3032ICT / 7230ICT / 1117ICT

Big Data Analytics and Social Media

Assignment Specifications

Report

Milestone 2

Prof Bela Stantic

Author: Thi Minh Oanh Luong

Student ID: s538355

School of Information and communication technology

Date of submission: 27th May 2023

Contents

Abstract

Following Milestone 1, this report continues exploring and elucidating the utilization of social media analysis crossing methods of management and improvement in Twice’s popularity. By leveraging APIs from Spotify, YouTube, and Twitter, this report combines various analytical approaches, such as degree, betweenness, closeness centrality, as well as Girvan-Newman and Louvain algorithms, to provide a vivid view of Twice's presence across different platforms. Thus, it would gain a comprehensive understanding of their popularity in order to manage and enhance it effectively.

Introduction

By some details in Milestone 1 including the study case setting and data selection and exploration techniques, the assignment completes data selection and exploration and presents serval new sections including Text pre-processing, social network analysis, machine learning models, and visualization. Spotify data retrieval results not only some details from the band collected to compare to the study case in Milestone 1, but also analysis of 2 prevalent features of their songs. With YouTube dataset, it can calculate the numbers like and views of 5-top videos involving into Twice, providing valuable insights and feedback. From the data source in Milestone 1, it details pre-processing, term-document matrix and top 10 terms in Twitter with some comparation with the last information. In terms of Social Network Analysis about Twice and 2 more related bands, Centrality analysis uses degree, betweenness, and closeness analysis, while Community analysis are depicted by Girvan-Newman and Louvain analysis. By the multiple datasets, Machine Learning models figure out and explain Sentiment analysis, decision tree, and topic modeling about the three bands. Additionally, Visualization in Dashboard presents four different charts detailing description and reasons for inclusion. The purpose of this report is to apply multiple sources and different tools (Rstudio and Tableau) to analysis and gain insight into how musicians can improve their popularity.

Data Selection & Exploration

2.1. Spotify data retrieval

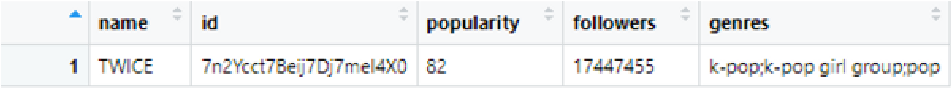
To begin, find out exactly the band name Twice whose primary genre is K-pop:

```
# Authentication for Rspotify package:

keys <- spotifyOAuth(token, app_id, app_secret)

# Get Spotify data on 'Twice'

find_my_artist <- searchArtist("Twice", token = keys)
```

In the `find_my_artist` table, there is precisely one record that is suitable with the demand above. To get further detail form the record, `getArtist` function is applied:

```
# Retrieve information about artist

my_artist <- getArtist("7n2Ycct7Beij7Dj7meI4X0", token = keys)
```

The result is



	name	id	popularity	followers	genres
1	TWICE	7n2Ycct7Beij7Dj7meI4X0	82	17447455	k-pop;k-pop girl group;pop

The number of years they have been active

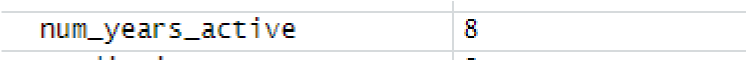
It gets all products owned by Twice and removes duplication after using `get_artist_audio_features` function.

```
Twice <-get_artist_audio_features(artist = "7n2Ycct7Beij7Dj7meI4X0",

                                  include_groups = c("album", "single"),

                                  return_closest_artist = TRUE,

                                  dedupe_albums = TRUE,

                                  market = NULL,
                        authorization = get_spotify_access_token())
```

```
Twice <- Twice[!duplicated(Twice$track_name)]
```

Next, the earliest year from the first single/album of the band is determined. It, then, calculate the number of years they have been active by subtracting the debut year from the current year.

```
earliest_year <- min(Twice$album_release_year)

num_years_active <- (as.integer(format(Sys.Date(), "%Y")) -
earliest_year)
```



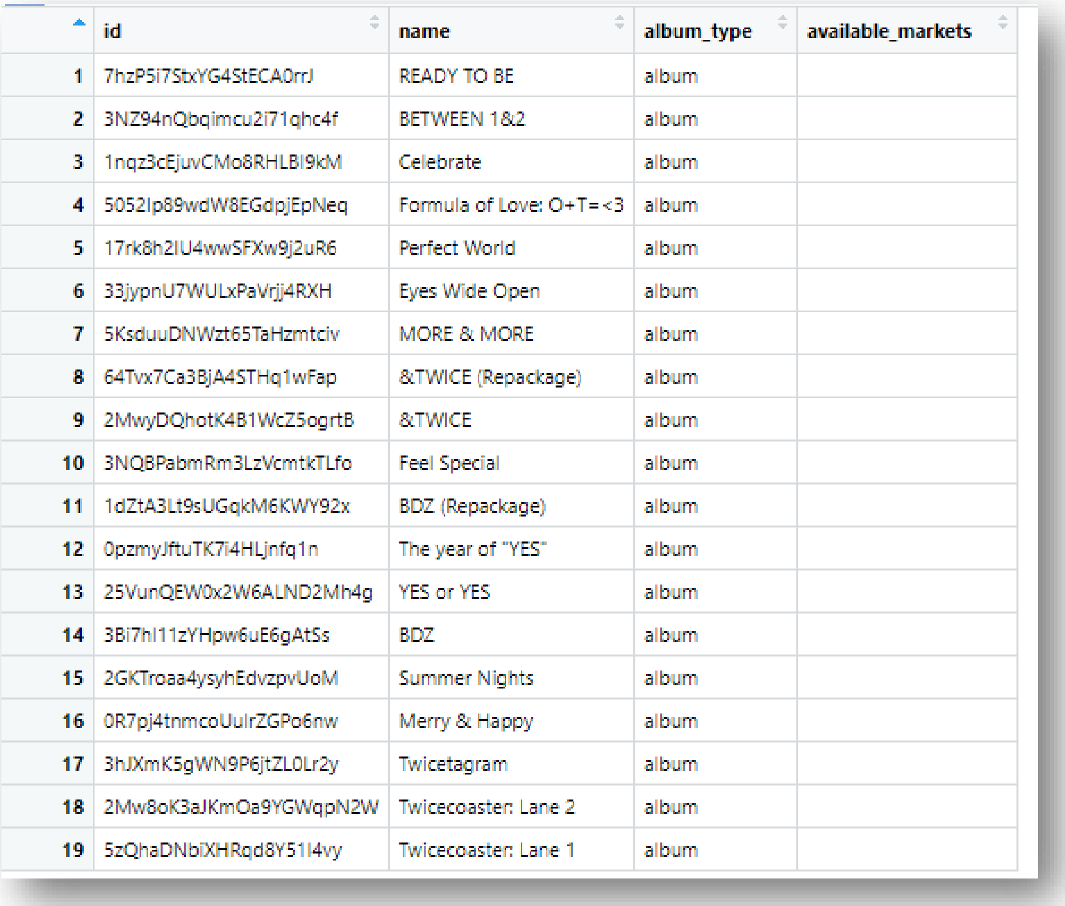
```
num_years_active        8
```

The result is **8 years**

The number of albums & songs have they published

By using *getAlbums* function:

```
# Retrieve album data of artist

albums <- getAlbums("7n2Ycct7Beij7Dj7meI4X0", token = keys)
```



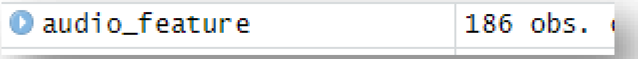
	id	name	album_type	available_markets
1	7hzP5i7StxYG4StECA0rrJ	READY TO BE	album	
2	3NZ94nQbqimcu2i71qhc4f	BETWEEN 1&2	album	
3	1nqz3cEjuvCMo8RHLBI9kM	Celebrate	album	
4	5052Ip89wdW8EGdpjEpNeq	Formula of Love: O+T=<3	album	
5	17rk8h2IU4wwSFXw9j2uR6	Perfect World	album	
6	33jypnU7WULxPaVrjj4RXH	Eyes Wide Open	album	
7	5KsduuDNWzt65TaHzmtciv	MORE & MORE	album	
8	64Tvx7Ca3BjA4STHq1wFap	&TWICE (Repackage)	album	
9	2MwyDQhotK4B1WcZ5ogrtB	&TWICE	album	
10	3NQBPabmRm3LzVcmtkTLfo	Feel Special	album	
11	1dZtA3Lt9sUGqkM6KWY92x	BDZ (Repackage)	album	
12	0pzmyJftuTK7i4HLjnfq1n	The year of "YES"	album	
13	25VunQEW0x2W6ALND2Mh4g	YES or YES	album	
14	3Bi7hl11zYHpw6uE6gAtSs	BDZ	album	
15	2GKTroaa4ysyhEdvzpvUoM	Summer Nights	album	
16	0R7pj4tnmcoUulrZGPo6nw	Merry & Happy	album	
17	3hJXmK5gWN9P6jtZL0Lr2y	Twicetagram	album	
18	2Mw8oK3aJKmOa9YGWqpN2W	Twicecoaster: Lane 2	album	
19	5zQhaDNbiXHRqd8Y51l4vy	Twicecoaster: Lane 1	album	

Table 1: Twice's albums

The number of albums is **19 albums** in total.

