Social Media Analytics for Canadian Banks v01

Chris Tan, Student ID 303428

22 July 2019

Abstract This is for the fulfillment of the York University's Advanced Analytics Course Capstone Project. The aim of this project is to uncover insights from the social media space through programmatic means.

Project Scope

Here are the boundaries of the project:

1. Social media channel: Twitter (to include Facebook if time permits)
2. Social media scope: Major Canadian Financial Institutions (FI) like BMO, CIBC, RBC, Scotiabank, TD (to include digital banks like Simplii, Tangerine, EQ if time permits)
3. Comparison of the following insights across the above FIs: Sentiment Analysis (polarity and categorical); Word Cloud (conversation drivers); Key-word dendrogram (blend of sentiment and conversation drivers); Network Analysis (demographics and product segmentation). Paraphrases of these insights are given in the "Research Questions" section below

Research Questions

Here are the research questions for this project:

1. Which bank has the most favourable / unfavourable trending opinion?
2. What are the current financial products being discussed?
3. What are the current emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) towards each bank?
4. What are the current sentiments towards trending financial product segments / categories (and the general network of terms being tweeted)?

Introduction and overall approach

Here is the general approach we adopted for Sprint#1:

1. Data Preparation
   1.1 Preliminaries (load libraries and set seed)
   1.2 Data Access
   1.3 Data extraction using twitteR
   1.4 Data Storage
2. Exploratory Data Analysis
   2.1 Create the term-document matrix
   2.2 Remove terms which have at least 98% of sparse elements
   2.3 Visualise term counts (or frequency)
   2.4 Data processing and normalization
      2.4.1 Remove stop words
      2.4.2 Remove terms which have at least 98% of sparse elements
      2.4.3 Visualise term counts (or frequency)
   2.5 Additional data processing and normalization
   2.6 Explore clustering of terms
   2.7 Plot the hierarchical clusters
   2.8 Nonhierarchical k-means clustering of words/tweets
   2.9 Find the pair of terms that appears frequently together
      2.9.1 Pair of terms that appears frequently together
      2.9.2 Find the network of terms
3. Observations and recommendation

Steps employed in Social Media (Twitter) Analytics using R

1. Data Preparation

1.1 Preliminaries

```r
options(warn=-1) # Suppress warnings to make output more readable

# This is tweets data extraction and other utilities specific to Twitter
suppressMessages(library(twitteR))

# Core data analytics packages like ggplot2, dplyr, tidyr, readr, purrr, tibb
le, stringr, forcats
suppressMessages(library(tidyverse))

# Primary text mining package
suppressMessages(library(tm))

set.seed(123)
```

1.2 Data Access

```r
# set the credentials
CONSUMER_SECRET <- 'Your CONSUMER_SECRET'
CONSUMER_KEY <- 'Your CONSUMER_KEY'
ACCESS_TOKEN <- 'Your ACCESS_TOKEN'
ACCESS_TOKEN_SECRET <- 'Your ACCESS_TOKEN_SECRET'

# Connect to twitter app. Select 2 in the console
setup_twitter_oauth(CONSUMER_KEY, CONSUMER_SECRET, ACCESS_TOKEN, ACCESS_TOKEN
_SECRET)
```

```
## [1] "Using direct authentication"
```

1.3 Data extraction using twitteR

```r
setwd("/Users/sgchr/Documents/CSDA1050/Data/")

# Get tweets separately for each bank for two reasons (1) Twitter's free deve
loper account has limits to search terms (2) Easier to do analysis by individ
ual banks later

# Get Cibc tweets
terms <- c("cibc","Canadian Imperial Bank of Commerce", "CanadianImperialBank
ofCommerce", "CIBCForTheFans")
terms_search <- paste(terms, collapse = " OR ") # Insert "OR between each ter
m
Cibc <- searchTwitter(terms_search, n=1000, since='2019-07-07', until='2019-0
7-22')

# Get Rbc tweets
terms <- c("rbc", "royal bank of canada", "royalbankofcanada")
terms_search <- paste(terms, collapse = " OR ") # Insert "OR between each ter
m
Rbc <- searchTwitter(terms_search, n=1000, lang="en", since='2019-07-07', unt
il='2019-07-22')
```

```
# Get Td tweets
#terms <- c("tdbank", "td bank", "tdgroup", "td group", "td bank group", "tdb
ankgroup", "td canada trust", "td canadatrust", "tdcanadatrust") ### no tweet
s
terms <- c("toronto dominion bank", "td bank", "TD bank", "TD Bank", "#tdbank
", "#TDBank")
terms_search <- paste(terms, collapse = " OR ") # Insert "OR between each ter
m
Td <- searchTwitter(terms_search, n=1000, lang="en", since='2019-07-07', unti
l='2019-07-22')

# Get Bmo tweets
terms <- c("bmo", "bank of montreal", "bankofmontreal", "bmoharris")
terms_search <- paste(terms, collapse = " OR ") # Insert "OR between each ter
m
Bmo <- searchTwitter(terms_search, n=1000, lang="en", since='2019-07-07', unt
il='2019-07-22')

# Get Bns tweets
terms <- c("scotiabank", "scotia bank")
terms_search <- paste(terms, collapse = " OR ") # Insert "OR between each ter
m
Bns <- searchTwitter(terms_search, n=1000, lang="en", since='2019-07-07', unt
il='2019-07-22')
```

1.4 Data storage

```
# Convert into dataframe for easier analysis later
Cibc_df <- twListToDF(Cibc) # rework logic for case when list is NULL
Rbc_df <- twListToDF(Rbc)
Td_df <- twListToDF(Td)
Bmo_df <- twListToDF(Bmo)
Bns_df <- twListToDF(Bns)

# Store the dataframed tweets
write.table(Cibc_df,"/Users/sgchr/Documents/CSDA1050/Data/Cibc.csv", append=T
, row.names=F, col.names=F,  sep=",")
write.table(Cibc_df,"/Users/sgchr/Documents/CSDA1050/Data/AllBanks.csv", appe
nd=T, row.names=F, col.names=F,  sep=",")

write.table(Rbc_df,"/Users/sgchr/Documents/CSDA1050/Data/Rbc.csv", append=T,
row.names=F, col.names=F,  sep=",")
write.table(Rbc_df,"/Users/sgchr/Documents/CSDA1050/Data/AllBanks.csv", appen
d=T, row.names=F, col.names=F,  sep=",")

write.table(Td_df,"/Users/sgchr/Documents/CSDA1050/Data/Td.csv", append=T, ro
w.names=F, col.names=F,  sep=",")
write.table(Td_df,"/Users/sgchr/Documents/CSDA1050/Data/AllBanks.csv", append
=T, row.names=F, col.names=F,  sep=",")
```

```
write.table(Bmo_df,"/Users/sgchr/Documents/CSDA1050/Data/Bmo.csv", append=T,
row.names=F, col.names=F,  sep=",")
write.table(Bmo_df,"/Users/sgchr/Documents/CSDA1050/Data/AllBanks.csv", appen
d=T, row.names=F, col.names=F,  sep=",")

write.table(Bns_df,"/Users/sgchr/Documents/CSDA1050/Data/Bns.csv", append=T,
row.names=F, col.names=F,  sep=",")
write.table(Bns_df,"/Users/sgchr/Documents/CSDA1050/Data/AllBanks.csv", appen
d=T, row.names=F, col.names=F,  sep=",")
```

2. Exploratory Data Analysis

Unlike spreasheets and typical databases, tweet contents are unstructured data. To aid in the analysis of unstructured data, we employ a technique to transform the tweets into structured data called the term-document matrix (or document-term matrix if we want the document to be displayed in rows). Documents here are the tweets; and terms are the words in the tweets. Each element in the matrix represents the number of times a particular term (words in the tweets) appears in a particular document (the tweets).

2.1 Create the term-document matrix

```
# Load the archived tweets
AllBanks_csv <- read.csv("/Users/sgchr/Documents/CSDA1050/Data/AllBanks.csv",
header = TRUE)

# Build the term-document matrix corpus
AllBankstext <- iconv(AllBanks_csv$text, to = 'UTF-8')
corpus <- Corpus(VectorSource(AllBankstext))

# Create term document matrix. Inf or infinity means to ingest everything
tdm <- TermDocumentMatrix(corpus, control = list(minWordLength=c(1,Inf)))
inspect(tdm)
```

```
## <<TermDocumentMatrix (terms: 6519, documents: 2710)>>
## Non-/sparse entries: 37393/17629097
## Sparsity           : 100%
## Maximal term length: 77
## Weighting          : term frequency (tf)
## Sample             :
##           Docs
## Terms      1704 1770 1798 1908 1942 1954 2001 2022 2234 2355
##   and         3    0    3    5    5    1    4    0    0    1
##   bank       15   11   20    5    3   10    6   11   10   12
##   canada     14    9    1    0    0    1    1   11   10    2
##   for         0    0    2    7    6    5    7    0    0    3
##   from        2    0    4    1    0    1    1    0    3    1
##   montreal    0    0    0    0    0    0    0    0    0    6
##   royal      15   10    1    0    0    0    0   11   10    0
##   scotia      0    0    0    3    0    7    5    0    0    0
##   the         6    4    4    6    2    4    2    2    4    7
##   you         0    0    0    0    4    0    5    0    0    1
```

There are 5487 terms in 1623 documents or tweets. 100% sparsity means there are lots of terms occuring zero times in a document.

2.2 Remove terms which have at least 98% of sparse elements

```
t <- removeSparseTerms(tdm, sparse=0.98)

# Check sparsity
inspect(t)
```

```
## <<TermDocumentMatrix (terms: 89, documents: 2710)>>
## Non-/sparse entries: 14143/227047
## Sparsity           : 94%
## Maximal term length: 34
## Weighting          : term frequency (tf)
## Sample             :
##           Docs
## Terms      1704 1770 1798 1908 1954 2001 2022 2234 2327 2355
##   and         3    0    3    5    1    4    0    0    0    1
##   bank       15   11   20    5   10    6   11   10    7   12
##   canada     14    9    1    0    1    1   11   10    0    2
##   for         0    0    2    7    5    7    0    0    1    3
##   from        2    0    4    1    1    1    0    3    2    1
##   montreal    0    0    0    0    0    0    0    0    7    6
##   royal      15   10    1    0    0    0   11   10    0    0
##   scotia      0    0    0    3    7    5    0    0    0    0
##   the         6    4    4    6    4    2    2    4    3    7
##   you         0    0    0    0    0    5    0    0    2    1
```

Number of terms has dropped to 82 and sparsity has dropped to 94%. We can experiment with sparse=? values to make sure we have sufficient term counts for analysis

2.3 Visualise term counts (or frequency)

```
# Convert into matrix for further analysis
m <- as.matrix(t)

# Plot frequent terms
freq <- rowSums(m) # Count number of times each of the 82 terms appears

# It will be a very busy group if we plot all 82 terms, so we restrict any te
rms that appears more > 24x
freq <- subset(freq, freq>=25)

# Visualise
barplot(freq,
        las=2, # list all words vertically
        col = rainbow(25))
```

