[head(ord)]
freq[tail(ord)]

#remove sparce terms
```

```
dtms <- removeSparseTerms(dtm, 0.7)

#perform Term frequency - Inverse Document Frequency to preserve important terms
tfidf <- weightTfIdf(dtms)
tfidf <- col_sums(as.matrix(tfidf))
finalTerms <- sort(tfidf, decreasing = TRUE)[1:20]

dtms <- dtms[, names(finalTerms)]

#show document term matrix dimensions
dim(dtms)

#convert to matrix
dtms <- as.matrix(dtms)

#write csv
write.csv(dtms, "dtms.csv")

#create distance matrix
distmatrix = proxy::dist(dtms, method = "cosine")

#fix plot margins
par(mar = c(10, 4, 4, 2))

#clustering and plot dendrogram
fit = hclust(distmatrix, method = "ward.D")
dendrogram <- as.dendrogram(fit)
plot(dendrogram, main = "Document Dendrogram")

#colour clusters and plot
dendrogramColour <- color_branches(dendrogram, k = 3)
plot(dendrogramColour, main = "Document Dendrogram Showing Clusters")

#analyse accuracy
topics = c("ClimChange","ClimChange","ClimChange", "ClimChange", "ClimChange", "Taiwan",
        "Taiwan", "Taiwan", "Taiwan", "Taiwan", "UFC", "UFC", "UFC", "UFC",
        "UFC")

groups = cutree(fit, k = 3)
TA = as.data.frame.matrix(table(GroupNames =topics, Clusters = groups))
TA = TA[,c(1, 2, 3)]
TA

##############################################################
# Accuracy = 93%
# calculated by correct classifications / number of documents
##############################################################

#create network data for single mode network for documents
dtmsx = dtms
dtmsx = as.matrix((dtmsx > 0) + 0)
ByDocMatrix = dtmsx %*% t(dtmsx)
diag(ByDocMatrix) = 0
```

```
#create and plot graph
ByDoc = graph_from_adjacency_matrix(ByDocMatrix,mode = "undirected", weighted = TRUE)
plot(ByDoc, main = "Single-Mode Document Network")

#get centrality
degree(ByDoc)

#community identification
set.seed(32471033)
cfb =  cluster_fast_greedy(ByDoc)
plot(cfb, ByDoc,vertex.label = V(ByDoc)$role,
     main = "Fast-Greedy Algorithm Community Clustering")

################################
#improved graph of documents
#shows communities
#shows strength of connections
################################


ByDoc.sp <- ByDoc

#graph layout
l <- layout_in_circle(ByDoc.sp)

#get communities
set.seed(32471033)
cl <- cluster_spinglass(ByDoc.sp)

#normalise weight
normalise <- function(x) (x - min(x)) / (max(x) - min(x))
normalisedWeights <- normalise(E(ByDoc.sp)$weight)

#set edge colour based on weight
edgeColour <- colorRampPalette(c("black", "red"))
E(ByDoc.sp)$color <- edgeColour(100)[(cut(normalisedWeights, breaks = 100))]

#plot graph
plot(cl, ByDoc.sp, layout = l, edge.width = 2, edge.color = E(ByDoc.sp)$color,
     main = "Communities and Connection Strengh in the Document Network")

#format legend
legendColours <- edgeColour(5)
legendWeights <- seq(min(E(ByDoc.sp)$weight), max(E(ByDoc.sp)$weight), length.out = 5)
legendLabels <- round(legendWeights, 1)

#add legend to graph
legend("topright", legend = legendLabels, col = legendColours, lwd = 3,
       title = "Edge Weight", bty = "n", cex = 1)

#################################################################################

#create network data for single mode network for tokens
dtmsx = dtms
dtmsx = as.matrix((dtmsx > 0) + 0)
byTokenMatrix = t(dtmsx) %*% dtmsx
diag(byTokenMatrix) = 0

#create and plot graph
byToken = graph_from_adjacency_matrix(byTokenMatrix,mode = "undirected", weighted = TRUE)
plot(byToken, main = "Single-Mode Token Network")

###############################
#improved graph of tokens
#shows communities
#shows strength of connections
###############################

#remove edges with low weight (below mean edge weight)
token.cut.off <- mean(E(byToken)$weight)
byToken.sp <- delete_edges(byToken, E(byToken)[weight<token.cut.off])