To get how many songs the band has published, I get audio features for Twice and remove duplication in the dataset.

```
# Get audio features for 'Twice'
audio_features <- get_artist_audio_features("Twice")
audio_feature <- audio_features[!duplicated(audio_features$track_name,
audio_features$album_name), ]
```



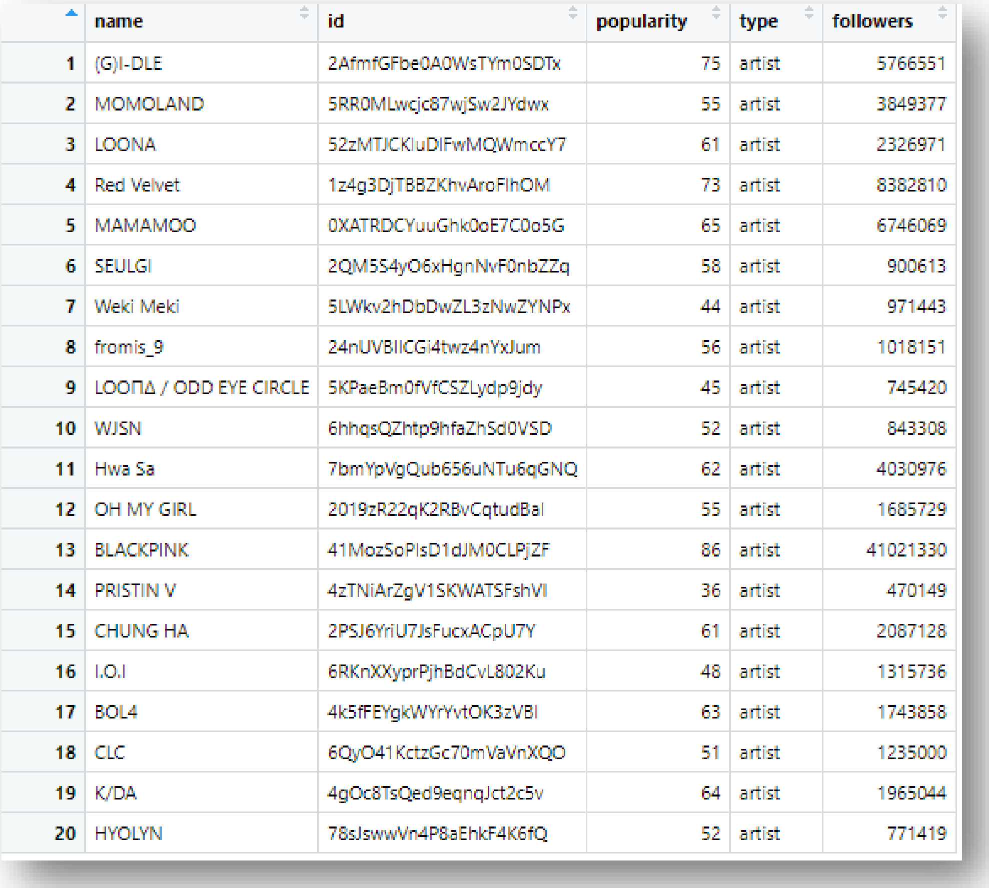
```
audio_feature    186 obs.
```

Thus, there are **186 songs.**

The list of artists/bands they have often collaborated with

To look for who have Twice often collaborated with, `getRelated` is applied:

```
# Retrieve information about related artists

related_bm <- getRelated("Twice", token = keys)
```



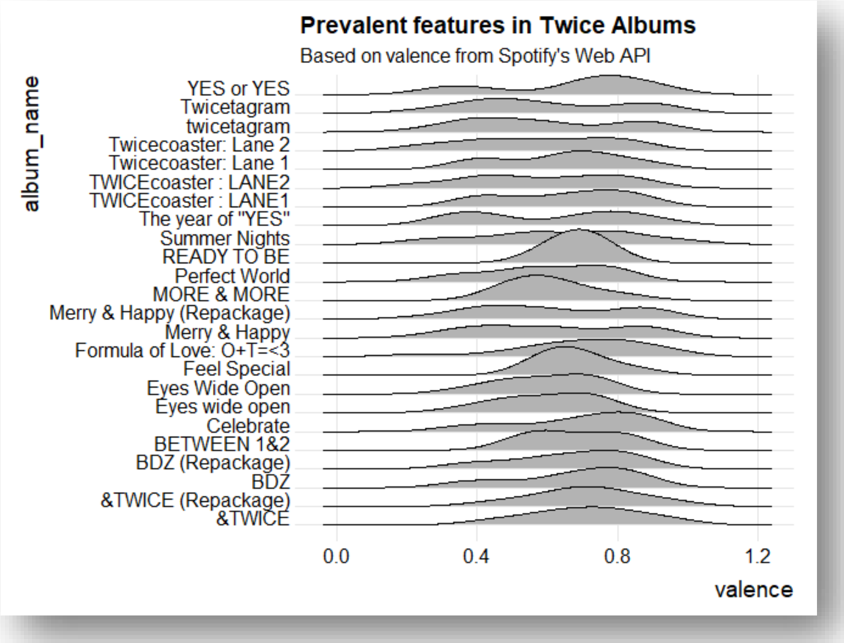
	name	id	popularity	type	followers
1	(G)I-DLE	2AfmfGFbe0A0WsTYm0SDTx	75	artist	5766551
2	MOMOLAND	5RR0MLwcjc87wjSw2JYdwx	55	artist	3849377
3	LOONA	52zMTJCKIuDlFwMQWmccY7	61	artist	2326971
4	Red Velvet	1z4g3DjTBBZKhvAroFlhOM	73	artist	8382810
5	MAMAMOO	0XATRDCYuuGhk0oE7C0o5G	65	artist	6746069
6	SEULGI	2QM5S4yO6xHgnNvF0nbZZq	58	artist	900613
7	Weki Meki	5LWkv2hDbDwZL3zNwZYNPx	44	artist	971443
8	fromis_9	24nUVBIICGi4twz4nYxJum	56	artist	1018151
9	LOOΠΔ / ODD EYE CIRCLE	5KPaeBm0fVfCSZLydp9jdy	45	artist	745420
10	WJSN	6hhqsQZhtp9hfaZhSd0VSD	52	artist	843308
11	Hwa Sa	7bmYpVgQub656uNTu6qGNQ	62	artist	4030976
12	OH MY GIRL	2019zR22qK2RBvCqtudBal	55	artist	1685729
13	BLACKPINK	41MozSoPIsD1dJM0CLPjZF	86	artist	41021330
14	PRISTIN V	4zTNiArZgV1SKWATSFshVI	36	artist	470149
15	CHUNG HA	2PSJ6YriU7JsFucxACpU7Y	61	artist	2087128
16	I.O.I	6RKnXXyprPjhBdCvL802Ku	48	artist	1315736
17	BOL4	4k5fFEYgkWYrYvtOK3zVBI	63	artist	1743858
18	CLC	6QyO41KctzGc70mVaVnXQO	51	artist	1235000
19	K/DA	4gOc8TsQed9eqnqJct2c5v	64	artist	1965044
20	HYOLYN	78sJswwVn4P8aEhkF4K6fQ	52	artist	771419

Table 2: List of Twice's related artists/bands

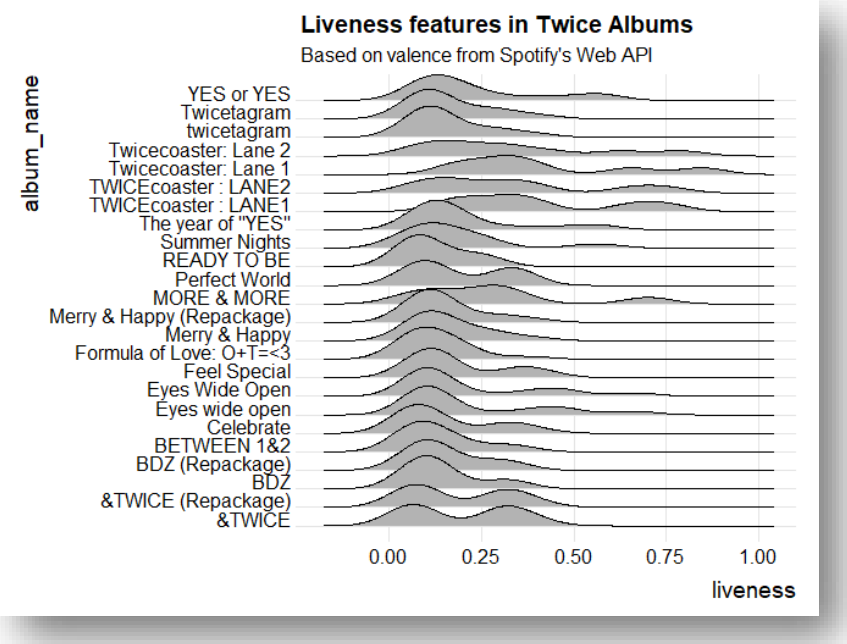
As the result, there are **20 artists/bands** that Twice have collaborated with.

