Table of Contents

Overview

In my assignment, I have chosen to make space exploration as my main area of interest, and have chosen my specific topics as space movies, dark matter, space missions, and space info articles. Due to my interest in seeing how well the clustering algorithm can separate these into their respective topics, and as a test I have also chosen to augment 3 documents from a field unrelated to space: NYC food reviews. I have collected 16 documents, which as according to the numbering in the references are arranged as:

1-4 : movie_sum 1-4 (references 1-4 corresponds with movie_sum_1, and so on)

5-6: mission_overview 1-2

7-8: info_article 1-2

9-11: nyc_food 1-3

12-16: dark_matter 1-5

My corpus contains a folder of text files, in the folder CorpusAbstracts.

Creating DTM

For the text transformations in my documents, one challenge I faced were the existence of alternate quotation marks, which look like. ‘ ’ and “ ” which instead of ' ", which the remove punctuation wasn’t picking up hence I was being left with stems such as wasn’ in the documents after tokenisation. I suspected this would reduce the accuracy, therefore I changed all the alternate quotations, to their respective form, which R’s removePuntuation would pick up. Additionally, upon inspection of the individual documents, I noticed that X-ray was being treated as two different tokens, so I changed it such that it would be treated as one (xray). After this and trial and error, I came to the using 0.4 in my remove sparse terms functions as it left me with 25 tokens in my DTM. The DTM is attached in the appendix. Upon inspection of the DTM we can see that the token star is the most common token, appearing 64 times.

Clustering

For clustering, I have chosen to use cosine distance, as it is more likely to give better clustering results. After inspection of the plot, and creating 5 clusters from the dendrogram alone, we can see the info articles are clustered together, and for the other 4 clusters, we can see that all of the clusters are always incorrect by one external document in the cluster. For example if we look at the 3rd cluster in the dendrogram(appendix), we can see that movie_sum_1, movie_sum_3, movie_sum_4 are clustered together, and dark_matter_2 is the only impure document in this cluster. This indicates that our clustering is able to differentiate effectively, but still contains a small margin of error. Alternatively, if we look at the confusion matrix (appendix), and manually assign cluster 1 to nyc food.
,cluster 2 to dark matter, cluster 3 to mission overviewcluster 4 to movie summary, cluster 5 to info_article, giving us a accuracy of 43.75%, as a quantitative measure of how the dendrogram Is performing at clustering. Whilst this number is low, it is not indicative of the true capability of the dendrogram is at clustering, as it was very close when observing the plot, and many of the clusters were close to being pure.

Single Mode Network(Documents)

In order to compute the strength of connections between each document, I utilised the method taught in week 12, where I created a binary matrix from the DTM, and then multiplied it by its transpose, and made the leading diagonal 0. Whilst inspecting this network (appendix), when analysing the plot for the abstract network we can see that dark matters is grouped together near the middle of the map, nyc foods at the top, the movie summaries near the right and mission overviews on the left. No clear inference can be made about the location of info articles. From a visual glance we can see that all dark matter texts, except dark matter 1 have high importance, as the width of their lines are larger, indicating that their connections are weighted more heavily, and that the other dark matters are situated more towards the middle.

When doing a numerical analysis of the nodes (see appendix for table) we can see all nodes have a degree of 15, meaning that they all share some connection. Mission overview 1 has the highest betweenness score at 23.5. Mission overview 1 also has the highest closeness at 0.00885. When inspecting Eigenvector centrality, dark matter 3 has the highest centrality at 1.00. This could be explained by the fact that the higher weighted neighbours of dark matter 3 are more central in the graph. It is interesting to note that Mission overview 1, which had the highest result for closeness and betweenness has the lowest Eigenvector centrality at 0.493. We can conclude that mission overview 1 is a highly important node, and so is dark matter 3. However, due to the fact that the result for the Eigenvector centrality aligns more closely with the visual conclusion, I will only consider the answer of that, hence dark matter 3 is a very important plot.

To improve my network plot, I made the edge widths be based on the weights of that edge, and the vertex size be based on the eigenvector centrality of that node. This means that it is easier to tell importance of specific nodes, as well as any specific relationships of those nodes.