There are many stop words that can be removed

2.4 Data processing and normalization

2.4.1 Remove stop words

```
corpus <- tm_map(corpus, removeWords, stopwords(kind="en"))
tdm <- TermDocumentMatrix(corpus, control = list(minWordLength=c(1,Inf)))
inspect(tdm)
```

```
## <<TermDocumentMatrix (terms: 6382, documents: 2710)>>
## Non-/sparse entries: 33090/17262130
## Sparsity           : 100%
## Maximal term length: 76
## Weighting          : term frequency (tf)
## Sample             :
##           Docs
## Terms      1704 1731 1770 1798 1954 2001 2022 2234 2327 2355
##   …           1    0    0    0    2    0    0    0    0    0
##   bank       15    6   11   20   10    6   11   10    7   12
##   canada     14    6    9    1    1    1   11   10    0    2
##   form        3    0    0    5    3    2    0    6    3    1
##   montreal    0    0    0    0    0    0    0    0    7    6
##   price       0    0    2    1    0    0    3    0    0    0
##   royal      15    6   10    1    0    0   11   10    0    0
##   scotia      0    0    0    0    7    5    0    0    0    0
##   see         0    0    0    0    3    2    0    0    0    0
##   the         2    0    0    1    1    0    1    3    1    2
```

Term count reduced from 5487 to 5352 (135 stop words have been removed)

2.4.2 Remove terms which have at least 98% of sparse elements

```
t <- removeSparseTerms(tdm, sparse=0.98)

# Check sparsity
inspect(t)
```

```
## <<TermDocumentMatrix (terms: 71, documents: 2710)>>
## Non-/sparse entries: 10838/181572
## Sparsity           : 94%
## Maximal term length: 34
## Weighting          : term frequency (tf)
## Sample             :
##           Docs
## Terms      1704 1770 1798 1954 2001 2022 2234 2327 2337 2355
##   …           1    0    0    2    0    0    0    0    0    0
##   bank       15   11   20   10    6   11   10    7    7   12
##   canada     14    9    1    1    1   11   10    0    0    2
##   form        3    0    5    3    2    0    6    3    0    1
##   montreal    0    0    0    0    0    0    0    7    4    6
##   price       0    2    1    0    0    3    0    0    0    0
##   royal      15   10    1    0    0   11   10    0    0    0
##   scotia      0    0    0    7    5    0    0    0    0    0
```

```
##   see          0    0    0    3    2    0    0    0    0    0
##   the          2    0    1    1    0    1    3    1    0    2
```

Term count has reduced to 64 (and with stop words also removed via the precedign code)

2.4.3 Visualise term counts (or frequency)

```
# Convert into matrix for further analysis
m <- as.matrix(t)

# Plot frequent terms
freq <- rowSums(m) # Count number of times each of the 64 terms appears

# It will be a very busy group if we plot all 64 terms, so we restrict any te
rms that appears more > 24x
freq <- subset(freq, freq>=25)

# Visualise
barplot(freq,
        las=2, # list all words vertically
        col = rainbow(25))
```

It is evident that more text cleaning are still needed

2.5 Additional data processing and normalization

```
corpus <- tm_map(corpus, tolower) # Not crucial for text analytics but good p
ractice to do so
corpus <- tm_map(corpus, removePunctuation) # Remove punctuations
corpus <- tm_map(corpus, removeNumbers) # Remove numbers

# Remove URL
removeURL <- function(x) gsub('http[[:alnum:]]*', '', x)
corpus <- tm_map(corpus, content_transformer(removeURL))

# Remove words
corpus <- tm_map(corpus, removeWords, c('bank', 'the', '$td', '424b2', '…', '
$ry', 'fwp', 'amp'))

# Replace words
corpus <- tm_map(corpus, gsub, pattern = '#stocks', replacement = 'stock')
corpus <- tm_map(corpus, gsub, pattern = 'stocks', replacement = 'stock')

corpus <- tm_map(corpus, stripWhitespace) # remove leftover from the precedin
g removal

# Repeat the preceding codes
tdm <- TermDocumentMatrix(corpus, control = list(minWordLength=c(1,Inf)))
t <- removeSparseTerms(tdm, sparse=0.98)
m <- as.matrix(t)
freq <- rowSums(m)
freq <- subset(freq, freq>=25)
```

```
barplot(freq,
        las=2, # list all words vertically
        col = rainbow(25))
```