The prevalent features of their songs (e.g., valence)

Based on valence from Spotify's Web API, below is the graphs of prevalent and liveness features in all albums of Twice:



The figure for prevalent features in albums is around 0.2 to 1. In there, *"READY TO BE"* and *"MORE & MORE"* are **more outstanding than others**.



The figure for liveness feature in albums most albums standing at the high level in range 0.1 to 0.25.

Comparation between the Spotify data and the information collected from other sources in Step 1.1 (Milestone 1)

The difference in information between Milestone 1 and 2:

Detail	Milestone 1	Milestone 2	Reason of difference
The years they have been active	8	8	
The number of albums	21	19	There are several albums which in other sources believed to be albums, but they are EPs as Spotify detail. This includes *The story begins, Page two, Signal, and What is love* [4].
The number of songs	202	186	The change of the number of albums *Twice* own.

Compared to Milestone 1, the Spotify data has a slight discrepancy in terms of the number of published albums and songs. The main reason for these inconsistencies in Milestone 1 is the fact that several EPs on Sportify are categorized as albums in the other sources. However, it is important to note that the data provided by Spotify is considered to be more accurate because they are provided by the bands' publisher and manger company.

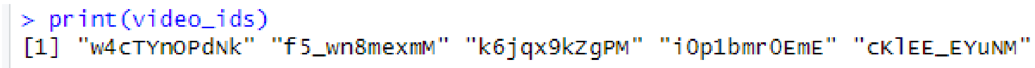
2.2. YouTube views/likes

Videos have the highest number of views and likes

To mitigate the potential rate-limit issue caused by the large dataset of 550 videos, a subset comprising only 5 videos has been created.

```
video_ids <- as.vector(video_search$video_id[1:5])
```

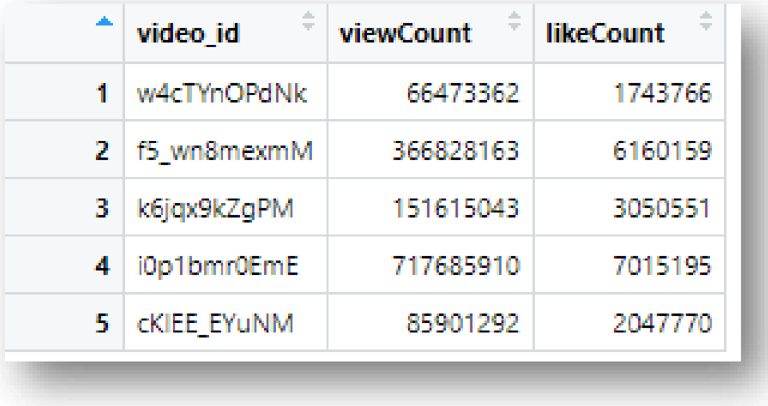
They are:



```
> print(video_ids)
[1] "w4cTYnOPdNk" "f5_wn8mexmM" "k6jqx9kZgPM" "iOp1bmrOEmE" "cKlEE_EYuNM"
```

From the source of statistics, it satisfies the number of views and likes of each video:

```
video_view_likes = data.frame()

for (i in video_ids){

 video_view_like <-c(i,get_stats(i)$viewCount,get_stats(i)$likeCount)

 video_view_likes<-rbind(video_view_likes, video_view_like)

}

columns <-c("video_id","viewCount","likeCount")

colnames(video_view_likes) = columns

#convert datatype from char to numeric

video_view_likes$viewCount =

as.numeric(as.character(video_view_likes$viewCount))

video_view_likes$likeCount =

as.numeric(as.character(video_view_likes$likeCount))
```

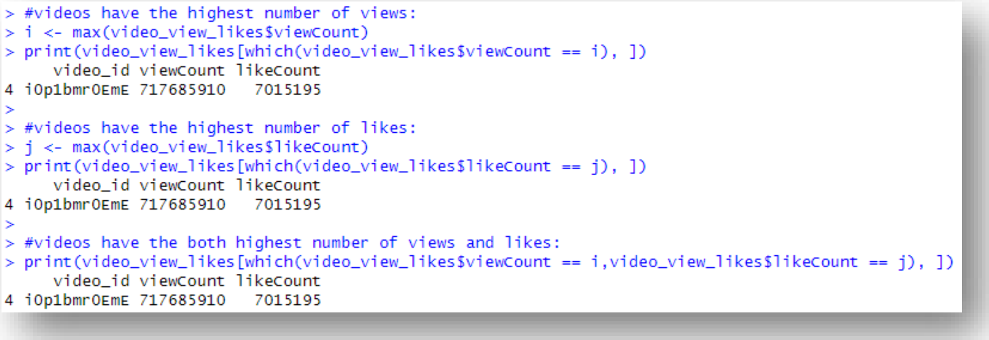


	video_id	viewCount	likeCount
1	w4cTYnOPdNk	66473362	1743766
2	f5_wn8mexmM	366828163	6160159
3	k6jqx9kZgPM	151615043	3050551
4	i0p1bmr0EmE	717685910	7015195
5	cKIEE_EYuNM	85901292	2047770

Table 3: the numbers of views and likes in top 5 videos

Next, find out the index of all videos which have maximum of views, likes, and both.

```
#videos have the highest number of views:

i <- max(video_view_likes$viewCount)

print(video_view_likes[which(video_view_likes$viewCount == i), ])

#videos have the highest number of likes:

j <- max(video_view_likes$likeCount)

print(video_view_likes[which(video_view_likes$likeCount == j), ])

#videos have the both highest number of views and likes:

print(video_view_likes[which(video_view_likes$viewCount ==

i,video_view_likes$likeCount == j), ])
```



```
> #videos have the highest number of views:
> i <- max(video_view_likes$viewCount)
> print(video_view_likes[which(video_view_likes$viewCount == i), ])
     video_id viewCount likeCount
4 iOp1bmrOEmE 717685910   7015195
>
> #videos have the highest number of likes:
> j <- max(video_view_likes$likeCount)
> print(video_view_likes[which(video_view_likes$likeCount == j), ])
     video_id viewCount likeCount
4 iOp1bmrOEmE 717685910   7015195
>
> #videos have the both highest number of views and likes:
> print(video_view_likes[which(video_view_likes$viewCount == i,video_view_likes$likeCount == j), ])
     video_id viewCount likeCount
4 iOp1bmrOEmE 717685910   7015195
```

The correlation between views and likes

The result figures out that the videos having the highest number views are also the videos having the highest number of likes.