Single Mode Network(Tokens)

I used same method to compute the strength of connections in this as I did before but used it for tokens instead. Whilst we cannot group the tokens in the same way we can with documents, we can till, investigate the plot from a visual perspective. Though majority of the texts were based on the topic of space exploration, the token “space” was not the most important node in this network, as the width of edges is lower than some of the others, and it does not occupy a highly central spot. The tokens that occupy the central spot are more generalised tokens which aren’t specific to topics such as “one”, “like”, “point”. One noteworthy token is “time” as it occupies a mildly centralized location despite not being a general term and not a direct link to space exploration.

Upon doing a numerical inspection of the nodes, we can see that all nodes have a degree of 24 and a betweenness of 0. “remain” has the highest closeness centrality at 0.0061, whereas “one” has the highest Eigenvector centrality at 1.00. This could be explained by the fact that the close neighbours of one have more importance. Due to this, I will be using eigenvector centrality as my measure of centrality as it supports my conclusion from my visual analysis.

To improve my network plot, I made the edge widths be based on the weights of that edge, and the vertex size be based on the eigenvector centrality of that node. This means that it is easier to tell importance of specific nodes, as well as any specific relationships of those nodes.

Bipartite Network

To transform the data into a suitable format, I initially used for loops to create the data into a suitable format. We can see that the nyc food is situated on the top left side, separated from the rest as its own community of documents. We can see that movie summaries and dark matter as the documents which crowd the centre, and tokens “space”, “one” and “star” as the most dominant token nodes. Whilst it is difficult to form proper communities, we can see that the edges of nyc food are not linked very strongly to then specific tokens about space exploration which is to be expected. Token “time” shows stronger connections towards documents that are not of topic dark matter.

When looking at the numerical attributes of the bipartite network, we can see that dark matter 3 has the highest overall degree at 23, dark matter 4 has highest betweenness at 73.9, dark matter 4 has the highest closeness at 0.0132, and token “space” has the highest eigenvector at 1.00.
This is interesting, because when looking at tokens only, the highest eigenvector was “one”. This could be due to the fact that a larger number of documents had space as their underlying theme, therefore when adding that as a factor, it could skew the importance of token “space” higher. This is also interesting because for the first 3 measures of centrality, degree, betweenness and closeness, the abstracts ranked the highest, however for eigenvector it was a token. This could indicate that the bipartite network is able to comprehend the importance of “space” to the collection of documents, when given the context of all documents and not just the tokens by themselves.

This network illustrates how the relationship between words and documents greatly affects the relations, had we looked only at one type at a time. Words which are more important throughout the corpus are shown to have increased importance in the bipartite graph, when compared to the single node graph. The addition of both documents and words strengthens the more important nodes, therefore is a better measure of importance of nodes, in a corpus.

To improve my network plot, I made the edge widths be based on the weights of that edge, and the vertex size be based on the eigenvector centrality of that node. This means that it is easier to tell importance of specific nodes, as well as any specific relationships of those nodes.

Summary

Throughout this investigation we can see that the dark matter topic has had high importance, as it is a large part of the collection of documents. Further, the movie summaries also have consistently shows significant importance throughout. The addition of the nyc foods topics have shown how the networks are able to comprehend that these specific topics are unlinked, and therefore have been separated upon review of the networks.

One very interesting outcome, was how the relative importance of certain nodes increased when going from the single node graph to the bipartite nodes.

References

1. Ebert, R. (2014, November 5). Interstellar. RogerEbert.com. Retrieved June 13, 2023, from https://www.rogerebert.com/reviews/interstellar-2014

2. Ebert, R. (2015, October 2). The Martian. RogerEbert.com. Retrieved June 13, 2023, from https://www.rogerebert.com/reviews/the-martian-2015

3. Ebert, R. (2018, October 12). First Man. RogerEbert.com. Retrieved June 13, 2023, from https://www.rogerebert.com/reviews/first-man-2018