The text is mostly cleaned up

2.6 Explore clustering of terms

```
# Hierarchical word/tweet clustering using dendrogram
# Find the distance, use scale to normalise the matrix
distance <- dist(scale(m))

# Print the terms, and calculate the distance between the words in each docum
ent (tweets)
print(distance, digits = 2)
```

```
##                   canada price dominion toronto stock  form filing   sec
## price              141.9
## dominion           150.9  67.8
## toronto            151.6  81.7     46.6
## stock              148.7  74.5     69.1    82.7
## form               149.6  78.0     58.9    74.3  78.6
## filing             147.8  68.7     49.9    67.4  69.9  36.7
## sec                147.8  68.7     49.9    67.4  69.9  36.7    0.0
## canadian           154.3  73.3     67.1    81.0  78.1  73.0   66.1  66.1
## new                151.3  71.6     62.1    77.0  72.9  73.0   63.0  63.0
## news               149.7  62.8     49.4    67.1  63.5  62.1   50.5  50.5
## …                  162.3 106.9     98.3   106.5 107.8 106.9  100.3 100.3
## energy             143.3  59.1     51.3    68.6  62.5  64.4   52.8  52.8
## sector             144.5  63.3     49.8    67.4  62.5  63.2   51.3  51.3
## investment         149.0  67.8     55.6    71.6  67.0  67.6   56.7  56.7
## research           149.9  66.7     54.2    70.8  65.8  66.5   55.3  55.3
## royal               65.1 132.9    143.5   144.5 141.1 141.8  140.0 140.0
## target             140.4  26.7     63.5    78.1  70.9  74.3   64.5  64.5
## rating             139.3  77.2     66.8    80.8  68.6  77.0   67.6  67.6
## perform            144.2  62.2     48.7    66.6  61.4  62.0   49.9  49.9
## analysts           146.0  54.1     50.5    67.9  61.0  63.5   51.7  51.7
## cut                144.5  60.7     51.0    68.4  63.9  63.9   52.2  52.2
## rbc                147.3  83.3     71.7    83.8  83.7  83.1   74.5  74.5
## opportunity        141.8  66.9     51.8    61.6  68.1  66.6   55.5  55.5
## will               148.9  69.2     57.3    72.8  70.3  68.9   58.2  58.2
## first              151.7  71.5     60.1    75.4  72.6  71.4   61.1  61.1
## see                157.1  78.2     67.4    81.3  79.3  78.1   68.8  68.8
## canadas            149.4  61.9     48.3    66.1  63.2  61.7   49.5  49.5
## bmo                156.6  77.5     67.6    80.8  73.8  71.9   64.6  64.6
## scotia             191.9 138.3    133.2   135.2 138.5 135.0  131.8 131.8
## files              148.0  64.5     51.6    68.8  65.7  45.3   52.7  52.7
## account            158.6  81.6     71.8    85.0  82.6  81.4   72.6  72.6
## money              155.7  80.8     70.9    84.2  81.8  80.6   71.7  71.7
## spotify            155.5  75.0     64.2    78.7  76.0  74.8   65.1  65.1
## imperial           149.3  61.5     53.2    70.0  66.7  60.5   52.0  52.0
```

```
## announced            149.0  60.2     46.0    64.7  61.5  60.0   47.3  47.3
## scotiabank           179.4 117.4    110.8   119.7 118.0 117.2  111.3 111.3
## alone                152.6  68.8     56.9    72.8  69.9  68.6   57.9  57.9
## charges              152.8  69.6     57.9    73.6  70.8  69.4   58.9  58.9
## strange              153.0  69.5     57.7    73.5  70.7  69.4   58.8  58.8
## bns                  149.9  61.7     49.2    67.0  62.7  55.1   45.4  45.4
## nova                 153.7  72.3     61.9    76.8  72.7  65.8   58.9  58.9
## montreal             176.9 115.1    108.7   115.7 112.4 110.7  106.8 106.8
## road                 149.8  62.1     48.6    66.6  63.4  61.9   49.8  49.8
## venture              149.2  60.9     47.0    65.4  62.2  60.7   48.2  48.2
## andrzejxa            148.8  59.7     45.4    64.3  61.0  59.5   46.7  46.7
## iqcorp               148.9  59.9     45.7    64.5  61.2  59.7   47.0  47.0
## joint                148.9  59.9     45.7    64.5  61.2  59.7   47.0  47.0
## mavennet             148.9  59.9     45.7    64.5  61.2  59.7   47.0  47.0
## stable               148.8  59.9     45.6    64.4  61.2  59.7   46.9  46.9
## approx               149.0  60.2     46.1    64.8  61.5  60.0   47.4  47.4
## cambridge…           148.9  60.0     45.9    64.6  61.3  59.8   47.1  47.1
## continuing           149.0  60.2     46.1    64.8  61.5  60.0   47.4  47.4
## hespeler             149.7  61.9     48.3    66.4  63.2  61.7   49.5  49.5
## investigate          149.1  60.6     46.6    65.1  61.9  60.4   47.9  47.9
## occurred             149.0  60.2     46.1    64.8  61.5  60.0   47.4  47.4
## robbery              150.3  63.4     50.2    67.8  64.6  63.2   51.4  51.4
## wrpstoday            150.2  63.2     49.9    67.5  64.4  63.0   51.1  51.1
## href                 147.4  73.3     61.4    71.8  74.6  72.2   62.5  62.5
## truenana             144.4  62.1     47.6    61.8  63.4  61.6   49.5  49.5
## relnofollowtwitter   149.5  64.3     51.3    67.4  65.5  63.9   52.4  52.4
## falsenana            148.2  62.7     49.2    65.8  64.1  61.8   50.0  50.0
##                     canadian   new  news     … energy sector investment
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new                    69.8
## news                   61.5  56.0
## …                     104.4 100.8  96.3
## energy                 63.7  58.5  44.7  97.6
## sector                 62.5  57.1  42.9  96.8   42.1
## investment             64.8  62.0  49.3  99.1   51.6   50.1
## research               65.8  59.2  47.7  99.0   50.1   48.5       34.7
## royal                 146.5 144.4 141.9 160.2  135.1  136.3      141.4
## target                 68.6  67.6  58.1 103.6   54.1   58.7       63.5
## rating                 75.2  72.2  61.6 106.2   61.2   59.2       64.7
## perform                61.3  55.8  41.2  96.0   40.8   12.0       48.6
## analysts               60.2  57.2  43.4  97.0   45.2   43.6       50.5
## cut                    62.1  57.9  44.0  96.7   46.5   41.9       50.9
## rbc                    81.9  78.6  69.1 108.6   70.7   69.1       73.3
```

```
## opportunity             66.1  61.0  48.0  97.2   50.4   48.8       53.3
## will                    67.8  63.5  51.1  97.8   53.3   51.8       56.6
## first                   70.2  66.1  54.3  99.6   56.4   55.0       60.1
## see                     77.4  73.3  55.0 103.2   64.7   63.5       68.0
## canadas                 59.4  55.5  40.8  95.8   43.5   41.7       48.3
## bmo                     76.7  72.9  62.5 102.3   64.3   63.1       61.7
## scotia                 132.1 134.7 129.1 148.6  131.5  130.9      133.2
## files                   60.5  58.4  44.7  97.5   47.2   45.5       51.6
## account                 80.8  76.2  59.6 100.7   68.7   67.5       71.8
## money                   80.0  75.9  66.0 105.0   67.7   66.6       70.8
## spotify                 74.2  69.7  50.1 101.6   60.7   59.4       64.1
## imperial                47.7  57.9  46.5  98.2   48.9   47.3       50.3
## announced               58.7  53.5  38.1  94.7   41.0   39.1       45.3
## scotiabank             115.8 111.0 106.7 134.5  108.8  108.1      109.8
## alone                   68.0  63.1  40.3  97.7   52.9   51.4       56.8
## charges                 68.8  64.0  41.8  97.6   53.9   52.5       57.8
## strange                 68.7  63.9  41.6  98.1   53.8   52.3       57.7
## bns                     61.6  56.2  41.8  96.3   44.5   42.7       49.1
## nova                    71.2  65.7  54.0 101.6   58.3   56.9       61.9
## montreal               114.1 109.8 105.6 132.3  106.7  105.9      105.1
## road                    61.2  55.7  41.1  93.9   43.8   42.0       48.5
## venture                 60.0  54.3  39.2  94.7   42.1   40.2       46.9
## andrzejxa               58.7  53.0  37.3  94.4   40.3   38.3       45.3
## iqcorp                  59.0  53.3  37.7  94.6   40.6   38.7       45.6
## joint                   59.0  53.3  37.7  94.6   40.6   38.7       45.6
## mavennet                59.0  53.3  37.7  94.6   40.6   38.7       45.6
## stable                  58.9  53.2  37.6  94.5   40.5   38.6       45.6
## approx                  59.3  53.6  38.1  94.7   41.1   39.1       46.0
## cambridge…              59.1  53.4  37.8  94.6   40.8   38.8       45.8
## continuing              59.3  53.6  38.1  94.7   41.1   39.1       46.0
## hespeler                61.0  55.5  40.8  94.0   43.6   41.7       48.3
## investigate             59.7  54.0  38.7  95.0   41.6   39.7       46.5
## occurred                59.3  53.6  38.1  94.7   41.1   39.1       46.0
## robbery                 62.5  57.2  43.0  95.0   45.6   43.9       50.1
## wrpstoday               62.3  56.9  42.7  94.8   45.3   43.6       49.8
## href                    70.9  68.0  57.0 101.9   59.0   57.5       61.3
## truenana                61.1  55.8  41.2  94.4   44.0   42.1       47.2
## relnofollowtwitter      61.4  57.9  44.4  96.8   47.0   45.3       51.4
## falsenana               61.3  56.4  42.4  96.1   45.0   43.1       49.4
##                    research royal target rating perform analysts   cut
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
```

```
## …
## energy
## sector
## investment
## research
## royal                142.1
## target                62.3 131.3
## rating                63.5 130.4   73.5
## perform               47.0 136.1   57.5   58.0
## analysts              48.9 137.8   48.6   59.9    41.9
## cut                   49.4 136.3   55.8   62.9    40.1     40.3
## rbc                   72.6 140.5   79.8   82.2    68.6     69.9  69.0
## opportunity           53.1 133.6   62.5   65.8    47.3     49.2  49.7
## will                  55.9 141.2   64.9   68.1    50.4     52.2  43.0
## first                 58.8 144.4   67.5   70.5    53.6     55.3  50.5
## see                   66.8 149.7   74.5   77.3    62.3     63.8  64.2
## canadas               46.6 141.0   57.2   60.7    39.9     42.2  42.8
## bmo                   60.5 149.6   73.8   75.2    61.9     61.8  63.8
## scotia               132.6 188.0  136.2  137.8   130.4    130.6 130.7
## files                 50.0 140.1   60.0   63.4    43.9     45.9  46.5
## account               70.7 151.3   78.1   80.7    66.5     67.8  68.2
## money                 69.7 151.1   77.2   79.9    65.5     66.9  67.2
## spotify               62.9 147.8   71.1   74.0    58.1     59.7  60.1
## imperial              51.7 142.5   57.2   63.2    45.7     45.7  48.2
## announced             44.3 141.1   55.3   58.9    37.2     39.6  40.2
## scotiabank           110.1 173.2  114.3  116.7   107.4    108.3 108.5
## alone                 55.4 145.0   64.5   67.7    49.9     51.7  52.2
## charges               56.4 145.4   65.4   68.6    51.1     52.8  53.3
## strange               56.3 145.4   65.3   68.5    50.9     52.7  53.2
## bns                   47.5 142.2   56.9   61.4    41.0     41.9  43.7
## nova                  60.6 147.1   68.2   71.4    55.6     56.3  57.7
## montreal             101.1 172.0  112.6  113.6   105.3    105.2 106.4
## road                  46.9 142.0   57.4   60.9    40.3     42.5  43.1
## venture               45.2 141.4   56.0   59.7    38.3     40.7  41.3
## andrzejxa             43.6 140.9   54.7   58.4    36.4     38.8  39.4
## iqcorp                43.9 141.0   55.0   58.7    36.8     39.2  39.8
## joint                 43.9 141.0   55.0   58.7    36.8     39.2  39.8
## mavennet              43.9 141.0   55.0   58.7    36.8     39.2  39.8
## stable                43.8 141.0   54.9   58.6    36.6     39.1  39.7
## approx                44.3 141.1   55.3   59.0    37.2     39.6  40.2
## cambridge…            44.1 141.1   55.1   58.8    36.9     39.3  40.0
## continuing            44.3 141.1   55.3   59.0    37.2     39.6  40.2
## hespeler              46.6 141.9   57.2   60.7    40.0     42.2  42.8
## investigate           44.8 141.3   55.7   59.4    37.9     40.2  40.8
## occurred              44.3 141.1   55.3   59.0    37.2     39.6  40.2
## robbery               48.6 142.5   58.8   62.2    42.2     44.3  44.9
## wrpstoday             48.3 142.4   58.5   62.0    41.9     44.0  44.6
## href                  61.2 141.0   69.3   72.3    56.3     57.7  58.4
## truenana              47.0 137.1   57.3   60.9    40.4     42.5  43.2
## relnofollowtwitter    49.8 142.4   59.7   63.2    43.6     45.5  46.2
```

```
## falsenana              47.8 141.0   58.1   61.6    41.4     43.4  44.2
##                      rbc opportunity  will first   see canadas   bmo
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
## …
## energy
## sector
## investment
## research
## royal
## target
## rating
## perform
## analysts
## cut
## rbc
## opportunity         68.4
## will                74.6        54.0
## first               76.9        59.0  57.1
## see                 83.2        67.1  69.1  68.1
## canadas             68.3        46.6  50.1  41.4  62.1
## bmo                 81.6        66.1  68.1  67.8  77.8    61.6
## scotia             140.9       132.7 131.8 133.3 131.2   130.3 137.3
## files               70.7        50.3  53.3  56.3  64.7    43.5  61.3
## account             85.7        70.5  72.1  75.3  56.5    66.2  81.4
## money               85.5        70.0  70.5  74.4  80.9    65.2  80.6
## spotify             80.3        63.1  65.3  68.0  45.2    57.8  74.7
## imperial            71.8        52.0  54.3  57.8  65.9    45.3  65.5
## announced           66.7        44.6  47.3  39.5  60.3    16.7  59.9
## scotiabank         116.5       110.2 110.3 111.7 117.4   106.3 116.7
## alone               74.6        55.7  58.2  61.2  37.3    49.6  68.5
## charges             75.4        56.7  59.2  61.5  37.9    50.7  69.4
## strange             75.3        56.6  59.0  62.0  35.9    50.6  69.3
## bns                 68.9        47.8  50.9  54.1  62.7    40.5  62.3
## nova                78.5        60.8  63.3  62.2  73.1    55.3  72.8
## montreal           118.9       108.1 107.2 110.7 115.5   104.8  96.4
## road                68.5        47.2  50.3  53.5  62.3    39.8  61.8
## venture             67.4        45.6  48.8  40.6  61.0    19.2  60.6
## andrzejxa           66.3        43.9  47.3  39.5  59.8    16.7  59.4
## iqcorp              66.5        44.3  47.6  39.1  60.1    15.8  59.6
## joint               66.5        44.3  47.6  39.1  60.1    15.8  59.6
## mavennet            66.5        44.3  47.6  39.1  60.1    15.8  59.6
```

```
## stable               66.4        44.2  47.5  39.2  60.0    15.9  59.5
## approx               66.8        44.7  48.0  51.3  60.4    36.8  59.9
## cambridge…           66.6        44.4  47.7  51.1  60.2    36.4  59.7
## continuing           66.8        44.7  48.0  51.3  60.4    36.8  59.9
## hespeler             68.3        47.0  50.1  53.3  62.1    39.5  61.6
## investigate          67.1        45.2  48.4  51.8  60.7    37.4  60.3
## occurred             66.8        44.7  48.0  51.3  60.4    36.8  59.9
## robbery              69.7        48.9  51.9  55.0  63.5    41.8  63.1
## wrpstoday            69.4        48.6  51.6  54.8  61.8    41.5  62.9
## href                 76.4        55.7  62.7  63.5  73.5    52.8  71.9
## truenana             66.2        40.3  50.0  53.6  62.3    39.9  61.2
## relnofollowtwitter   69.9        49.9  52.1  52.6  64.3    39.1  63.4
## falsenana            68.8        47.5  50.3  51.2  62.9    37.1  61.7
##                     scotia files account money spotify imperial announced
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
## …
## energy
## sector
## investment
## research
## royal
## target
## rating
## perform
## analysts
## cut
## rbc
## opportunity
## will
## first
## see
## canadas
## bmo
## scotia
## files               130.0
## account             137.2  68.7
## money               137.9  67.7    81.3
## spotify             136.9  60.6    49.7  77.4
## imperial            132.1  44.6    69.8  68.9    62.0
## announced           129.4  41.0    64.6  63.5    55.9     42.2
## scotiabank          162.4 108.8   116.9 114.1   115.3    109.5     106.2
```

```
## alone                133.7  52.8    43.9  71.8    29.6     54.3      47.4
## charges              134.1  53.9    44.5  72.6    27.6     55.4      48.6
## strange              134.0  53.8    43.5  72.5    27.5     55.3      48.4
## bns                  123.7  39.9    66.8  65.8    58.5     46.2      37.8
## nova                 117.8  54.8    76.7  75.7    69.6     59.6      53.3
## montreal             162.3 104.9   117.8 117.2   113.3    107.4     103.9
## road                 124.2  43.8    66.4  65.4    58.0     45.6      37.0
## venture              129.8  42.0    65.2  64.2    56.7     43.9      12.1
## andrzejxa            129.2  40.2    64.1  63.1    55.4     42.2       7.7
## iqcorp               129.3  40.6    64.3  63.3    55.7     42.5       5.4
## joint                129.3  40.6    64.3  63.3    55.7     42.5       5.4
## mavennet             129.3  40.6    64.3  63.3    55.7     42.5       5.4
## stable               129.3  40.5    64.3  63.2    55.6     42.5       8.4
## approx               125.0  41.0    64.6  63.6    56.0     43.0      33.7
## cambridge…           125.0  40.7    64.4  63.4    55.8     42.7      33.4
## continuing           125.0  41.0    64.6  63.6    56.0     43.0      33.7
## hespeler             124.1  43.5    66.2  65.2    57.8     45.3      36.7
## investigate          125.1  41.6    65.0  64.0    56.4     43.5      34.4
## occurred             125.0  41.0    64.6  63.6    56.0     43.0      33.7
## robbery              124.9  45.6    67.6  66.6    59.4     47.3      39.1
## wrpstoday            124.0  45.3    67.4  66.4    59.2     47.0      38.8
## href                 134.5  58.7    77.1  76.0    70.2     60.1      51.4
## truenana             130.2  43.8    66.4  65.1    58.2     45.6      37.1
## relnofollowtwitter   129.9  46.8    68.4  67.1    60.4     48.5      37.0
## falsenana            129.7  44.7    67.2  66.2    58.9     46.6      34.6
##                   scotiabank alone charges strange   bns  nova montreal
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
## …
## energy
## sector
## investment
## research
## royal
## target
## rating
## perform
## analysts
## cut
## rbc
## opportunity
## will
```

```
## first
## see
## canadas
## bmo
## scotia
## files
## account
## money
## spotify
## imperial
## announced
## scotiabank
## alone                  111.3
## charges                111.9  15.8
## strange                111.8  10.1    12.2
## bns                    107.6  50.4    51.5    51.4
## nova                   113.9  62.9    63.8    63.7  37.5
## montreal               144.4 109.3   109.8   109.7 105.5 112.0
## road                   107.1  49.8    51.0    50.8  40.8  55.5    105.2
## venture                106.6  48.3    49.5    49.3  38.9  54.1    103.2
## andrzejxa              106.0  46.7    48.0    47.8  37.0  52.8    103.6
## iqcorp                 106.1  47.0    48.3    48.1  37.4  53.0    103.8
## joint                  106.1  47.0    48.3    48.1  37.4  53.0    103.8
## mavennet               106.1  47.0    48.3    48.1  37.4  53.0    103.8
## stable                 106.1  47.0    48.2    48.0  37.3  53.0    103.6
## approx                 106.3  47.4    48.6    48.5  37.9  53.4    104.1
## cambridge…             106.1  47.2    48.4    48.2  37.6  53.2    104.0
## continuing             106.3  47.4    48.6    48.5  37.9  53.4    104.1
## hespeler               107.2  49.6    50.7    50.6  40.6  55.3    105.1
## investigate            106.0  47.9    49.1    49.0  38.5  53.8    104.3
## occurred               106.3  47.4    48.6    48.5  37.9  53.4    104.1
## robbery                107.7  51.4    52.5    52.4  42.8  57.0    106.0
## wrpstoday              106.9  51.1    52.3    52.1  42.4  56.7    105.8
## href                   112.6  63.7    64.5    64.5  56.5  67.5    109.7
## truenana               107.1  50.0    51.1    51.0  40.8  55.4    104.2
## relnofollowtwitter     107.0  52.6    53.7    53.6  43.8  57.4    105.3
## falsenana              107.0  51.0    52.0    51.9  41.9  56.1    104.1
##                     road venture andrzejxa iqcorp joint mavennet stable
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
## …
## energy
```

```
## sector
## investment
## research
## royal
## target
## rating
## perform
## analysts
## cut
## rbc
## opportunity
## will
## first
## see
## canadas
## bmo
## scotia
## files
## account
## money
## spotify
## imperial
## announced
## scotiabank
## alone
## charges
## strange
## bns
## nova
## montreal
## road
## venture             38.2
## andrzejxa           36.2    12.1
## iqcorp              36.6    10.8       5.4
## joint               36.6    10.8       5.4    0.0
## mavennet            36.6    10.8       5.4    0.0   0.0
## stable              36.5    12.6       5.2    6.4   6.4      6.4
## approx              15.3    35.0      32.8   33.3  33.3     33.3   33.1
## cambridge…          16.0    34.6      32.5   32.9  32.9     32.9   32.8
## continuing          15.3    35.0      32.8   33.3  33.3     33.3   33.1
## hespeler             4.9    37.9      35.9   36.3  36.3     36.3   36.2
## investigate         16.8    35.7      33.5   34.0  34.0     34.0   33.8
## occurred            15.3    35.0      32.8   33.3  33.3     33.3   33.1
## robbery             14.4    40.2      38.4   38.8  38.8     38.8   38.6
## wrpstoday           17.1    39.9      38.0   38.4  38.4     38.4   38.3
## href                56.3    52.3      51.0   51.1  51.1     51.1   50.8
## truenana            40.3    38.3      36.3   36.7  36.7     36.7   36.7
## relnofollowtwitter  43.4    38.2      36.4   36.6  36.6     36.6   36.1
## falsenana           41.3    35.9      34.0   34.2  34.2     34.2   33.7
##                    approx cambridge… continuing hespeler investigate
```

```
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
## …
## energy
## sector
## investment
## research
## royal
## target
## rating
## perform
## analysts
## cut
## rbc
## opportunity
## will
## first
## see
## canadas
## bmo
## scotia
## files
## account
## money
## spotify
## imperial
## announced
## scotiabank
## alone
## charges
## strange
## bns
## nova
## montreal
## road
## venture
## andrzejxa
## iqcorp
## joint
## mavennet
## stable
## approx
```

```
## cambridge…            4.8
## continuing            0.0        4.8
## hespeler             14.5       15.3       14.5
## investigate           6.9        8.4        6.9     16.1
## occurred              0.0        4.8        0.0     14.5         6.9
## robbery              19.9       20.5       19.9     13.6        18.6
## wrpstoday            20.1       19.7       20.1     16.3        21.3
## href                 54.2       54.1       54.2     56.1        54.6
## truenana             37.3       37.0       37.3     40.0        37.9
## relnofollowtwitter   40.6       40.5       40.6     43.1        41.2
## falsenana            38.4       38.3       38.4     41.0        39.0
##                    occurred robbery wrpstoday  href truenana
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
## …
## energy
## sector
## investment
## research
## royal
## target
## rating
## perform
## analysts
## cut
## rbc
## opportunity
## will
## first
## see
## canadas
## bmo
## scotia
## files
## account
## money
## spotify
## imperial
## announced
## scotiabank
## alone
## charges
```

```
## strange
## bns
## nova
## montreal
## road
## venture
## andrzejxa
## iqcorp
## joint
## mavennet
## stable
## approx
## cambridge…
## continuing
## hespeler
## investigate
## occurred
## robbery                19.9
## wrpstoday              20.1    21.3
## href                   54.2    57.7      57.4
## truenana               37.3    42.3      41.9  35.8
## relnofollowtwitter     40.6    45.2      44.9  32.5     39.6
## falsenana              38.4    43.2      42.9  33.6     39.2
##                    relnofollowtwitter
## price
## dominion
## toronto
## stock
## form
## filing
## sec
## canadian
## new
## news
## …
## energy
## sector
## investment
## research
## royal
## target
## rating
## perform
## analysts
## cut
## rbc
## opportunity
## will
## first
## see
```

```
## canadas
## bmo
## scotia
## files
## account
## money
## spotify
## imperial
## announced
## scotiabank
## alone
## charges
## strange
## bns
## nova
## montreal
## road
## venture
## andrzejxa
## iqcorp
## joint
## mavennet
## stable
## approx
## cambridge…
## continuing
## hespeler
## investigate
## occurred
## robbery
## wrpstoday
## href
## truenana
## relnofollowtwitter
## falsenana                          22.9
```