Text Pre-Processing

2.3. Pre-processing & term-document matrix & top 10 terms

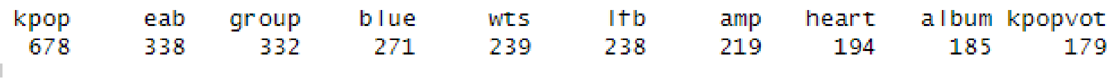
Using *sort* function after applying matrix:

dtm_df <- as.data.frame(as.matrix(doc_term_matrix))

freq <- sort(colSums(dtm_df), decreasing = TRUE)

head(freq, n = 10)

As a result:

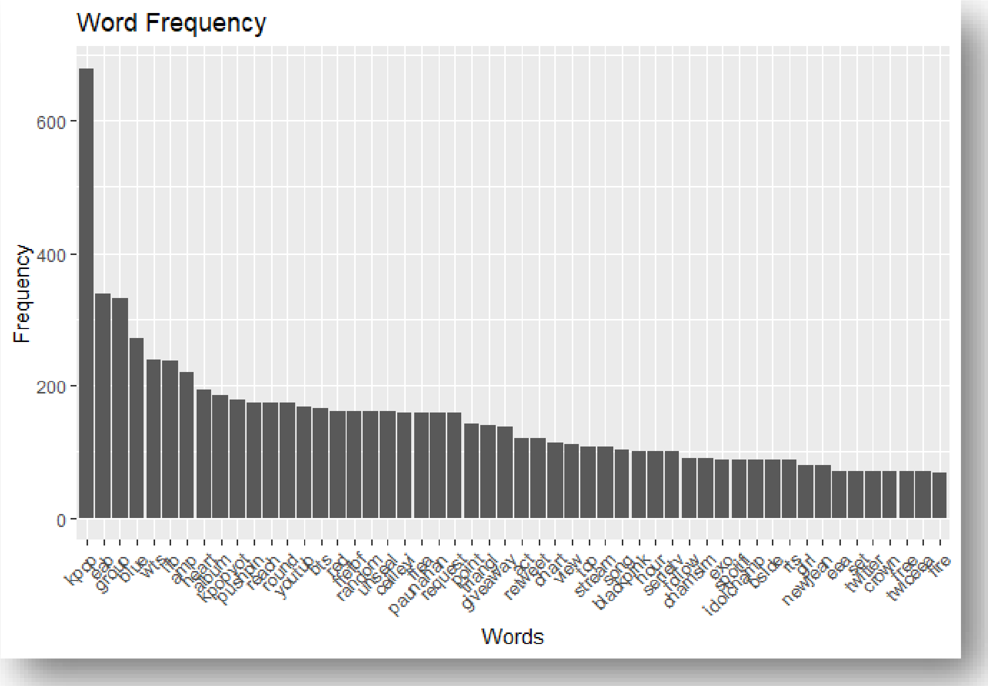


```
   kpop     eab   group    blue     wts     ltb     amp   heart   album kpopvot
    678     338     332     271     239     238     219     194     185     179
```

Term	Description
kpop	primary genre of band
eab	trending key word
group	group
blue	a word of “baby blue love” - one of their songs

wts	trending key word
lfb	trending key word
amp	trending key word
heart	trending key word
album	their products
kpopvot	kpop vote

The bar chart of the result:



Difference to step 1.4 (Milestone 1)

In the Milestone 1, the top 10 terms are:

top
kpop
songs
#kpop
mv
#twice
blue
#pasabuys
group
views

When comparing the top 10 terms from Milestone 1 and 2, several similarities and differences can be observed. The terms from both provide a glimpse of discussion, top-titles, and other contents relating to Twice band including their musical career and personal lives, are shared on Twitter. However, there are notable differences between the two Milestone studies. In Milestone 2, the top 10 terms consist only of text without any hashtags, distinguishing it from the previous study. Moreover, the terms in Milestone 2 tend to reflect trending words on Twitter, making them more general, objective, and potentially valuable for improving the band's popularity.

Social Network Analysis

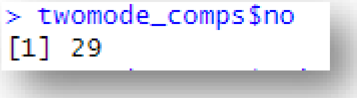
In this task, I created a network as twomode which includes two different modes of nodes and remove hashtag. It can calculate the centrality if and only if the graph is connected (a path between each pair of nodes). Because the graph is fully connected, its components are separated.

In the related artists/band list with *Twice*, I choose *Blackpink* and *Momoland* in the purpose of comparation. Since my Twitter account is suspended, I used YouTube data for Blackpink and Momoland.

2.4. Centrality analysis

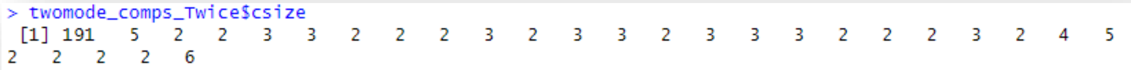
Twice

The number of connected components in our graph:



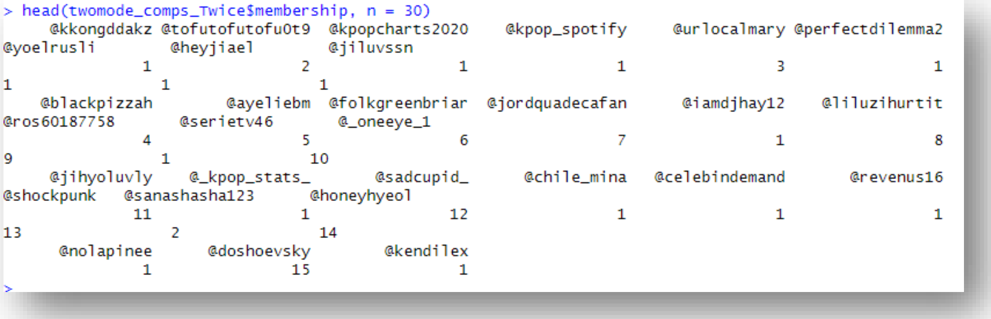
```
> twomode_comps$no
[1] 29
```

The number of nodes in each component:



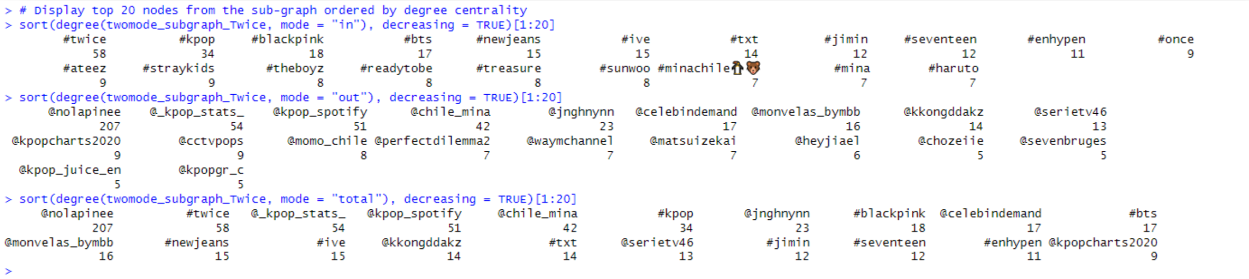
```
> twomode_comps_Twice$csize
 [1] 191   5   2   2   3   3   2   2   2   3   2   3   3   2   3   3   3   2   2   2   3   2   4   5
2   2   2   2   6
```

The first 30 nodes and which component they belong to:



```
> head(twomode_comps_Twice$membership, n = 30)
     @kkongddakz @tofutofutofu0t9  @kpopcharts2020     @kpop_spotify      @urlocalmary @perfectdilemma2
@yoelrusli         @heyjiael         @jiluvssn
               1                2                1                1                3                1
1                1                1
    @blackpizzah        @ayeliebm  @folkgreenbriar  @jordquadecafan      @iamdjhay12     @liluzihurtit
@ros60187758       @serietv46        @_oneeye_1
               4                5                6                7                1                8
9                1               10
     @jihyoluvly    @_kpop_stats_       @sadcupid_      @chile_mina    @celebindemand       @revenus16
@shockpunk    @sanashasha123       @honeyhyeol
              11                1               12                1                1                1
13               2                14
      @nolapinee       @doshoevsky        @kendilex
               1               15                1
>
```

Degree centrality

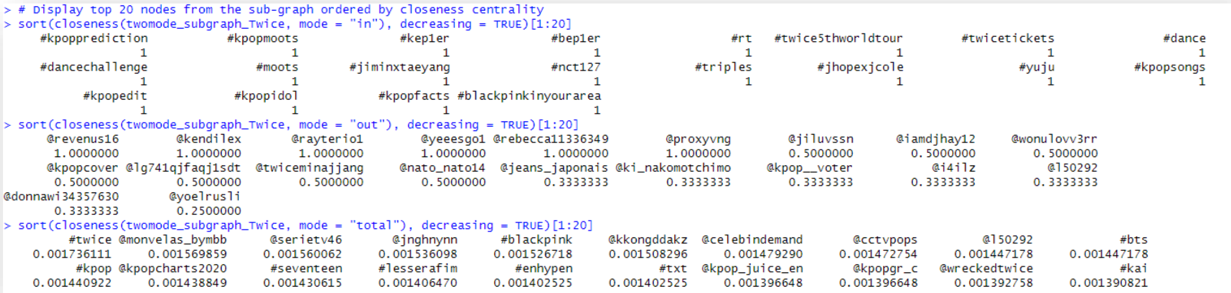


```
> # Display top 20 nodes from the sub-graph ordered by degree centrality
> sort(degree(twomode_subgraph_Twice, mode = "in"), decreasing = TRUE)[1:20]
     #twice       #kpop  #blackpink        #bts   #newjeans        #ive        #txt      #jimin  #seventeen    #enhypen       #once 
         58          34          18          17          15          15          14          12          12          11           9 
     #ateez  #straykids    #theboyz  #readytobe   #treasure     #sunwoo #minachile🐧🐻      #mina     #haruto 
          9           9           8           8           8           8           7           7           7 
> sort(degree(twomode_subgraph_Twice, mode = "out"), decreasing = TRUE)[1:20]
      @nolapinee     @_kpop_stats_    @kpop_spotify      @chile_mina        @jnghnynn  @celebindemand  @monvelas_bymbb      @kkongddakz       @serietv46 
             207               54               51               42               23               17               16               14               13 
 @kpopcharts2020        @cctvpops      @momo_chile @perfectdilemma2     @waymchannel      @matsuizekai        @heyjiael        @chozeiie     @sevenbruges 
               9                9                8                7                7                7                6                5                5 
   @kpop_juice_en        @kpopgr_c 
               5                5 
> sort(degree(twomode_subgraph_Twice, mode = "total"), decreasing = TRUE)[1:20]
     @nolapinee          #twice   @_kpop_stats_   @kpop_spotify     @chile_mina           #kpop       @jnghnynn      #blackpink  @celebindemand            #bts 
            207              58              54              51              42              34              23              18              17              17 
@monvelas_bymbb        #newjeans            #ive     @kkongddakz            #txt      @serietv46          #jimin      #seventeen        #enhypen @kpopcharts2020 
             16              15              15              14              14              13              12              12              11               9 
> 
```