4. Ebert, R. (1968, April 3). 2001: A Space Odyssey. RogerEbert.com. Retrieved June 13, 2023, from https://www.rogerebert.com/reviews/great-movie-2001-a-space-odyssey-1968

5. NASA. (2022, January 6). Apollo 11 mission overview. NASA.gov. Retrieved June 13, 2023, from https://www.nasa.gov/mission_pages/apollo/missions/apollo11.html

6. NASA. (2018, October 31). Kepler mission overview. NASA.gov. Retrieved June 13, 2023, from https://www.nasa.gov/mission_pages/kepler/overview/index.html

7. Britannica. (2017, October 19). The Fermi Paradox: Where Are All the Aliens? Retrieved June 13, 2023, from https://www.britannica.com/story/the-fermi-paradox-where-are-all-the-aliens

8. Space.com. (2022, January 29). How Big is the Universe? Retrieved June 13, 2023, from https://www.space.com/24073-how-big-is-the-universe.html

9. The Infatuation. (2023, March 1). Lucali. The Infatuation. Retrieved June 13, 2023, from https://www.theinfatuation.com/new-york/reviews/lucali

10. The Infatuation. (2023, March 1). Joe's Pizza. The Infatuation. Retrieved June 13, 2023, from https://www.theinfatuation.com/new-york/reviews/joes-pizza

11. The Infatuation. (2023, March 1). Katz's Deli. The Infatuation. Retrieved June 13, 2023, from https://www.theinfatuation.com/new-york/reviews/katzs-deli

12. Britannica. (2022, February 3). Dark Matter. Retrieved June 13, 2023, from https://www.britannica.com/science/dark-matter

13. NASA Science. (n.d.). What is Dark Energy? Retrieved June 13, 2023, from https://science.nasa.gov/astrophysics/focus-areas/what-is-dark-energy

14. Smithsonian Magazine. (2020, March 16). New Generation of Dark Matter Experiments Gear Up to Search for Elusive Particle. Retrieved June 13, 2023, from https://www.smithsonianmag.com/science-nature/new-generation-dark-matter-experiments-gear-search-elusive-particle-180974111/

15. Physics World. (2023, February 24). New Theory Links Supermassive Black Holes and Dark Energy. Retrieved June 13, 2023, from https://physicsworld.com/a/new-theory-links-supermassive-black-holes-and-dark-energy/#:~:text=A%20controversial%20new%20theory%20suggests,accelerating%20expansion%20of%20the%20universe

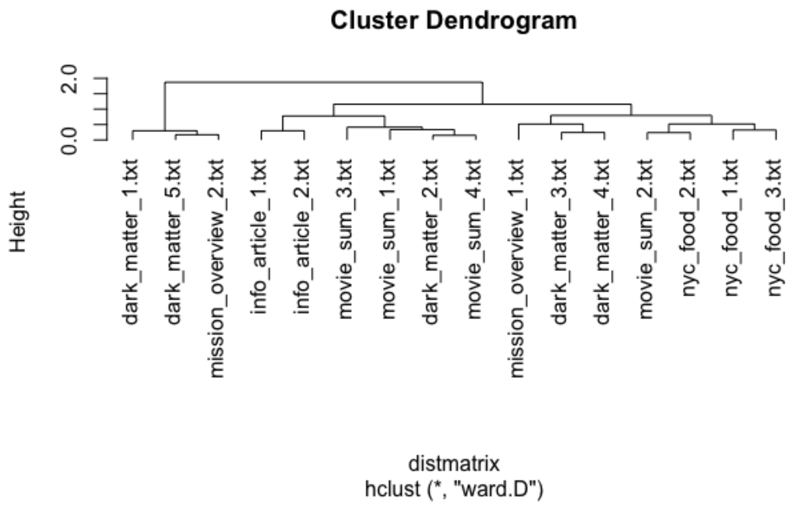
16. Gizmodo. (2018, January 29). What's the Speed of Dark Matter? Retrieved June 13, 2023, from https://gizmodo.com/whats-the-speed-of-dark-matter-1822465813