If distance is high, it means those two words should not be in the same cluster; likewise, if distance is low, they should be in the same cluster

2.7 Plot the hierarchical clusters

```
hc <- hclust(distance, method = "ward.D")
plot(hc, hang=-1)
rect.hclust(hc, k=10) # 10 clusters
```

2.8 Nonhierarchical k-means clustering of words/tweets

```
m1 <- t(m) # Transpose
k <- 10 # 10 clusters
```

```
kc <- kmeans(m1, k)
kc

## K-means clustering with 10 clusters of sizes 1412, 742, 97, 228, 8, 4, 126
, 1, 64, 28
##
## Cluster means:
##         canada      price    dominion     toronto      stock       form
## 1   0.03116147 0.01133144  0.00000000  0.02266289 0.03611898 0.04745042
## 2   1.05929919 0.20485175  0.02021563  0.10377358 0.07008086 0.09703504
## 3   0.04123711 0.06185567  0.00000000  0.00000000 0.00000000 0.19587629
## 4   0.14473684 0.00000000  0.00000000  0.00000000 0.00000000 0.00000000
## 5   1.00000000 0.00000000  0.00000000  0.00000000 0.50000000 1.37500000
## 6  11.25000000 1.25000000  0.00000000  0.50000000 0.00000000 2.25000000
## 7   1.00000000 0.00000000  0.00000000  0.00000000 0.27777778 0.00000000
## 8   1.00000000 1.00000000 11.00000000 11.00000000 0.00000000 5.00000000
## 9   0.06250000 0.00000000  0.93750000  1.06250000 0.00000000 0.51562500
## 10  0.64285714 0.07142857  0.07142857  0.14285714 0.03571429 0.10714286
##        filing        sec   canadian        new        news          …
## 1  0.02549575 0.02549575 0.00000000 0.02337110 0.037535411 0.00000000
## 2  0.06603774 0.06603774 0.02964960 0.02830189 0.005390836 0.07142857
## 3  0.08247423 0.08247423 1.00000000 0.09278351 0.010309278 0.05154639
## 4  0.00000000 0.00000000 0.00877193 0.04385965 0.004385965 1.00000000
## 5  0.87500000 0.87500000 0.00000000 0.25000000 0.125000000 1.25000000
## 6  1.50000000 1.50000000 0.25000000 0.50000000 0.250000000 0.50000000
## 7  0.00000000 0.00000000 0.00000000 0.00000000 0.000000000 0.01587302
## 8  3.00000000 3.00000000 0.00000000 1.00000000 0.000000000 0.00000000
## 9  0.39062500 0.39062500 0.00000000 0.00000000 0.000000000 0.12500000
## 10 0.07142857 0.07142857 0.28571429 0.03571429 0.000000000 0.14285714
##        energy      sector investment   research        royal      target
## 1  0.00000000 0.001416431 0.02691218 0.03186969  0.004249292 0.005665722
## 2  0.07277628 0.056603774 0.03773585 0.01886792  1.049865229 0.200808625
## 3  0.00000000 0.000000000 0.07216495 0.00000000  0.041237113 0.092783505
## 4  0.00000000 0.000000000 0.00000000 0.00000000  0.021929825 0.000000000
## 5  0.00000000 0.000000000 0.00000000 0.12500000  0.000000000 0.000000000
## 6  0.50000000 1.500000000 0.00000000 0.25000000 11.750000000 1.250000000
## 7  0.07936508 0.103174603 0.00000000 0.00000000  1.047619048 0.000000000
## 8  0.00000000 0.000000000 4.00000000 4.00000000  1.000000000 0.000000000
## 9  0.01562500 0.015625000 0.00000000 0.00000000  0.000000000 0.000000000
## 10 0.00000000 0.035714286 0.07142857 0.03571429  0.607142857 0.071428571
##        rating    perform    analysts         cut         rbc opportunity
## 1  0.01062323 0.00000000 0.009206799 0.002832861 0.024787535 0.001416431
## 2  0.00000000 0.05390836 0.056603774 0.082210243 0.163072776 0.133423181
## 3  0.04123711 0.00000000 0.051546392 0.000000000 0.010309278 0.000000000
## 4  0.00000000 0.00000000 0.000000000 0.004385965 0.039473684 0.000000000
## 5  0.00000000 0.00000000 0.250000000 0.000000000 0.000000000 0.000000000
## 6  1.25000000 1.25000000 0.250000000 1.000000000 0.250000000 1.000000000
## 7  1.01587302 0.10317460 0.047619048 0.000000000 0.007936508 0.000000000
## 8  1.00000000 0.00000000 1.000000000 0.000000000 1.000000000 1.000000000
## 9  0.00000000 0.00000000 0.000000000 0.000000000 0.000000000 0.000000000
```

```
## 10 0.14285714 0.03571429 0.071428571 0.035714286 0.035714286 0.035714286
##          will      first        see     canadas         bmo      scotia
## 1  0.01770538 0.06373938 0.11189802 0.047450425 0.075070822 0.275495751
## 2  0.04986523 0.02560647 0.00000000 0.002695418 0.002695418 0.005390836
## 3  0.02061856 0.01030928 0.01030928 0.041237113 0.020618557 0.288659794
## 4  0.07456140 0.05263158 0.13157895 0.000000000 0.070175439 0.228070175
## 5  0.62500000 0.25000000 0.62500000 0.125000000 1.625000000 1.625000000
## 6  0.00000000 0.00000000 0.00000000 0.250000000 0.000000000 0.000000000
## 7  0.00000000 0.00000000 0.00000000 0.000000000 0.000000000 0.000000000
## 8  0.00000000 0.00000000 0.00000000 0.000000000 0.000000000 0.000000000
## 9  0.00000000 0.00000000 0.03125000 0.000000000 0.000000000 0.031250000
## 10 0.10714286 0.17857143 0.03571429 0.142857143 0.142857143 0.571428571
##         files     account      money    spotify    imperial  announced
## 1  0.01841360 0.111898017 0.04957507 0.09773371 0.002832861 0.04390935
## 2  0.03099730 0.002695418 0.00000000 0.00000000 0.016172507 0.00000000
## 3  0.11340206 0.000000000 0.00000000 0.00000000 0.474226804 0.02061856
## 4  0.00000000 0.219298246 0.07456140 0.13157895 0.000000000 0.00000000
## 5  0.37500000 0.125000000 0.12500000 0.00000000 0.000000000 0.12500000
## 6  0.75000000 0.000000000 0.00000000 0.00000000 0.000000000 0.00000000
## 7  0.00000000 0.000000000 0.00000000 0.00000000 0.000000000 0.00000000
## 8  0.00000000 0.000000000 0.00000000 0.00000000 0.000000000 0.00000000
## 9  0.00000000 0.000000000 0.00000000 0.00000000 0.000000000 0.00000000
## 10 0.03571429 0.071428571 0.07142857 0.03571429 0.142857143 0.14285714
##    scotiabank      alone    charges   strange        bns       nova
## 1  0.14376771 0.08356941 0.08569405 0.0878187 0.03753541 0.06798867
## 2  0.00000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 3  0.06185567 0.00000000 0.00000000 0.0000000 0.00000000 0.05154639
## 4  0.07017544 0.10964912 0.13157895 0.1140351 0.00000000 0.03508772
## 5  0.37500000 0.00000000 0.12500000 0.0000000 0.50000000 0.75000000
## 6  0.00000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 7  0.00000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 8  0.00000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 9  0.00000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 10 0.57142857 0.00000000 0.00000000 0.0000000 0.07142857 0.10714286
##       montreal       road    venture andrzejxa     iqcorp      joint
## 1  0.164305949 0.05099150 0.04532578 0.0417847 0.04390935 0.04390935
## 2  0.004043127 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 3  0.051546392 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 4  0.096491228 0.06578947 0.00877193 0.0000000 0.00000000 0.00000000
## 5  4.250000000 0.00000000 0.12500000 0.1250000 0.12500000 0.12500000
## 6  0.000000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 7  0.000000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 8  0.000000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 9  0.000000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
## 10 0.428571429 0.03571429 0.14285714 0.1428571 0.14285714 0.14285714
##      mavennet     stable     approx cambridge… continuing   hespeler
## 1  0.04390935 0.04320113 0.04957507 0.04815864 0.04957507 0.05099150
## 2  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## 3  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## 4  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06140351
```

```
## 5  0.12500000 0.12500000 0.00000000 0.00000000 0.00000000 0.00000000
## 6  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## 7  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## 8  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## 9  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
## 10 0.14285714 0.14285714 0.03571429 0.00000000 0.03571429 0.03571429
##    investigate   occurred    robbery  wrpstoday        href    truenana
## 1   0.05099150 0.04957507 0.05382436 0.05382436  0.03257790 0.006373938
## 2   0.00000000 0.00000000 0.00000000 0.00000000  0.09433962 0.083557951
## 3   0.00000000 0.00000000 0.00000000 0.00000000  0.05154639 0.000000000
## 4   0.00000000 0.00000000 0.06140351 0.06140351  0.00877193 0.008771930
## 5   0.00000000 0.00000000 0.00000000 0.00000000  7.25000000 2.125000000
## 6   0.00000000 0.00000000 0.00000000 0.00000000 11.50000000 1.750000000
## 7   0.00000000 0.00000000 0.00000000 0.00000000  0.00000000 0.000000000
## 8   0.00000000 0.00000000 0.00000000 0.00000000 19.00000000 2.000000000
## 9   0.00000000 0.00000000 0.00000000 0.00000000  0.00000000 0.000000000
## 10  0.03571429 0.03571429 0.03571429 0.03571429  3.03571429 0.500000000
##    relnofollowtwitter    falsenana
## 1         0.028328612  0.023371105
## 2         0.004043127  0.009433962
## 3         0.051546392  0.020618557
## 4         0.004385965  0.000000000
## 5         3.250000000  4.750000000
## 6         1.500000000  9.250000000
## 7         0.000000000  0.000000000
## 8         1.000000000 16.000000000
## 9         0.000000000  0.000000000
## 10        2.035714286  1.821428571
##
## Clustering vector:
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##    1    1    1    9    1    9    9    9    9    1    9    1    9    9    1
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30
##    3    9    9    9    4    4    9    1    1    1    9    1    9    1    1
##   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45
##    1    1    1    1    1    1    9    1    1    1    2    2    2    2    2
##   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60
##    2    2    7    2    2    2    7    7    2    2    4    2    2    2    7
##   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75
##    2    2    2    7    2    7    7    2    2    2    2    2    2    2    2
##   76   77   78   79   80   81   82   83   84   85   86   87   88   89   90
##    4    2    2    2    2    2    2    2    1    2    3    1    2    2    2
##   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105
##    2    2    2    2    1    2    2    1    2    2    4    2    2    2    1
##  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120
##    2    2    7    2    2    7    7    2    2    2    2    2    2    7    1
##  121  122  123  124  125  126  127  128  129  130  131  132  133  134  135
##    1    1    2    2    2    2    2    2    7    7    7    2    2    7    2
##  136  137  138  139  140  141  142  143  144  145  146  147  148  149  150
##    2    2    2    2    2    2    1    7    2    2    2    2    7    2    2
```

```
##  151  152  153  154  155  156  157  158  159  160  161  162  163  164  165
##    2    2    2    1    2    7    2    2    2    2    2    2    2    7    2
##  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180
##    2    2    2    2    2    2    2    2    2    2    2    2    2    2    2
##  181  182  183  184  185  186  187  188  189  190  191  192  193  194  195
##    2    2    2    2    2    1    4    2    1    1    1    2    2    2    2
##  196  197  198  199  200  201  202  203  204  205  206  207  208  209  210
##    1    4    2    2    2    2    2    1    2    2    2    2    2    2    2
##  211  212  213  214  215  216  217  218  219  220  221  222  223  224  225
##    2    2    2    2    2    2    2    2    2    2    2    2    2    2    2
##  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240
##    2    2    3    2    7    7    2    7    2    2    2    2    2    7    2
##  241  242  243  244  245  246  247  248  249  250  251  252  253  254  255
##    2    2    2    2    2    2    2    2    2    2    2    2    2    2    2
##  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270
##    2    2    2    1    7    2    2    2    2    2    2    2    2    7    7
##  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285
##    2    2    2    2    2    2    2    2    2    2    1    2    2    2    2
##  286  287  288  289  290  291  292  293  294  295  296  297  298  299  300
##    2    2    2    2    2    2    2    2    7    2    2    2    2    2    2
##  301  302  303  304  305  306  307  308  309  310  311  312  313  314  315
##    2    2    2    2    2    2    2    2    2    2    1    2    2    2    2
##  316  317  318  319  320  321  322  323  324  325  326  327  328  329  330
##    2    2    2    2    2    2    2    2    2    2    2    1    2    2    2
##  331  332  333  334  335  336  337  338  339  340  341  342  343  344  345
##    2    2    2    2    