The top hashtag: #twice #kpop #blackpink #bts #newjeans #ive #ateez #straykids #theboyz #readytobe #treasure #sunwoo #minachile #txt #jimin#seventeen #enhypen #once #mina #haruto #newjeans

The top users: @nolapinee *@kpop_stats* @kpop_spotify @chile_mina@kpopcharts 2020 @cctvpops @momo_chile @perfect dilemma2 @jnghnynn @waymchannel @celebindemand @monvelas_bymbb @kkongddakz @serietv46 @matsuizekai @heyjiael @chozeiie @sevenbruges @kpop_juice_en @kpopgr_c @nolapinee *@kpop_stats* @kpop_spotify

Closeness centrality



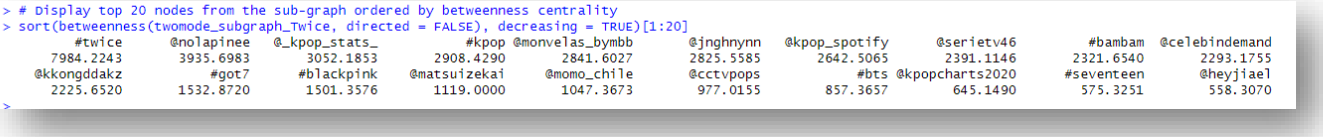
```
> # Display top 20 nodes from the sub-graph ordered by closeness centrality
> sort(closeness(twomode_subgraph_Twice, mode = "in"), decreasing = TRUE)[1:20]
      #kpopprediction            #kpopmoots               #kep1er               #bep1er                   #rt    #twice5thworldtour          #twicetickets                #dance 
                    1                     1                     1                     1                     1                     1                      1                     1 
      #dancechallenge                #moots         #jiminxtaeyang               #nct127              #triples           #jhopexjcole                  #yuju            #kpopsongs 
                    1                     1                     1                     1                     1                     1                      1                     1 
            #kpopedit             #kpopidol            #kpopfacts #blackpinkinyourarea 
                    1                     1                     1                     1 
> sort(closeness(twomode_subgraph_Twice, mode = "out"), decreasing = TRUE)[1:20]
      @revenus16         @kendilex        @rayterio1         @yeeesgo1  @rebecca11336349         @proxyvng         @jiluvssn       @iamdjhay12       @wonulovv3rr 
       1.0000000         1.0000000         1.0000000         1.0000000         1.0000000         1.0000000         0.5000000         0.5000000         0.5000000 
      @kpopcover  @lg741qjfaqj1sdt   @twiceminajjang      @nato_nato14   @jeans_japonais @ki_nakomotchimo     @kpop__voter            @i4ilz           @l50292 
       0.5000000         0.5000000         0.5000000         0.5000000         0.3333333         0.3333333         0.3333333         0.3333333         0.3333333 
 @donnawi34357630        @yoelrusli 
       0.3333333         0.2500000 
> sort(closeness(twomode_subgraph_Twice, mode = "total"), decreasing = TRUE)[1:20]
         #twice @monvelas_bymbb       @serietv46       @jnghnynn      #blackpink     @kkongddakz  @celebindemand       @cctvpops         @l50292            #bts 
    0.001736111     0.001569859     0.001560062     0.001536098     0.001526718     0.001508296     0.001479290     0.001472754     0.001447178     0.001447178 
          #kpop @kpopcharts2020       #seventeen      #lesserafim        #enhypen            #txt  @kpop_juice_en       @kpopgr_c   @wreckedtwice            #kai 
    0.001440922     0.001438849     0.001430615     0.001406470     0.001402525     0.001402525     0.001396648     0.001396648     0.001392758     0.001390821 
```

The top hashtag: #kpopedit #kpopprediction #dancechallenge #kpopmoots #moots #kpopidol #kepler #bepler #rt #twice5thworldtour #twicetickets #dance #jiminxtaeyang #nct127 #triples

#jhopexjcole #yuju #kpopsongs #kpopfacts #blackpinkinyourarea #kai #bts #seventeen #lesserafim #blackpink #enhypen #txt

The top users: @revenus16 @kendilex @rayterio1 @kpopcover @1g741qjfaqj1sdt @twiceminajjang @donnawi @yoelrusli @yeeesgo1 @rebecca @nato_nato14 @jeans_japonais @ki_nakomotchimo @proxyvng @jiluvssn @kpop__voter @iamdjhay12 @i4ilz @monvelas_bymbb @serietv46 @kpopcharts2020 @jnghnynn @kkongddakz @celebindemand @kpop_juice_en @wonulovv3rr @150292 @cctvpops @150292 @kpopgr_c @wreckedtwice

Betweenness centrality



```
> # Display top 20 nodes from the sub-graph ordered by betweenness centrality
> sort(betweenness(twomode_subgraph_Twice, directed = FALSE), decreasing = TRUE)[1:20]
         #twice      @nolapinee   @_kpop_stats_           #kpop @monvelas_bymbb       @jnghnynn   @kpop_spotify      @serietv46         #bambam  @celebindemand 
      7984.2243       3935.6983       3052.1853       2908.4290       2841.6027       2825.5585       2642.5065       2391.1146       2321.6540       2293.1755 
    @kkongddakz           #got7      #blackpink    @matsuizekai     @momo_chile       @cctvpops            #bts @kpopcharts2020      #seventeen       @heyjiael 
      2225.6520       1532.8720       1501.3576       1119.0000       1047.3673        977.0155        857.3657        645.1490        575.3251        558.3070 
> 
```

The top hashtag: #twice #got7 #bts #blackpink#kpop #seventeen #bambam

The top users: @nolapinee @kkongddakz *@kpop_stats* @monvelas_bymbb @matsuizekai @momo_chile @jnghnynn @kpop_spotify @cctvpops @serietv46 @celebindemand @kpopcharts2020 @heyjiael

In all centrality, there are some hashtags and users are similar.

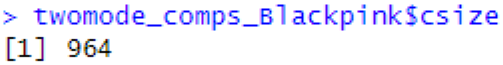
Blackpink

Because 2-mode is not supported for YouTube, I changed bimodal to actor network in comments at a video posted by Blackpink official channel.

```
youtubeAuth <- Authenticate("youtube", apiKey = api_key)

videoID <- "https://www.youtube.com/watch?v=hR1gMWQS-ws"
```