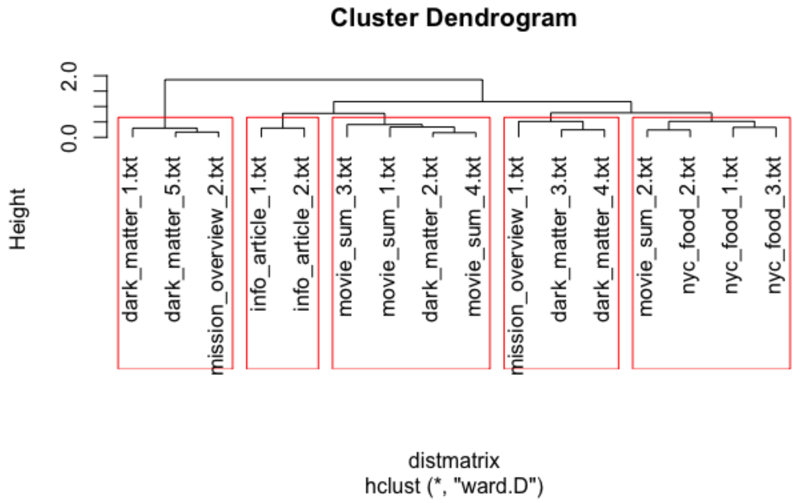
Appendix

DTM

	x
star	64
space	60
one	58
like	45
can	43
year	40
make	37
first	36
time	35
might	28
two	26
will	26
even	25
know	25
way	24
just	24
look	24
made	23
possibl	22
still	22
also	20
point	17
long	15
cant	15
remain	13

Cluster Dendrograms





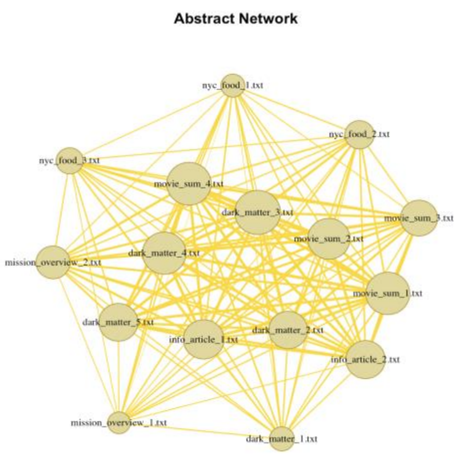
Confusion.Matrix	1	2	3	4	5
dark_matter	1	2	1	0	1
info_article	0	0	0	0	2
mission_overview	0	1	0	0	1
movie_summary	1	0	1	2	0
nyc_food	1	1	1	0	0

Importance of each node in document network.

	Degree	Betweenness	Closeness	Eigenvector
dark_matter_1.txt	15	0.333	0.00806	0.543
dark_matter_2.txt	15	0	0.00515	0.846
dark_matter_3.txt	15	0	0.00427	1
dark_matter_4.txt	15	0	0.0045	0.955
dark_matter_5.txt	15	0	0.0051	0.854

info_article_1.txt	15	0	0.00505	0.889
info_article_2.txt	15	0	0.00513	0.861
mission_overview_1.txt	15	23.5	0.00885	0.493
mission_overview_2.txt	15	0	0.00592	0.743
movie_sum_1.txt	15	0	0.00463	0.959
movie_sum_2.txt	15	0	0.00465	0.924
movie_sum_3.txt	15	0	0.00532	0.811
movie_sum_4.txt	15	0	0.00446	0.969
nyc_food_1.txt	15	8.67	0.00833	0.522
nyc_food_2.txt	15	0	0.00671	0.642
nyc_food_3.txt	15	0	0.0073	0.591