2    2    2    2    2    2    2    2    2    3    2
##  346  347  348  349  350  351  352  353  354  355  356  357  358  359  360
##    2    1    1    2    1    3    3    3    3    3    3    3    3    3    3
##  361  362  363  364  365  366  367  368  369  370  371  372  373  374  375
##    3    3    3    3    3    3    1    1    4    3    3    3    3    3    3
##  376  377  378  379  380  381  382  383  384  385  386  387  388  389  390
##    3    3    3    3    1    1    4    1    1    1    1    1    1    4    1
##  391  392  393  394  395  396  397  398  399  400  401  402  403  404  405
##    1    1    4    1    1    1    4    1    4    4    1    1    1    1    1
##  406  407  408  409  410  411  412  413  414  415  416  417  418  419  420
##    1    1    4    1    1    1    1    1    1    4    1    4    1    1    4
##  421  422  423  424  425  426  427  428  429  430  431  432  433  434  435
##    4    1    4    1    1    4    4    4    4    1    1    1    1    1    1
##  436  437  438  439  440  441  442  443  444  445  446  447  448  449  450
##    1    1    1    1    1    1    3    1    1    4    4    1    1    1    1
##  451  452  453  454  455  456  457  458  459  460  461  462  463  464  465
##    1    1    4    1    1    1    1    1    1    1    1    1    1    1    1
##  466  467  468  469  470  471  472  473  474  475  476  477  478  479  480
##    1    1    1    1    1    4    1    1    1    1    1    1    1    1    1
##  481  482  483  484  485  486  487  488  489  490  491  492  493  494  495
##    1    1    4    1    4    1    1    4    4    1    1    1    4    1    1
##  496  497  498  499  500  501  502  503  504  505  506  507  508  509  510
##    1    1    1    1    1    1    4    3    1    4    1    1    1    1    4
##  511  512  513  514  515  516  517  518  519  520  521  522  523  524  525
##    1    1    1    1    1    1    1    1    3    1    1    1    1    1    1
```

```
##  526  527  528  529  530  531  532  533  534  535  536  537  538  539  540
##    1    1    1    1    1    1    4    1    1    1    1    4    4    1    1
##  541  542  543  544  545  546  547  548  549  550  551  552  553  554  555
##    1    4    1    1    1    1    1    4    1    1    1    1    4    1    1
##  556  557  558  559  560  561  562  563  564  565  566  567  568  569  570
##    1    1    1    1    4    1    1    1    4    4    4    4    1    1    1
##  571  572  573  574  575  576  577  578  579  580  581  582  583  584  585
##    1    1    1    1    1    4    4    1    4    1    1    1    1    1    1
##  586  587  588  589  590  591  592  593  594  595  596  597  598  599  600
##    4    1    1    1    1    1    1    1    4    1    1    1    1    1    1
##  601  602  603  604  605  606  607  608  609  610  611  612  613  614  615
##    4    4    1    1    1    4    1    1    4    1    4    4    1    4    1
##  616  617  618  619  620  621  622  623  624  625  626  627  628  629  630
##    1    1    1    1    1    1    1    1    4    1    1    1    1    1    1
##  631  632  633  634  635  636  637  638  639  640  641  642  643  644  645
##    1    1    4    1    1    1    1    1    1    1    1    1    1    1    1
##  646  647  648  649  650  651  652  653  654  655  656  657  658  659  660
##    3    4    1    1    1    1    4    1    1    1    1    1    1    1    4
##  661  662  663  664  665  666  667  668  669  670  671  672  673  674  675
##    1    4    4    1    1    4    1    4    1    1    1    1    1    3    1
##  676  677  678  679  680  681  682  683  684  685  686  687  688  689  690
##    1    1    1    4    1    1    1    1    1    1    1    1    1    1    1
##  691  692  693  694  695  696  697  698  699  700  701  702  703  704  705
##    1    1    4    4    1    1    1    1    1    1    1    1    1    1    4
##  706  707  708  709  710  711  712  713  714  715  716  717  718  719  720
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  721  722  723  724  725  726  727  728  729  730  731  732  733  734  735
##    1    1    1    4    1    1    1    1    1    1    4    1    1    3    4
##  736  737  738  739  740  741  742  743  744  745  746  747  748  749  750
##    1    4    1    1    3    1    1    1    1    1    3    1    1    3    3
##  751  752  753  754  755  756  757  758  759  760  761  762  763  764  765
##    3    4    1    3    1    1    1    1    1    1    1    1    1    3    1
##  766  767  768  769  770  771  772  773  774  775  776  777  778  779  780
##    4    4    1    4    1    4    1    4    1    1    1    1    4    1    4
##  781  782  783  784  785  786  787  788  789  790  791  792  793  794  795
##    1    4    4    4    1    1    4    4    1    1    1    1    1    1    1
##  796  797  798  799  800  801  802  803  804  805  806  807  808  809  810
##    1    1    1    1    1    1    1    1    1    1    1    4    1    4    4
##  811  812  813  814  815  816  817  818  819  820  821  822  823  824  825
##    1    1    1    1    4    1    4    1    1    4    1    1    1    1    1
##  826  827  828  829  830  831  832  833  834  835  836  837  838  839  840
##    1    1    1    1    1    1    1    1    3    1    1    1    1    1    1
##  841  842  843  844  845  846  847  848  849  850  851  852  853  854  855
##    1    4    1    1    1    1    1    1    1    1    1    1    1    1    1
##  856  857  858  859  860  861  862  863  864  865  866  867  868  869  870
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
##  871  872  873  874  875  876  877  878  879  880  881  882  883  884  885
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    4
##  886  887  888  889  890  891  892  893  894  895  896  897  898  899  900
##    1    1    1    1    4    1    1    1    4    1    1    1    1    1    2
```

```
##  901  902  903  904  905  906  907  908  909  910  911  912  913  914  915
##    2    4    1    1    4    1    1    4    3    1    1    3    9    9    9
##  916  917  918  919  920  921  922  923  924  925  926  927  928  929  930
##    1    9    9    9    9    9    9    1    1    9    2    1    2    2    2
##  931  932  933  934  935  936  937  938  939  940  941  942  943  944  945
##    7    2    7    2    1    2    2    2    2    2    2    2    2    2    7
##  946  947  948  949  950  951  952  953  954  955  956  957  958  959  960
##    2    2    2    2    2    2    2    2    2    2    2    2    1    4    1
##  961  962  963  964  965  966  967  968  969  970  971  972  973  974  975
##    1    1    1    1    1    1    1    1    1    1    1    4    1    1    1
##  976  977  978  979  980  981  982  983  984  985  986  987  988  989  990
##    1    1    1    1    1    1    1    1    3    3    3    1    1    1    3
##  991  992  993  994  995  996  997  998  999 1000 1001 1002 1003 1004 1005
##    1    3    1    1    1    1    1    4    1    2    1    1    1    4    1
## 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020
##    4    1    4    1    1    1    1    1    1    1    4    1    1    1    4
## 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035
##    1    1    1    1    1    1    4    1    2    1    1    1    1    1    1
## 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1050
##    1    1    1    1    1    1    1    1    1    1    1    1    1    4    1
## 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1062 1063 1064 1065
##    1    1    1    1    1    1    1    1    1    1    1    1    4    1    4
## 1066 1067 1068 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1079 1080
##    1    1    1    1    3    1    4    1    1    1    4    1    4    1    1
## 1081 1082 1083 1084 1085 1086 1087 1088 1089 1090 1091 1092 1093 1094 1095
##    1    4    1    1    1    4    1    1    1    4    4    4    1    4    4
## 1096 1097 1098 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1110
##    1    1    1    1    1    4    3    4    1    1    1    1    1    1    1
## 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125
##    1    4    1    1    4    1    1    3    1    3    1    1    1    1    1
## 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140
##    1    1    4    1    1    1    1    4    4    1    1    1    1    1    1
## 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155
##    1    1    1    1    4    1    1    1    1    4    1    1    1    1    4
## 1156 1157 1158 1159 1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170
##    1    4    4    4    1    1    4    1    1    3    1    1    3    2    2
## 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185
##    7    7    7    2    2    2    2    2    2    2    2    2    2    2    7
## 1186 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1200
##    2    2    2    2    7    2    2    2    2    1    1    7    2    2    2
## 1201 1202 1203 1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215
##    2    7    7    7    2    2    2    7    2    1    1    4    7    7    7
## 1216 1217 1218 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230
##    2    7    2    7    2    2    2    2    1    2    1    2    2    2    2
## 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245
##    2    2    2    2    2    2    2    2    2    2    2    2    2    2    2
## 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260
##    2    2    2    2    2    2    2    2    2    4    2    2    2    9    1
## 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1272 1273 1274 1275
##    2    9    9    2    2    1    1    1    1    1    3    3    3    1    1
```

```
## 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290
##    1    4    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300 1301 1302 1303 1304 1305
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1306 1307 1308 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1320
##    1    1    1    1    1    1    1    1    4    1    4    1    1    1    4
## 1321 1322 1323 1324 1325 1326 1327 1328 1329 1330 1331 1332 1333 1334 1335
##    1    1    1    1    4    1    1    1    1    1    1    1    1    1    1
## 1336 1337 1338 1339 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350
##    1    1    1    1    1    1    1    1    1    1    4    4    4    1    1
## 1351 1352 1353 1354 1355 1356 1357 1358 1359 1360 1361 1362 1363 1364 1365
##    1    4    4    4    4    1    1    1    1    1    1    1    1    1    9
## 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380
##    1    1    1    1    1    2    2    2    2    2    2    2    1    2    2
## 1381 1382 1383 1384 1385 1386 1387 1388 1389 1390 1391 1392 1393 1394 1395
##    2    7    2    7    7    7    2    2    7    7    7    1    2    2    2
## 1396 1397 1398 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1410
##    2    2    7    7    2    2    4    2    9    9    9    1    9    9    9
## 1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425
##    1    2    2    2    2    2    2    2    1    2    1    2    2    2    2
## 1426 1427 1428 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1440
##    2    2    2    2    2    2    7    7    7    7    7    7    7    2    2
## 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455
##    7    7    7    7    2    7    7    7    7    7    2    2    2    2    2
## 1456 1457 1458 1459 1460 1461 1462 1463 1464 1465 1466 1467 1468 1469 1470
##    2    2    2    2    2    2    2    2    2    2    7    1    1    1    1
## 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500
##    1    1    1    1    1    1    1    4    1    1    2    1    4    4    1
## 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    2
## 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530
##    2    4    2    2    2    2    2    2    2    2    2    2    2    2    2
## 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545
##    2    2    2    2    2    2    2    7    2    7    2    2    2    7    2
## 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560
##    2    2    2    2    2    2    2    2    2    2    2    2    2    2    2
## 1561 1562 1563 1564 1565 1566 1567 1568 1569 1570 1571 1572 1573 1574 1575
##    2    2    2    2    2    2    2    2    2    4    2    7    2    1    1
## 1576 1577 1578 1579 1580 1581 1582 1583 1584 1585 1586 1587 1588 1589 1590
##    1    9    1    9    1    1    2    1    1    1    1    1    1    1    1
## 1591 1592 1593 1594 1595 1596 1597 1598 1599 1600 1601 1602 1603 1604 1605
##    1    1    1    1    1    4    1    1    1    1    1    1    1    1    1
## 1606 1607 1608 1609 1610 1611 1612 1613 1614 1615 1616 1617 1618 1619 1620
##    4    1    1    1    1    1    1    3    1    1    1    4    1    1    1
## 1621 1622 1623 1624 1625 1626 1627 1628 1629 1630 1631 1632 1633 1634 1635
##    1    1    1    3    3    3    3    3    3    1    1    3    3    3    3
## 1636 1637 1638 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1650
##    3    3    3    2    2    4    2    1    2    1    2    1    2    2    1
```