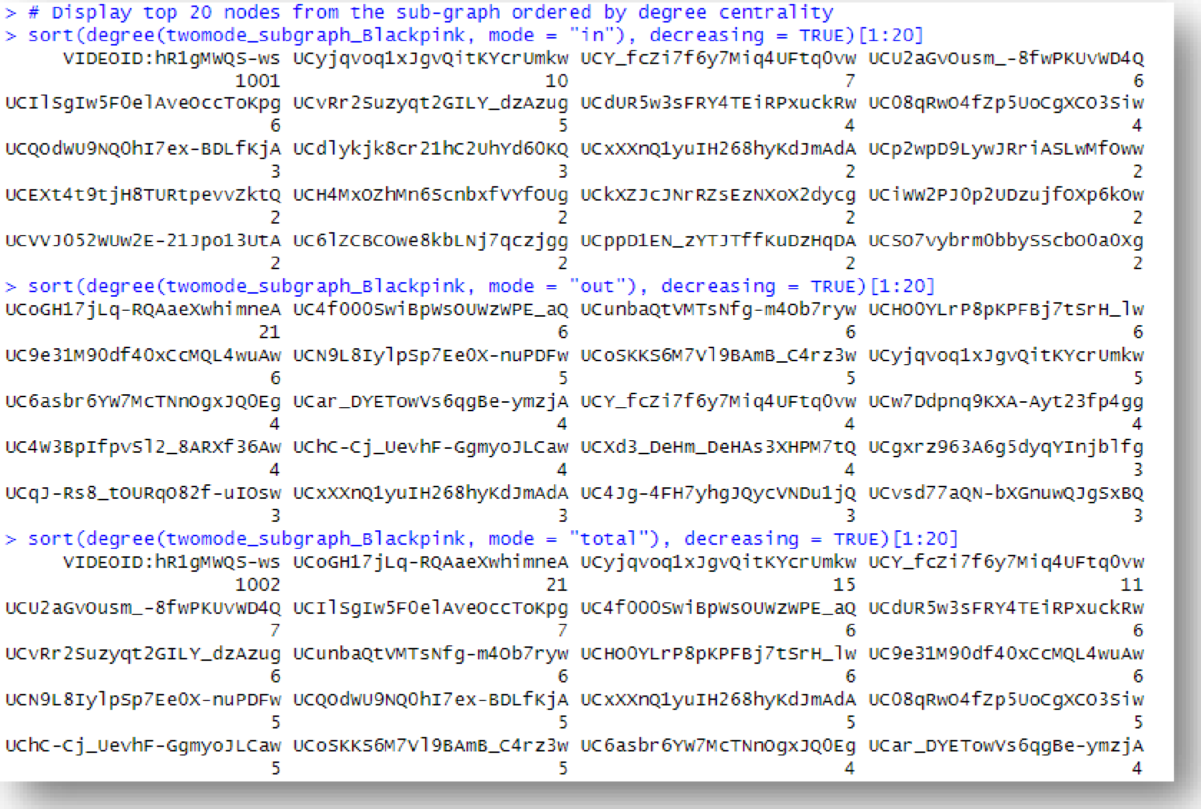
```
Blackpink_data <- youtubeAuth %>%Collect(videoID, maxComments = 1000,

writeToFile = TRUE)

twomode_network_Blackpink <- Blackpink_data |> Create("actor") |>

AddText(Blackpink_data)

twomode_graph_Blackpink <- twomode_network_Blackpink %>% Graph()
```

The number of nodes in each component:



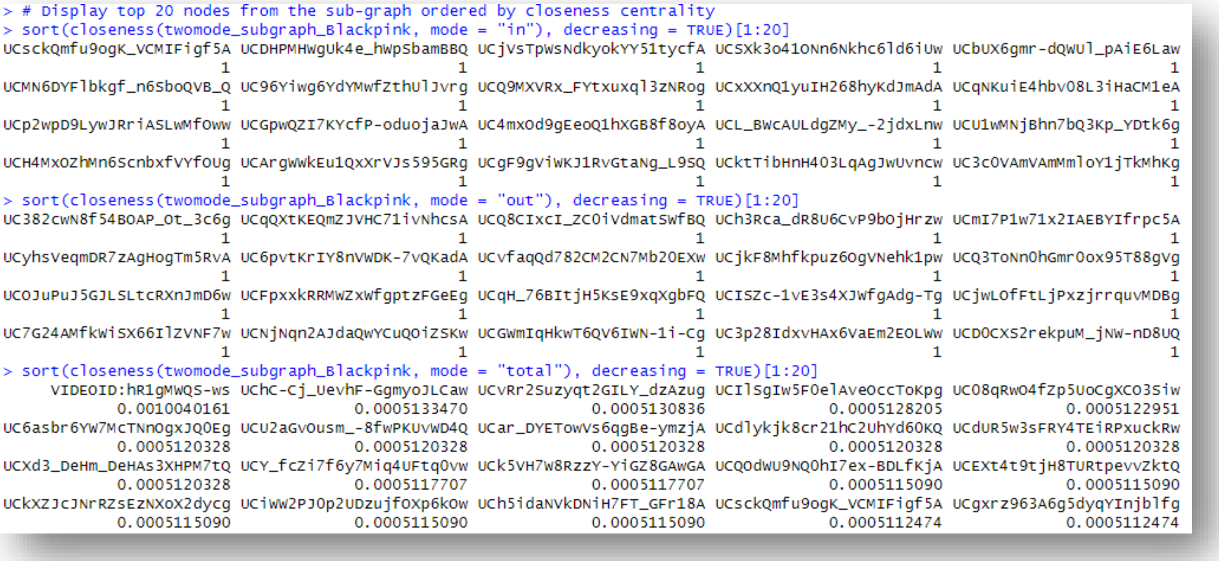
```
> twomode_comps_Blackpink$csize
[1] 964
```

Degree centrality



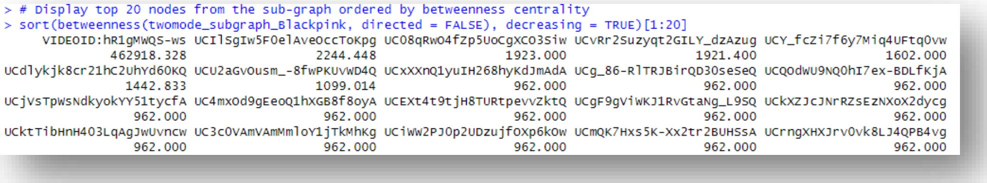
```
> # Display top 20 nodes from the sub-graph ordered by degree centrality
> sort(degree(twomode_subgraph_Blackpink, mode = "in"), decreasing = TRUE)[1:20]
     VIDEOID:hR1gMWQS-ws UCyjqvoq1xJgvQitKYcrUmkw UCY_fcZi7f6y7Miq4UFtq0vw UCU2aGvOusm_-8fwPKUvWD4Q
                    1001                       10                        7                        6
UCIlSgIw5FOelAveOccToKpg UCvRr2Suzyqt2GILY_dzAzug UCdUR5w3sFRY4TEiRPxuckRw UC08qRwO4fZp5UoCgXCO3Siw
                       6                        5                        4                        4
UCQOdWU9NQ0hI7ex-BDLfKjA UCdlykjk8cr21hC2UhYd60KQ UCxXXnQ1yuIH268hyKdJmAdA UCp2wpD9LywJRriASLwMfOww
                       3                        3                        2                        2
UCEXt4t9tjH8TURtpevvZktQ UCH4MxOZhMn6ScnbxfVYfOUg UCkXZJcJNrRZsEzNXoX2dycg UCiWw2PJ0p2UDzujfOXp6kOw
                       2                        2                        2                        2
UCVVJ052WUw2E-21Jpo13UtA UC6lZCBCOwe8kbLNj7qczjgg UCppD1EN_zYTJTffKuDzHqDA UCSO7vybrm0bbySScbO0a0Xg
                       2                        2                        2                        2
> sort(degree(twomode_subgraph_Blackpink, mode = "out"), decreasing = TRUE)[1:20]
UCoGH17jLq-RQAaeXwhimneA UC4f0O0SwiBpWsOUWzWPE_aQ UCunbaQtVMTsNfg-m4Ob7ryw UCHOOYLrP8pKPFBj7tSrH_lw
                      21                        6                        6                        6
UC9e31M90df40xCcMQL4wuAw UCN9L8IylpSp7Ee0X-nuPDFw UCoSKKS6M7Vl9BAmB_C4rz3w UCyjqvoq1xJgvQitKYcrUmkw
                       6                        5                        5                        5
UC6asbr6YW7McTNnOgxJQ0Eg UCar_DYETowVs6qgBe-ymzjA UCY_fcZi7f6y7Miq4UFtq0vw UCw7Ddpnq9KXA-Ayt23fp4gg
                       4                        4                        4                        4
UC4W3BpIfpvSl2_8ARXf36Aw UChC-Cj_UevhF-GgmyoJLCaw UCXd3_DeHm_DeHAs3XHPM7tQ UCgxrz963A6g5dyqYInjblfg
                       4                        4                        4                        3
UCqJ-Rs8_tOURqO82f-uIOsw UCxXXnQ1yuIH268hyKdJmAdA UC4Jg-4FH7yhgJQycVNDu1jQ UCvsd77aQN-bXGnuwQJgSxBQ
                       3                        3                        3                        3
> sort(degree(twomode_subgraph_Blackpink, mode = "total"), decreasing = TRUE)[1:20]
     VIDEOID:hR1gMWQS-ws UCoGH17jLq-RQAaeXwhimneA UCyjqvoq1xJgvQitKYcrUmkw UCY_fcZi7f6y7Miq4UFtq0vw
                    1002                       21                       15                       11
UCU2aGvOusm_-8fwPKUvWD4Q UCIlSgIw5FOelAveOccToKpg UC4f0O0SwiBpWsOUWzWPE_aQ UCdUR5w3sFRY4TEiRPxuckRw
                       7                        7                        6                        6
UCvRr2Suzyqt2GILY_dzAzug UCunbaQtVMTsNfg-m4Ob7ryw UCHOOYLrP8pKPFBj7tSrH_lw UC9e31M90df40xCcMQL4wuAw
                       6                        6                        6                        6
UCN9L8IylpSp7Ee0X-nuPDFw UCQOdWU9NQ0hI7ex-BDLfKjA UCxXXnQ1yuIH268hyKdJmAdA UC08qRwO4fZp5UoCgXCO3Siw
                       5                        5                        5                        5
UChC-Cj_UevhF-GgmyoJLCaw UCoSKKS6M7Vl9BAmB_C4rz3w UC6asbr6YW7McTNnOgxJQ0Eg UCar_DYETowVs6qgBe-ymzjA
                       5                        5                        4                        4
```