Abstract Network

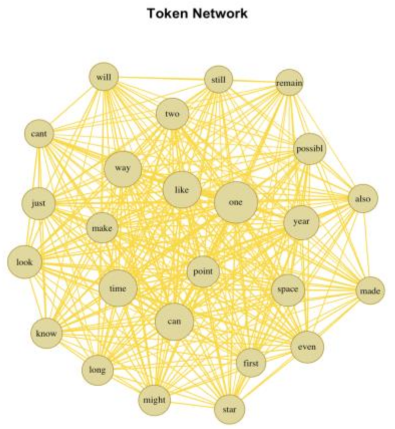


Importance of each node in token network

	Degree	Betweenness	Closeness	Eigenvector
also	24	0	0.00562	0.683
first	24	0	0.00559	0.686
made	24	0	0.00585	0.657
make	24	0	0.00526	0.73
one	24	0	0.00376	1
possibl	24	0	0.0051	0.75
remain	24	0	0.0061	0.632
space	24	0	0.005	0.768
star	24	0	0.00546	0.704
two	24	0	0.00505	0.758
way	24	0	0.00441	0.864
year	24	0	0.00465	0.82
can	24	0	0.00426	0.893

even	24	0	0.00503	0.762
just	24	0	0.00498	0.77
know	24	0	0.00529	0.726
like	24	0	0.00429	0.887
long	24	0	0.00532	0.724
might	24	0	0.00524	0.736
point	24	0	0.0051	0.751
still	24	0	0.00592	0.65
time	24	0	0.00431	0.881
cant	24	0	0.00575	0.671
look	24	0	0.00485	0.79
will	24	0	0.00575	0.671

Token Network

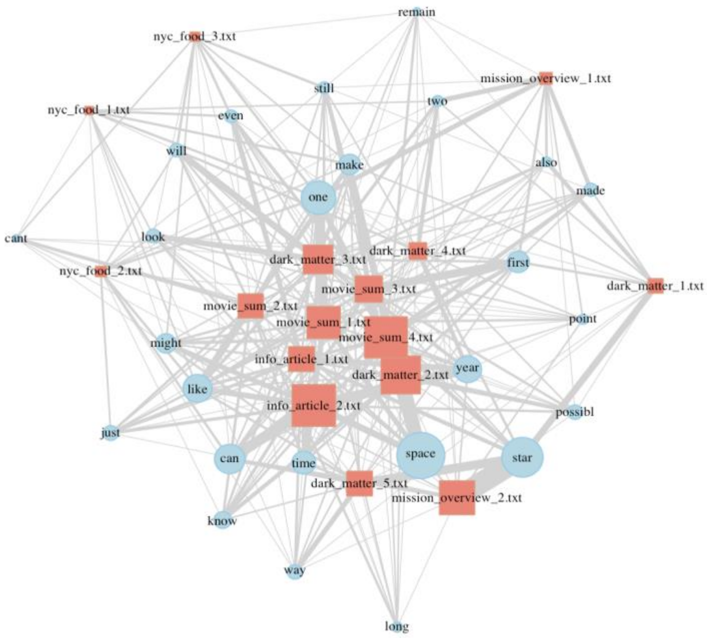


Importance of each node in bipartite network

	Degree	Betweenness	Closeness	Eigenvector
dark_matter_1.txt	12	20.3	0.0109	0.313
dark_matter_2.txt	19	25.4	0.0105	0.819
dark_matter_3.txt	23	10.4	0.0105	0.618
dark_matter_4.txt	22	73.9	0.0132	0.362
dark_matter_5.txt	19	70.5	0.0123	0.535
info_article_1.txt	20	47.8	0.0118	0.536
info_article_2.txt	19	16.4	0.0106	0.904
mission_overview_1.txt	11	17.4	0.0099	0.257
mission_overview_2.txt	16	58.8	0.0122	0.733
movie_sum_1.txt	22	39.6	0.0112	0.692
movie_sum_2.txt	21	48.3	0.0115	0.522
movie_sum_3.txt	18	21.8	0.0109	0.568
movie_sum_4.txt	22	8.24	0.0105	0.891
nyc_food_1.txt	11	12.8	0.0101	0.176