#graph layout
l <- layout_with_fr(byToken.sp)

#get communities
clp <- cluster_optimal(byToken.sp)

#normalise weight
normalise <- function(x) (x - min(x)) / (max(x) - min(x))
normalisedWeights <- normalise(E(byToken.sp)$weight)

#set edge colour based on weight
edgeColour <- colorRampPalette(c("black", "green"))
E(byToken.sp)$color <- edgeColour(100)[(cut(normalisedWeights, breaks = 100))]

#plot graph
plot(clp, byToken.sp, layout = l, vertex.size = 20, edge.width = 2, edge.color = E(byToken.sp)$color,
     main = "Communities and Connection Strengh in the Token Network")

#format legend
legendColours <- edgeColour(5)
legendWeights <- seq(min(E(byToken.sp)$weight), max(E(byToken.sp)$weight), length.out = 5)
legendLabels <- round(legendWeights, 2)

#add legend to graph
legend("topright", legend = legendLabels, col = legendColours, lwd = 3,
       title = "Edge Weights", bty = "n", cex = 1)

###############################################

#2 mode graph
#code taken from lecture slides
dtmsa = as.data.frame(dtms)
dtmsa$ABS = rownames(dtmsa)
dtmsb = data.frame()
for (i in 1:nrow(dtmsa)){
  for (j in 1:(ncol(dtmsa)-1)){
    touse = cbind(dtmsa[i,j], dtmsa[i,ncol(dtmsa)],
                  colnames(dtmsa[j]))
    dtmsb = rbind(dtmsb, touse ) } }
colnames(dtmsb) = c("weight", "doc", "token")
dtmsc = dtmsb[dtmsb$weight != 0,]
dtmsc = dtmsc[,c(2,3,1)]

#plot graph
g <- graph.data.frame(dtmsc, directed=FALSE)
bipartite.mapping(g)
V(g)$type <- bipartite_mapping(g)$type
V(g)$color <- ifelse(V(g)$type, "lightgreen", "pink")
V(g)$shape <- ifelse(V(g)$type, "circle", "square")
E(g)$color <- "gray"

vertex.label <- V(g)$name

l <- layout_nicely(g)
plot(g, layout = l, main = "2-Mode Network of Documents and Tokens")

################################
#improved 2 mode Graph
#shows communities
#shows strength of connections
################################

#get data
g.sp <- g

#set margins
par(mar = c(3, 3, 3, 3))

#graph layout
l <- layout_with_graphopt(g.sp)

#get communities
set.seed(32471033)
clp <- cluster_spinglass(g.sp)

#normalise weight
normalise <- function(x) (x - min(x)) / (max(x) - min(x))
normalisedWeights <- normalise(as.numeric(E(g.sp)$weight))

#set edge colour based on weight
edgeColour <- colorRampPalette(c("red", "green"))
E(g.sp)$color <- edgeColour(100)[(cut(normalisedWeights, breaks = 100))]

#change size based on centrality
V(g.sp)$size <- degree(g.sp) * 1.75
#V(g.sp)$label.cex <- degree(g.sp) * 0.2

#plot graph
V(g.sp)$label.color <- "black"
plot(clp, g.sp, layout = layout.auto, edge.width = 2, edge.color = E(g.sp)$color,
    main = "Communities and Connection Strength in 2 Mode Network with Node Size Corresponding to
Centrality")

#format weight legend
legendColours <- edgeColour(5)
legendWeights <- seq(min(E(g.sp)$weight), max(E(g.sp)$weight), length.out = 5)
legendLabels <- round(legendWeights, 2)

#add legend
legend("topright", legend = legendLabels, col = legendColours, lwd = 3,
       title = "Edge Weights", bty = "n", cex = 1)

#format node colour legend
nodeCommunity <- unique(membership(clp))
nodeColour <- c("yellow", "skyblue", "darkgreen", "orange")

#add legend
legend("right", legend = paste("Community", nodeCommunity),
       col = nodeColour[nodeCommunity], pch = 19, pt.cex = 2, bty = "n")

#add node shape legend
legend("bottomright", legend = c("Document", "Token"),
       pch = c(0, 1), pt.cex = 2, bty = "n")
```