```
## 1651 1652 1653 1654 1655 1656 1657 1658 1659 1660 1661 1662 1663 1664 1665
##    2    1    2    2    2    2    2    2    2    2    7    2    7    2    2
## 1666 1667 1668 1669 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1680
##    2    7    2    2    2    2    2    2    2    2    2    2    2    2    2
## 1681 1682 1683 1684 1685 1686 1687 1688 1689 1690 1691 1692 1693 1694 1695
##    2    2    2    2    2    2    2    2    2    2    2    2    4    2    7
## 1696 1697 1698 1699 1700 1701 1702 1703 1704 1705 1706 1707 1708 1709 1710
##    2    2    2    2    1    2    2    1    6   10   10    1    1    2    1
## 1711 1712 1713 1714 1715 1716 1717 1718 1719 1720 1721 1722 1723 1724 1725
##    2    2    2    1    2    1    2    1    2    2    2    2    2    1    2
## 1726 1727 1728 1729 1730 1731 1732 1733 1734 1735 1736 1737 1738 1739 1740
##    2    1    2    2    1   10    1    2    1    1    2    1    2    2    2
## 1741 1742 1743 1744 1745 1746 1747 1748 1749 1750 1751 1752 1753 1754 1755
##    2    2    2    2    2    2    2    2    2    1    2    1    2    1    2
## 1756 1757 1758 1759 1760 1761 1762 1763 1764 1765 1766 1767 1768 1769 1770
##    1    2    2    1    2    1    2    1    2    1    2    1    1    1    6
## 1771 1772 1773 1774 1775 1776 1777 1778 1779 1780 1781 1782 1783 1784 1785
##    1    1    2    1    2    2    2    2    2    2    2    7    2    2    2
## 1786 1787 1788 1789 1790 1791 1792 1793 1794 1795 1796 1797 1798 1799 1800
##    7    7    2    2    1    1    1    9    1    1    1    1    8    1    1
## 1801 1802 1803 1804 1805 1806 1807 1808 1809 1810 1811 1812 1813 1814 1815
##    2    1    2   10    1    1    1    1    1    1    1    1    1    1    1
## 1816 1817 1818 1819 1820 1821 1822 1823 1824 1825 1826 1827 1828 1829 1830
##    1    1    1    1    1    1    1    1    1    1    1    4    1    1    1
## 1831 1832 1833 1834 1835 1836 1837 1838 1839 1840 1841 1842 1843 1844 1845
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1846 1847 1848 1849 1850 1851 1852 1853 1854 1855 1856 1857 1858 1859 1860
##    1    4    1    2    1    1    1    1    1    1    1    1    1    1    1
## 1861 1862 1863 1864 1865 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875
##    1    1    1    1    1    1    1    1    5    1    1    1    1    1    1
## 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890
##    1    1    1    1    1    1    1    4    1    2    1    1    1    5    1
## 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905
##    1   10    1    1    4    1    1    4    1    1    1    1    2    1    2
## 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920
##    5   10   10    1    1    1    1    1    1    1    1    1    1    1    1
## 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935
##    3   10    1    1    1    1    1   10    1    1    1    1   10    1    1
## 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950
##    1    4    1    1    1   10   10    1    1    1    1    1    1    1    1
## 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965
##    1    1    1    5    3    1    1    1    3   10   10    1    1    1    1
## 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980
##    1    1    1    1    1    1    1    1    1    1   10    1    1   10    1
## 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
##    1    1    1    1    1    5    3    3    1   10    1   10    1    1    1
## 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
##   10   10    1    2    1    2    1    2    2    1    2    6    1    2    2
```

```
## 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040
##    1    2    2    2    1    2    1    2    2    2    2    2    2    2    2
## 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055
##    2    2    7    7    7    7    7    7    7    2    2    7    7    7    7
## 2056 2057 2058 2059 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070
##    2    7    7    7    7    7    2    2    2    2    2    2    2    2    2
## 2071 2072 2073 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085
##    2    2    2    2    2    2    7    1    2    2    1    2    2    2    2
## 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100
##    1    2    2    2    7    2    7    7    7    2    7    7    7    1    2
## 2101 2102 2103 2104 2105 2106 2107 2108 2109 2110 2111 2112 2113 2114 2115
##    2    2    2    2    7    7    2    2    4    2    2    2    7    7    7
## 2116 2117 2118 2119 2120 2121 2122 2123 2124 2125 2126 2127 2128 2129 2130
##    2    2    2    2    2    2    2    2    2    2    2    7    2    2    7
## 2131 2132 2133 2134 2135 2136 2137 2138 2139 2140 2141 2142 2143 2144 2145
##    2    2    2    2    1    1    7    2    2    2    2    7    7    7    2
## 2146 2147 2148 2149 2150 2151 2152 2153 2154 2155 2156 2157 2158 2159 2160
##    2    2    7    2    1    1    4    7    7    7    2    7    2    7    2
## 2161 2162 2163 2164 2165 2166 2167 2168 2169 2170 2171 2172 2173 2174 2175
##    2    2    2    1    2    2    2    2    1   10    1    2    1    2    2
## 2176 2177 2178 2179 2180 2181 2182 2183 2184 2185 2186 2187 2188 2189 2190
##    1    2    2    1    2    2    2    1    2    2    1    2    1    2    2
## 2191 2192 2193 2194 2195 2196 2197 2198 2199 2200 2201 2202 2203 2204 2205
##    2    2    2    2    1    2    1    2    4    2    2    2    9    2    1
## 2206 2207 2208 2209 2210 2211 2212 2213 2214 2215 2216 2217 2218 2219 2220
##    2    2    2    7    2    7    2    1    2    2    2    2    2    2    2
## 2221 2222 2223 2224 2225 2226 2227 2228 2229 2230 2231 2232 2233 2234 2235
##    7    2    2    2    2    2    2    2    2    2    2    2    2    6    2
## 2236 2237 2238 2239 2240 2241 2242 2243 2244 2245 2246 2247 2248 2249 2250
##   10    1    2    1    2    1    2    1    2    2    1    2    2    1    2
## 2251 2252 2253 2254 2255 2256 2257 2258 2259 2260 2261 2262 2263 2264 2265
##    1    2    1    1    2    2    2    2    2    2    2    2    2    3    2
## 2266 2267 2268 2269 2270 2271 2272 2273 2274 2275 2276 2277 2278 2279 2280
##    1   10    9    1    9    9    9    1    9    9    9    9    9    9    9
## 2281 2282 2283 2284 2285 2286 2287 2288 2289 2290 2291 2292 2293 2294 2295
##    1    9    1    2    9    9    2    1    9    9    9    1    9    9    9
## 2296 2297 2298 2299 2300 2301 2302 2303 2304 2305 2306 2307 2308 2309 2310
##    9    9    9    1    1    9    1    1    1    1    1    1    1    1    1
## 2311 2312 2313 2314 2315 2316 2317 2318 2319 2320 2321 2322 2323 2324 2325
##    1    2    1    1    1    1    2    1    1    1    1    1    1   10   10
## 2326 2327 2328 2329 2330 2331 2332 2333 2334 2335 2336 2337 2338 2339 2340
##    1    5    1   10   10    1    1    1    1    1    1    5   10    1    1
## 2341 2342 2343 2344 2345 2346 2347 2348 2349 2350 2351 2352 2353 2354 2355
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    5
## 2356 2357 2358 2359 2360 2361 2362 2363 2364 2365 2366 2367 2368 2369 2370
##    1    1    4    1    4    1    1    1    1    1    1    4    1    1    1
## 2371 2372 2373 2374 2375 2376 2377 2378 2379 2380 2381 2382 2383 2384 2385
##    1    1    1    4    1    1    1    1    1    2    1    1    4    1    1
## 2386 2387 2388 2389 2390 2391 2392 2393 2394 2395 2396 2397 2398 2399 2400
##    4    1    1    4    3    1    1    1    1    4    1    1    1    1    1
```

```
## 2401 2402 2403 2404 2405 2406 2407 2408 2409 2410 2411 2412 2413 2414 2415
##    1    1    1    1    4    1    1    1    1    1    1    3    1    1    1
## 2416 2417 2418 2419 2420 2421 2422 2423 2424 2425 2426 2427 2428 2429 2430
##    4    1    1    1    1    1    1    1    1    1    1    1    4    1    1
## 2431 2432 2433 2434 2435 2436 2437 2438 2439 2440 2441 2442 2443 2444 2445
##    2    1    4    4    1    1    1    1    1    1    1    1    1    1    1
## 2446 2447 2448 2449 2450 2451 2452 2453 2454 2455 2456 2457 2458 2459 2460
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 2461 2462 2463 2464 2465 2466 2467 2468 2469 2470 2471 2472 2473 2474 2475
##    1    1    1    1    1    1    1    1    1    1    1    1    4    1    4
## 2476 2477 2478 2479 2480 2481 2482 2483 2484 2485 2486 2487 2488 2489 2490
##    1    1    4    1    1    1    1    4    1    1    1    1    1    1    1
## 2491 2492 2493 2494 2495 2496 2497 2498 2499 2500 2501 2502 2503 2504 2505
##    1    1    1    1    1    1    1    1    1    1    1    1    1    4    4
## 2506 2507 2508 2509 2510 2511 2512 2513 2514 2515 2516 2517 2518 2519 2520
##    4    1    1    1    4    4    4    4    1    1    1    1    1    1    1
## 2521 2522 2523 2524 2525 2526 2527 2528 2529 2530 2531 2532 2533 2534 2535
##    1    1    9    1    1    1    1    1    4    1    1    1    1    3    1
## 2536 2537 2538 2539 2540 2541 2542 2543 2544 2545 2546 2547 2548 2549 2550
##    4    1    1    1    4    1    4    1    1    1    4    1    1    1    1
## 2551 2552 2553 2554 2555 2556 2557 2558 2559 2560 2561 2562 2563 2564 2565
##    4    1    1    1    4    4    4    1    4    4    1    1    1    1    1
## 2566 2567 2568 2569 2570 2571 2572 2573 2574 2575 2576 2577 2578 2579 2580
##    4    3    4    1    1    1    1    1    1    1    1    1    4    1    1
## 2581 2582 2583 2584 2585 2586 2587 2588 2589 2590 2591 2592 2593 2594 2595
##    4    1    1    3    1    3    1    1    1    1    1    1    1    4    1
## 2596 2597 2598 2599 2600 2601 2602 2603 2604 2605 2606 2607 2608 2609 2610
##    1    1    1    4    4    1    1    1    1    1    1    1    1    1    1
## 2611 2612 2613 2614 2615 2616 2617 2618 2619 2620 2621 2622 2623 2624 2625
##    4    1    1    1    1    4    1    1    1    1    4    1    4    4    4
## 2626 2627 2628 2629 2630 2631 2632 2633 2634 2635 2636 2637 2638 2639 2640
##    4    1    1    4    1    1    4    1    1    1    1    1    1    1    1
## 2641 2642 2643 2644 2645 2646 2647 2648 2649 2650 2651 2652 2653 2654 2655
##    1    1    1    1    1    4    1    1    1    1    1    1    1    1    1
## 2656 2657 2658 2659 2660 2661 2662 2663 2664 2665 2666 2667 2668 2669 2670
##    1    3    3    3    1    1    1    3    1    3    1    1    1    1    1
## 2671 2672 2673 2674 2675 2676 2677 2678 2679 2680 2681 2682 2683 2684 2685
##    3    1    1    1    1    1    3    1    1    3    3    3    4    1    3
## 2686 2687 2688 2689 2690 2691 2692 2693 2694 2695 2696 2697 2698 2699 2700
##    1    1    1    1    1    1    1    1    1    1    1    1    1    4    1
## 2701 2702 2703 2704 2705 2706 2707 2708 2709 2710
##    1    1    1    1    1    4    1    1    1    4
##
## Within cluster sum of squares by cluster:
##  [1] 3698.95042 1523.76954  177.48454  442.48246  406.62500  167.00000
##  [7]   84.15873    0.00000   70.54688  506.03571
##  (between_SS / total_SS =  47.4 %)
##
## Available components:
##
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Observations:

1. There are 10 clusters of sizes 397, 130, 35, etc
2. Cluster means: That is, average of each term being analysed. High average means that word has appeared in the particular cluster with higher frequency. For example, "charges" = 0.986577181, that is the term "charges" appear the highest in cluster 6
3. Clustering vector: We have 1623 terms. This shows which cluster each term has gone to. Example, term 3 went to cluster 7, term 992 went to cluster 2
4. Within cluster sum of squares by cluster: We want this to be low, which means the elements within each cluster are close to each other
5. between_SS / total_SS: We want this to be high, that is, distances between clusters are high

we can experiment with k = ? to maximise between_SS / total_SS

2.9 Find the pair of terms that appears frequently together