nyc_food_2.txt	14	30.2	0.0115	0.224
nyc_food_3.txt	13	27.9	0.0106	0.198
also	10	16.9	0.0115	0.215
first	10	0	0.0087	0.463
made	10	6.11	0.0098	0.284
make	12	0.548	0.00855	0.442
one	16	5.51	0.0106	0.72
possibl	11	32.6	0.0115	0.314
remain	10	50.5	0.0123	0.154
space	11	3.85	0.0099	1
star	10	1.57	0.00893	0.861
two	12	34	0.0122	0.239
way	13	34.4	0.0122	0.313
year	12	14.3	0.0111	0.578
can	13	3.86	0.00962	0.635
even	11	14.8	0.011	0.277
just	11	11.9	0.011	0.321
know	10	4.17	0.0103	0.362
like	13	0.699	0.00952	0.603
long	10	19.2	0.0118	0.218
might	11	1.19	0.00962	0.42
point	11	25.7	0.0115	0.217
still	10	5.92	0.0103	0.256
time	13	7.97	0.0109	0.495
cant	10	30.3	0.0118	0.157
look	12	20.4	0.012	0.297
will	10	3.26	0.0099	0.288

Bipartite Network



R CODE

```
#By Prakrit Dayal

setwd("/Users/prakrit/Desktop/Data Analytics/Assignment3")
library(dplyr)
library(tm); library(NLP) ; library(slam); library(SnowballC)
rm(list = ls())

cname = file.path(".", "CorpusAssign3")
dir(cname)

docs = Corpus(DirSource((cname)))

#fixing up the probelm quotations
toSpace <- content_transformer(function(x, pattern)
  gsub(pattern, '"', x))
docs <- tm_map(docs, toSpace, '”')
docs <- tm_map(docs, toSpace, '“')

#x-ray was coming as x ray, which is incorrect
toSpace <- content_transformer(function(x, pattern)
gsub(pattern, 'xray', x))
docs <- tm_map(docs, toSpace, 'X-ray')
docs <- tm_map(docs, toSpace, 'x—ray')

#remove dashes
toSpace <- content_transformer(function(x, pattern)
  gsub(pattern, ' ', x))
docs <- tm_map(docs, toSpace, '-')
docs <- tm_map(docs, toSpace, '—')

#remove other type  of problem quotations
toSpace <- content_transformer(function(x, pattern)
  gsub(pattern, "'", x))
docs <- tm_map(docs, toSpace, '’')
docs <- tm_map(docs, toSpace, '‘')
```

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

#creating base dtm
dtm <- DocumentTermMatrix(docs)

freq <- colSums(as.matrix(dtm))
ord1 = order(freq)
#freq[tail(ord1,15)]


dtms <- removeSparseTerms(dtm, 0.4) # rem. 40% empty
#dim(dtms)

#creating DTM matrix as readable format
freq2 <- colSums(as.matrix(dtms))
ord2 = order(freq2)
freq2[tail(ord2,15)]

#creating the DTM in readable format
dtm_final = as.data.frame(freq2[order(freq2,decreasing = TRUE)])
colnames(dtm_final) = "Frequencies"

#write.csv(freq2[order(freq2,decreasing = TRUE)], "dtms_Space.csv")

#calculating the cosine distance
distmatrix = proxy::dist(as.matrix(dtms), method = "cosine") #cosine distance
fit = hclust(distmatrix, method = "ward.D")
plot(fit)
plot(fit, hang = -1)
#clustering into 4 groups
rect.hclust(fit, k=5, border = "red")

topics = c("nyc_food","nyc_food","nyc_food",
        "movie_summary", "movie_summary", "movie_summary", "movie_summary",
        "dark_matter","dark_matter","dark_matter",
        "dark_matter", "dark_matter","mission_overview","mission_overview",
        "info_article", "info_article")
groups = cutree(fit, k = 5)
conf.mtx = table(GroupNames = topics, Clusters = groups)
accuracy = (conf.mtx[1,2]+conf.mtx[2,5]+conf.mtx[3,3]+conf.mtx[4,4]+conf.mtx[5,1])*100 /length(topics)
#write.csv(conf.mtx, "conf.mtx.csv")

 library(igraph); library(igraphdata)
#5
dtmsx = as.matrix(dtms)
dtmsx = as.matrix((dtmsx > 0 ) + 0)
ByAbsMatrix = dtmsx%*%t(dtmsx)
diag(ByAbsMatrix) = 0

#write.csv(ByAbsMatrix, "abs.mtx.csv")