```
# Term document matrix to convert unstructure text into structured for easier analysis
tdm <- TermDocumentMatrix(corpus)
tdm <- as.matrix(tdm)
tdm[1:25,1:25] # See the first 10 terms in the first 10 documents (tweets)
```

```
##                  Docs
## Terms             1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
##   back            1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   canada          1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   potus           1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   prosperous      1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   removed         1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   sent            1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   states          1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   torontodominion 1 0 0 0 1 0 0 0 0  1  0  1  0  0  0  0  0  0  0  0  0  0
##   united          1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   apompliano      0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   barclays        0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   china           0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   communicatio…   0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   current         0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   fixing          0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   gold            0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   participants    0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   price           0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   com…            0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
##   good            0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   officialmcafee  0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   set             0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   david           0 0 0 2 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   dominion        0 0 0 1 0 1 1 1 1  0  1  0  1  1  0  0  1  1  1  0  0  1
##   gas             0 0 0 1 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##                  Docs
## Terms             23 24 25
##   back             0  0  0
##   canada           0  0  0
##   potus            0  0  0
##   prosperous       0  0  0
##   removed          0  0  0
##   sent             0  0  0
##   states           0  0  0
##   torontodominion  1  1  0
##   united           0  0  0
##   apompliano       0  0  0
##   barclays         0  0  0
##   china            0  0  0
##   communicatio…    0  0  0
##   current          0  0  0
##   fixing           0  0  0
##   gold             0  0  0
##   participants     0  0  0
##   price            0  0  0
##   com…             0  0  0
##   good             0  0  0
##   officialmcafee   0  0  0
##   set              0  0  0
##   david            0  0  0
##   dominion         0  0  0
##   gas              0  0  0
```

Above show how many times each term appears in each tweet. Example “david” appears twice in tweet 4

We can use this table to determine what other terms to remove from analysis

```
# Network of terms
library(igraph) # Network Analysis and Visualization
```

```
##
## 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
```

```
tdm[tdm>1] <- 1 # Convert matrix into binary, that is whether a term appears
(1) or not (0)
```