ByAbs = graph_from_adjacency_matrix(ByAbsMatrix, mode="undirected", weighted = TRUE)
edge_widths_abs <- E(ByAbs)$weight/5 # edge width is a measure of edge weight
vertex_sizes_abs <- (evcent(ByAbs)$vector)*25 #vertex size is a measure of eig centrality

vertex_names_abs <- V(ByAbs)$media
plot(ByAbs,edge.width = edge_widths_abs,vertex.size = vertex_sizes_abs,
    edge.color="gold",vertex.frame.color= "#B59410",vertex.color="#E1D898",
    vertex.label=vertex_names_abs, vertex.label.color="black",vertex.label.cex = 0.8, main = "Abstract Network" )

##next line creates the df containing the degree, betweenness, closeness, and eig of all nodes in the network
summary.abs = cbind(as.data.frame(degree(ByAbs)),as.data.frame(betweenness(ByAbs))
        ,as.data.frame(closeness(ByAbs)),as.data.frame(evcent(ByAbs)$vector) )
colnames(summary.abs)= c("Degree", "Betweenness", "Closeness", "Eigenvector")
#round to 3 sigfigs
summary.abs = summary.abs %>%
  mutate_all(~signif(., digits = 3))
#write.csv(summary.abs, "abs.network.summary.csv")
summary.abs
```

```
#6
ByTokenMatrix = t(dtmsx)%*%dtmsx
diag(ByTokenMatrix) = 0
#write.csv(ByTokenMatrix, "token.mtx.csv")
ByToken = graph_from_adjacency_matrix(ByTokenMatrix, mode="undirected", weighted = TRUE)

edge_widths_token <- E(ByToken)$weight/5 # edge width is a measure of edge weight
vertex_sizes_token <- (evcent(ByToken)$vector)*25 #vertex size is a measure of eig centrality
vertex_names_token <- V(ByToken)$media


plot(ByToken,edge.width = edge_widths_token,vertex.size = vertex_sizes_token,
     edge.color="gold",vertex.frame.color= "#B59410",vertex.color="#E1D898",
     vertex.label=vertex_names_token, vertex.label.color="black" ,vertex.label.cex = 0.8, main = "Token Network" )
#plot(ByToken)

##next line creates the df containing the degree, betweenness, closeness, and eig of all nodes in the network
summary.token = cbind(as.data.frame(degree(ByToken)),as.data.frame(betweenness(ByToken))
                ,as.data.frame(closeness(ByToken)),as.data.frame(evcent(ByToken)$vector) )
colnames(summary.token)= c("Degree", "Betweenness", "Closeness", "Eigenvector")
#round to 3 sigfigs
summary.token = summary.token %>% mutate_all(~signif(., digits = 3))
#write.csv(summary.token, "token.network.summary.csv")
summary.token

#7
dtmsa = as.data.frame(as.matrix(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", "abs", "token")

dtmsc = dtmsb[dtmsb$weight!=0,]
dtmsc = dtmsc[,c(2,3,1)]
dtmsc
g <- graph.data.frame(dtmsc, directed=FALSE)

V(g)$type = bipartite_mapping(g)$type
V(g)$color = ifelse(V(g)$type, "lightblue", "salmon")
V(g)$frame.color = ifelse(V(g)$type, "skyblue", "darksalmon")
V(g)$shape = ifelse(V(g)$type, "circle", "square")
V(g)$size = (evcent(g)$vector)*15

E(g)$color = "lightgrey"
E(g)$width = as.numeric(E(g)$weight)*0.65

plot(g, vertex.label.color = "Black",vertex.label.cex = 0.65)

summary.bi = cbind(as.data.frame(degree(g)),as.data.frame(betweenness(g))
                 ,as.data.frame(closeness(g)),as.data.frame(evcent(g)$vector) )
colnames(summary.bi)= c("Degree", "Betweenness", "Closeness", "Eigenvector")
#round to 3 sigfigs
summary.bi = summary.bi %>% mutate_all(~signif(., digits = 3))
#write.csv(summary.bi, "bi.network.summary.csv")
summary.bi
```