```
tdm[1:25,1:25]
```

```
##                  Docs
## Terms             1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
##   back            1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   canada          1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   potus           1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   prosperous      1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   removed         1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   sent            1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   states          1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   torontodominion 1 0 0 0 1 0 0 0 0  1  0  1  0  0  0  0  0  0  0  0  0  0
##   united          1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   apompliano      0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   barclays        0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   china           0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   communicatio…   0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   current         0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   fixing          0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   gold            0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   participants    0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   price           0 1 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   com…            0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   good            0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   officialmcafee  0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   set             0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   david           0 0 0 1 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   dominion        0 0 0 1 0 1 1 1 1  0  1  0  1  1  0  0  1  1  1  0  0  1
```

```
##   gas              0 0 0 1 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0
##                  Docs
## Terms             23 24 25
##   back             0  0  0
##   canada           0  0  0
##   potus            0  0  0
##   prosperous       0  0  0
##   removed          0  0  0
##   sent             0  0  0
##   states           0  0  0
##   torontodominion  1  1  0
##   united           0  0  0
##   apompliano       0  0  0
##   barclays         0  0  0
##   china            0  0  0
##   communicatio…    0  0  0
##   current          0  0  0
##   fixing           0  0  0
##   gold             0  0  0
##   participants     0  0  0
##   price            0  0  0
##   com…             0  0  0
##   good             0  0  0
##   officialmcafee   0  0  0
##   set              0  0  0
##   david            0  0  0
##   dominion         0  0  0
##   gas              0  0  0
```

Note matrix values are now all binary

2.9.1 Pair of terms that appears frequently together

```
# Create term-term matrix
termM <- tdm %*% t(tdm) # Multiply tdm and transpose of tdm
termM[1:25,1:25]
```

```
##                  Terms
## Terms             back canada potus prosperous removed sent states
##   back              18      7     3          3       3    3      3
##   canada             7    970     3          3       3    4      3
##   potus              3      3     3          3       3    3      3
##   prosperous         3      3     3          3       3    3      3
##   removed            3      3     3          3       3    3      3
##   sent               3      4     3          3       3    6      3
##   states             3      3     3          3       3    3      3
##   torontodominion    3      9     3          3       3    3      3
##   united             3      3     3          3       3    3      3
##   apompliano         1      2     1          1       1    1      1
##   barclays           1      8     1          1       1    1      1
##   china              0      0     0          0       0    0      0
```

```
##   communicatio…      0      0     0          0       0    0      0
##   current            1      4     1          1       1    1      1
##   fixing             1      1     1          1       1    1      1
##   gold               1      2     1          1       1    1      1
##   participants       1      1     1          1       1    1      1
##   price              2    155     1          1       1    1      1
##   com…               0      0     0          0       0    0      0
##   good               0      9     0          0       0    0      0
##   officialmcafee     0      0     0          0       0    0      0
##   set                0      2     0          0       0    0      0
##   david              0      5     0          0       0    0      0
##   dominion           1     17     1          1       1    1      1
##   gas                0      8     0          0       0    0      0
##                  Terms
## Terms             torontodominion united apompliano barclays china
##   back                          3      3          1        1     0
##   canada                        9      3          2        8     0
##   potus                         3      3          1        1     0
##   prosperous                    3      3          1        1     0
##   removed                       3      3          1        1     0
##   sent                          3      3          1        1     0
##   states                        3      3          1        1     0
##   torontodominion              48      3          1        1     0
##   united                        3      3          1        1     0
##   apompliano                    1      1          8        5     4
##   barclays                      1      1          5       18     8
##   china                         0      0          4        8    12
##   communicatio…                 0      0          4        4     4
##   current                       1      1          5       11     8
##   fixing                        1      1          5       11     8
##   gold                          1      1          5        5     4
##   participants                  1      1          5       11     8
##   price                         4      1          5       16     8
##   com…                          0      0          0        4     4
##   good                          0      0          0        4     4
##   officialmcafee                0      0          0        4     4
##   set                           0      0          0        4     4
##   david                         0      0          0        0     0
##   dominion                      2      1          1        1     0
##   gas                           0      0          0        0     0
##                  Terms
## Terms             communicatio… current fixing gold participants price
##   back                        0       1      1    1            1     2
##   canada                      0       4      1    2            1   155
##   potus                       0       1      1    1            1     1
##   prosperous                  0       1      1    1            1     1
##   removed                     0       1      1    1            1     1
##   sent                        0       1      1    1            1     1
##   states                      0       1      1    1            1     1
##   torontodominion             0       1      1    1            1     4
```

```
##   united                    0       1      1    1            1     1
##   apompliano                4       5      5    5            5     5
##   barclays                  4      11     11    5           11    16
##   china                     4       8      8    4            8     8
##   communicatio…             4       4      4    4            4     4
##   current                   4      20     11    5           11     9
##   fixing                    4      11     11    5           11     9
##   gold                      4       5      5   22            5     5
##   participants              4      11     11    5           11     9
##   price                     4       9      9    5            9   179
##   com…                      0       4      4    0            4     4
##   good                      0       5      4    0            4     4
##   officialmcafee            0       4      4    0            4     4
##   set                       0       4      4    0            4    10
##   david                     0       0      0    0            0     0
##   dominion                  0       1      1    1            1     2
##   gas                       0       0      0    0            0     4
##                  Terms
## Terms              com… good officialmcafee set david dominion gas
##   back                0    0              0   0     0        1   0
##   canada              0    9              0   2     5       17   8
##   potus               0    0              0   0     0        1   0
##   prosperous          0    0              0   0     0        1   0
##   removed             0    0              0   0     0        1   0
##   sent                0    0              0   0     0        1   0
##   states              0    0              0   0     0        1   0
##   torontodominion     0    0              0   0     0        2   0
##   united              0    0              0   0     0        1   0
##   apompliano          0    0              0   0     0        1   0
##   barclays            4    4              4   4     0        1   0
##   china               4    4              4   4     0        0   0
##   communicatio…       0    0              0   0     0        0   0
##   current             4    5              4   4     0        1   0
##   fixing              4    4              4   4     0        1   0
##   gold                0    0              0   0     0        1   0
##   participants        4    4              4   4     0        1   0
##   price               4    4              4  10     0        2   4
##   com…                9    4              4   4     0        0   0
##   good                4   29              4   4     0        0   0
##   officialmcafee      4    4              4   4     0        0   0
##   set                 4    4              4  16     0        0   0
##   david               0    0              0   0     8        1   1
##   dominion            0    0              0   0     1       76   1
##   gas                 0    0              0   0     1        1   9
```

Example: price and canada appear together in 88 tweets

2.9.2 Find the network of terms

```
g <- graph.adjacency(termM, weighted = T, mode = 'undirected')
g
```

```
## IGRAPH 1feed27 UNW- 3702 75469 --
## + attr: name (v/c), weight (e/n)
## + edges from 1feed27 (vertex names):
##  [1] back--back            back--canada          back--potus
##  [4] back--prosperous      back--removed         back--sent
##  [7] back--states          back--torontodominion back--united
## [10] back--apompliano      back--barclays        back--current
## [13] back--fixing          back--gold            back--participants
## [16] back--price           back--dominion        back--toronto
## [19] back--anonymously     back--article         back--edited
## [22] back--wikipedia       back--form            back--filing
## + ... omitted several edges
```

```
g <- simplify(g) # To prevent looping of same terms
V(g)$label <- V(g)$name # Labels for the terms
V(g)$degree <- degree(g) # How often each term appears

# Histogram of node degree
hist(V(g)$degree,
     breaks = 100, # how many bars
     col = 'green',
     main = 'Histogram of Node Degree',
     ylab = 'Frequency',
     xlab = 'Degree of Vertices')
```

Above is a right skewed histogram and most terms appears less than 100 times

```
# Network diagram
plot(g)
```

Above graph is too busy. One method is to reduce the size of the matrix

```
tdm <- tdm[rowSums(tdm)>30,] # rowSums counts the total frequency; that is, k
eep terms that appear > 30 times

# Re-run earlier code
tdm[tdm>1] <- 1
termM <- tdm %*% t(tdm)
termM[1:10,1:10]
```

```
##                  Terms
## Terms             canada torontodominion price dominion toronto stock
##   canada             970               9   155       17      80    88
##   torontodominion      9              48     4        2       2     8
##   price              155               4   179        2       3    28
##   dominion            17               2     2       76      72     0
##   toronto             80               2     3       72     174     0
##   stock               88               8    28        0       0   137
##   anonymously         10               1     1       12      12     0
##   article             10               1     1       12      12     0
##   edited              10               1     1       12      12     0
```

```
##   wikipedia           10               1     1       12      12     0
##                   Terms
## Terms              anonymously article edited wikipedia
##   canada                    10      10     10        10
##   torontodominion            1       1      1         1
##   price                      1       1      1         1
##   dominion                  12      12     12        12
##   toronto                   12      12     12        12
##   stock                      0       0      0         0
##   anonymously               32      32     32        32
##   article                   32      39     32        32
##   edited                    32      32     32        32
##   wikipedia                 32      32     32        32
```

```
g <- graph.adjacency(termM, weighted = T, mode = 'undirected')
g
```

```
## IGRAPH 2c546e7 UNW- 118 2529 --
## + attr: name (v/c), weight (e/n)
## + edges from 2c546e7 (vertex names):
##  [1] canada--canada          canada--torontodominion
##  [3] canada--price           canada--dominion
##  [5] canada--toronto         canada--stock
##  [7] canada--anonymously     canada--article
##  [9] canada--edited          canada--wikipedia
## [11] canada--form            canada--filing
## [13] canada--sec             canada--canadian
## [15] canada--new             canada--news
## + ... omitted several edges
```

```
g <- simplify(g) # To prevent looping of same terms
V(g)$label <- V(g)$name
V(g)$degree <- degree(g)

# Histogram of node degree
hist(V(g)$degree,
     breaks = 100, # how many bars
     col = 'green',
     main = 'Histogram of Node Degree',
     ylab = 'Frequency',
     xlab = 'Degree of Vertices')
```

NOte the historgram is less busy

```
# Network diagram
set.seed(222)
plot(g)
```

```
plot(g,
     vertex.color='green',
```

```
     vertex.size = 8, # can experiment with this
     vertex.label.dist = 1.5,
     vertex.label = NA)
```

Much less busy than earlier. We can experiment with the vertex size to find the optimal network of terms

3. Observations and recommendation

Based on this exploratory data analysis, we will assess whether the stated research questions are feasible, and as such, whether an edit of the research question is needed

Research Q#1. Which bank has the most favourable / unfavourable trending opinion? Comments: ABout 1,623 tweets has been collected since July 7th, 2019, with close 5,500 terms. The collection will increase in the next several weeks. It should be feasible to answer this research question. The main drawback is for the low count of CIBC tweets (40 tweets) versus that of Scotia Bank (661 tweets). The wide difference will skew the analysis, especially that of CIBC's

Research Q#2. What are the current financial products being discussed? Comments: Frequent terms related to banking products are generic ones like, stock, charges, account. Unless we have a much more collection of tweets, it will be difficult to objectibely address this reserach question

Research Q#3. What are the current emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) towards each bank? Comments: Not shown in this Sprint#1 report as the codes are still experimental, the author managed to "see" these emotional terms at the AllBanks level. Again, dur to the low tweet count for CIBC, it may be difficult to pin down the sentiments, especially in these 8 categories towards CIBC

Research Q#4. What are the current sentiments towards trending financial product segments / categories (and the general network of terms being tweeted)? Comments: As stated above, frequent terms related to banking products are generic ones, hence it will be difficult to assess sentiments towards product segments. Network of terms is certainly a possibility