Closeness centrality



```
> # Display top 20 nodes from the sub-graph ordered by closeness centrality
> sort(closeness(twomode_subgraph_Blackpink, mode = "in"), decreasing = TRUE)[1:20]
UCsckQmfu9ogK_VCMIFigf5A UCDHPMHWgUk4e_hWpSbamBBQ UCjVsTpWsNdkyokYY51tycfA UCSXk3o41ONn6Nkhc6ld6iUw UCbUX6gmr-dQWUl_pAiE6Law 
                       1                        1                        1                        1                        1 
UCMN6DYFlbkgf_n6SboQVB_Q UC96Yiwg6YdYMwfZthUlJvrg UCQ9MXVRx_FYtxuxql3zNRog UCxXXnQ1yuIH268hyKdJmAdA UCqNKuiE4hbvO8L3iHaCM1eA 
                       1                        1                        1                        1                        1 
UCp2wpD9LywJRriASLwMfOww UCGpwQZI7KYcfP-oduojaJwA UC4mxOd9gEeoQ1hXGB8f8oyA UCL_BWcAULdgZMy_-2jdxLnw UCU1wMNjBhn7bQ3Kp_YDtk6g 
                       1                        1                        1                        1                        1 
UCH4MxOZhMn6ScnbxfVYfOUg UCArgWWkEu1QxXrVJs595GRg UCgF9gViWKJ1RvGtaNg_L9SQ UCktTibHnH403LqAgJwUvncw UC3cOVAmVAmMmloY1jTkMhKg 
                       1                        1                        1                        1                        1 
> sort(closeness(twomode_subgraph_Blackpink, mode = "out"), decreasing = TRUE)[1:20]
UC382cwN8f54BOAP_Ot_3c6g UCqQXtKEQmZJVHC71ivNhcsA UCQ8CIxcI_ZC0iVdmatSWfBQ UCh3Rca_dR8U6CvP9bOjHrzw UCmI7P1w71x2IAEBYIfrpc5A 
                       1                        1                        1                        1                        1 
UCyhsVeqmDR7zAgHogTm5RvA UC6pvtKrIY8nVWDK-7vQKadA UCvfaqQd782CM2CN7Mb20EXw UCjkF8Mhfkpuz6OgVNehk1pw UCQ3ToNn0hGmr0ox95T88gVg 
                       1                        1                        1                        1                        1 
UCOJuPuJ5GJLSLtcRXnJmD6w UCFpxxkRRMWZxWfgptzFGeEg UCqH_76BItjH5KsE9xqXgbFQ UCISZc-1vE3s4XJWfgAdg-Tg UCjwLOfFtLjPxzjrrquvMDBg 
                       1                        1                        1                        1                        1 
UC7G24AMfkWiSX66IlZVNF7w UCNjNqn2AJdaQwYCuQOiZSKw UCGWmIqHkwT6QV6IWN-1i-Cg UC3p28IdxvHAx6VaEm2EOLWw UCDOCXS2rekpuM_jNW-nD8UQ 
                       1                        1                        1                        1                        1 
> sort(closeness(twomode_subgraph_Blackpink, mode = "total"), decreasing = TRUE)[1:20]
     VIDEOID:hR1gMWQS-ws UChC-Cj_UevhF-GgmyoJLCaw UCvRr2Suzyqt2GILY_dzAzug UCIlSgIw5F0elAveOccToKpg UC08qRwO4fZp5UoCgXCO3Siw 
            0.0010040161             0.0005133470             0.0005130836             0.0005128205             0.0005122951 
UC6asbr6YW7McTNnOgxJQ0Eg UCU2aGvOusm_-8fwPKUvWD4Q UCar_DYETowVs6qgBe-ymzjA UCdlykjk8cr21hC2UhYd60KQ UCdUR5w3sFRY4TEiRPxuckRw 
            0.0005120328             0.0005120328             0.0005120328             0.0005120328             0.0005120328 
UCXd3_DeHm_DeHAs3XHPM7tQ UCY_fcZi7f6y7Miq4UFtq0vw UCk5VH7W8RzzY-YiGZ8GAwGA UCQOdWU9NQ0hI7ex-BDLfKjA UCEXt4t9tjH8TURtpevvZktQ 
            0.0005120328             0.0005117707             0.0005117707             0.0005115090             0.0005115090 
UCkXZJcJNrRZsEzNXoX2dycg UCiWW2PJOp2UDzujfOXp6kOw UCh5idaNVkDNiH7FT_GFr18A UCsckQmfu9ogK_VCMIFigf5A UCgxrz963A6g5dyqYInjblfg 
            0.0005115090             0.0005115090             0.0005115090             0.0005112474             0.0005112474 
```

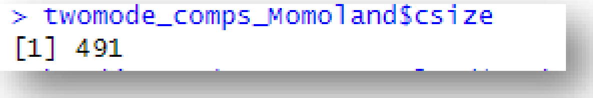
Betweenness centrality



```
> # Display top 20 nodes from the sub-graph ordered by betweenness centrality
> sort(betweenness(twomode_subgraph_Blackpink, directed = FALSE), decreasing = TRUE)[1:20]
     VIDEOID:hR1gMWQS-ws UCIlSgIw5F0elAveOccToKpg UC08qRwO4fZp5UoCgXCO3Siw UCvRr2Suzyqt2GILY_dzAzug UCY_fcZi7f6y7Miq4UFtq0vw 
              462918.328                 2244.448                 1923.000                 1921.400                 1602.000 
UCdlykjk8cr21hC2UhYd60KQ UCU2aGvOusm_-8fwPKUvWD4Q UCxXXnQ1yuIH268hyKdJmAdA UCg_86-RlTRJBirQD3OseSeQ UCQOdWU9NQ0hI7ex-BDLfKjA 
                1442.833                 1099.014                  962.000                  962.000                  962.000 
UCjVsTpWsNdkyokYY51tycfA UC4mxOd9gEeoQ1hXGB8f8oyA UCEXt4t9tjH8TURtpevvZktQ UCgF9gViWKJ1RvGtaNg_L9SQ UCkXZJcJNrRZsEzNXoX2dycg 
                 962.000                  962.000                  962.000                  962.000                  962.000 
UCktTibHnH403LqAgJwUvncw UC3cOVAmVAmMmloY1jTkMhKg UCiWW2PJOp2UDzujfOXp6kOw UCmQK7Hxs5K-Xx2tr2BUHSsA UCrngXHXJrvOvk8LJ4QPB4vg 
                 962.000                  962.000                  962.000                  962.000                  962.000 
```

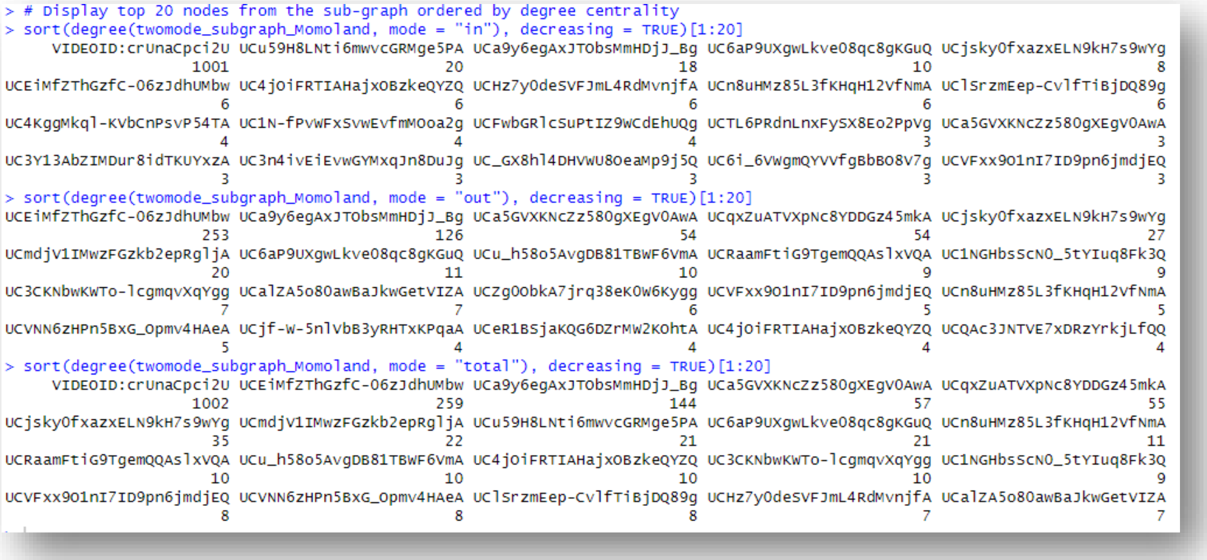
Momoland

The number of nodes in each component:



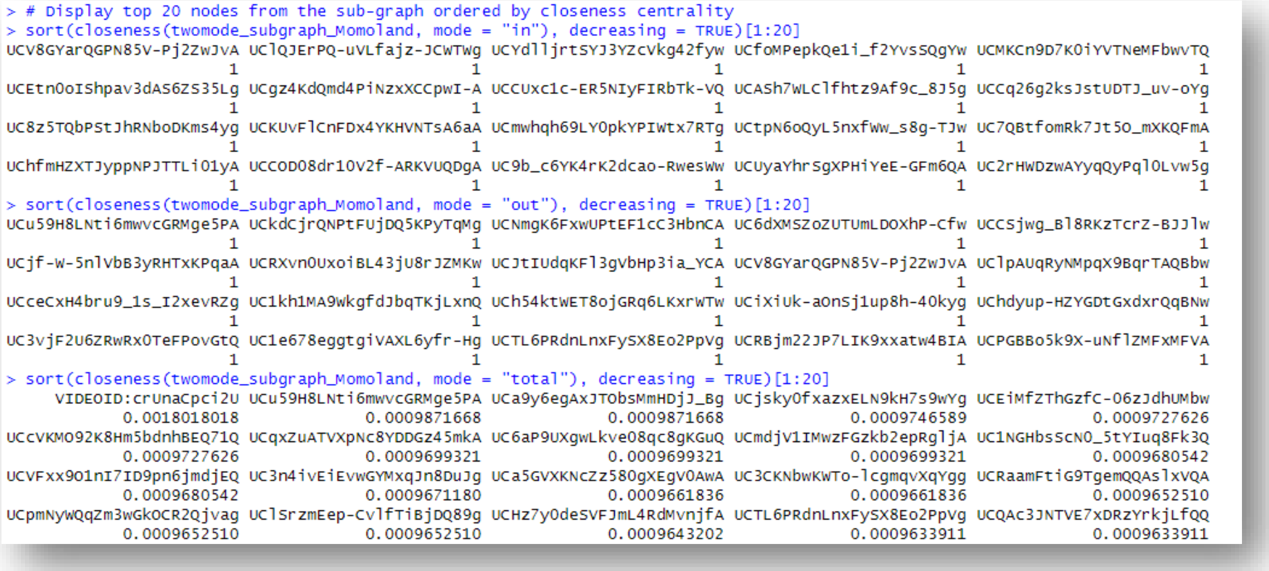
```
> twomode_comps_Momoland$csize
[1] 491
```

Degree centrality



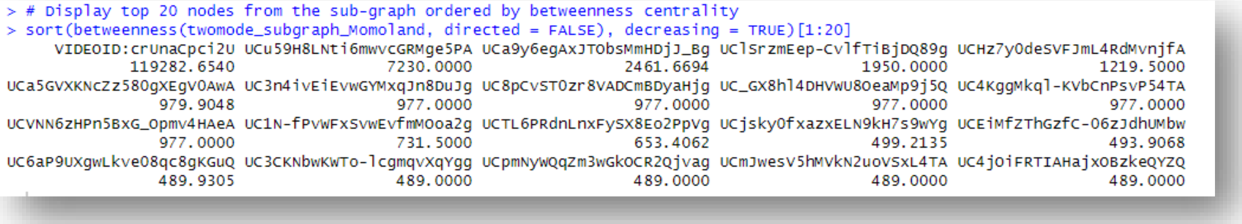
```
> # Display top 20 nodes from the sub-graph ordered by degree centrality
> sort(degree(twomode_subgraph_Momoland, mode = "in"), decreasing = TRUE)[1:20]
     VIDEOID:crUnaCpci2U UCu59H8LNti6mwvcGRMge5PA UCa9y6egAxJTObsMmHDjJ_Bg UC6aP9UXgwLkve08qc8gKGuQ UCjsky0fxazxELN9kH7s9wYg 
                    1001                       20                       18                       10                        8 
UCEiMfZThGzfC-06zJdhUMbw UC4jOiFRTIAHajxOBzkeQYZQ UCHz7y0deSVFJmL4RdMvnjfA UCn8uHMz85L3fKHqH12VfNmA UClSrzmEep-CvlfTiBjDQ89g 
                       6                        6                        6                        6                        6 
UC4KggMkql-KVbCnPsvP54TA UC1N-fPvWFxSvwEvfmMOoa2g UCFwbGRlcSuPtIZ9WCdEhUQg UCTL6PRdnLnxFySX8Eo2PpVg UCa5GVXKNcZz580gXEgV0AwA 
                       4                        4                        4                        3                        3 
UC3Y13AbZIMDur8idTKUYxzA UC3n4ivEiEvwGYMxqJn8DuJg UC_GX8hl4DHVWU8OeaMp9j5Q UC6i_6VWgmQYVVfgBbBO8V7g UCVFxx901nI7ID9pn6jmdjEQ 
                       3                        3                        3                        3                        3 
> sort(degree(twomode_subgraph_Momoland, mode = "out"), decreasing = TRUE)[1:20]
UCEiMfZThGzfC-06zJdhUMbw UCa9y6egAxJTObsMmHDjJ_Bg UCa5GVXKNcZz580gXEgV0AwA UCqxZuATVXpNc8YDDGz45mkA UCjsky0fxazxELN9kH7s9wYg 
                     253                      126                       54                       54                       27 
UCmdjV1IMwzFGzkb2epRgljA UC6aP9UXgwLkve08qc8gKGuQ UCu_h58o5AvgDB81TBWF6VmA UCRaamFtiG9TgemQQAslxVQA UC1NGHbsScN0_5tYIuq8Fk3Q 
                      20                       11                       10                        9                        9 
UC3CKNbwKWTo-lcgmqvXqYgg UCalZA5o80awBaJkwGetVIZA UCZg0ObkA7jrq38eK0W6Kygg UCVFxx901nI7ID9pn6jmdjEQ UCn8uHMz85L3fKHqH12VfNmA 
                       7                        7                        6                        5                        5 
UCVNN6zHPn5BxG_Opmv4HAeA UCjf-W-5nlVbB3yRHTxKPqaA UCeR1BSjaKQG6DZrMW2KOhtA UC4jOiFRTIAHajxOBzkeQYZQ UCQAc3JNTVE7xDRzYrkjLfQQ 
                       5                        4                        4                        4                        4 
> sort(degree(twomode_subgraph_Momoland, mode = "total"), decreasing = TRUE)[1:20]
     VIDEOID:crUnaCpci2U UCEiMfZThGzfC-06zJdhUMbw UCa9y6egAxJTObsMmHDjJ_Bg UCa5GVXKNcZz580gXEgV0AwA UCqxZuATVXpNc8YDDGz45mkA 
                    1002                      259                      144                       57                       55 
UCjsky0fxazxELN9kH7s9wYg UCmdjV1IMwzFGzkb2epRgljA UCu59H8LNti6mwvcGRMge5PA UC6aP9UXgwLkve08qc8gKGuQ UCn8uHMz85L3fKHqH12VfNmA 
                      35                       22                       21                       21                       11 
UCRaamFtiG9TgemQQAslxVQA UCu_h58o5AvgDB81TBWF6VmA UC4jOiFRTIAHajxOBzkeQYZQ UC3CKNbwKWTo-lcgmqvXqYgg UC1NGHbsScN0_5tYIuq8Fk3Q 
                      10                       10                       10                       10                        9 
UCVFxx901nI7ID9pn6jmdjEQ UCVNN6zHPn5BxG_Opmv4HAeA UClSrzmEep-CvlfTiBjDQ89g UCHz7y0deSVFJmL4RdMvnjfA UCalZA5o80awBaJkwGetVIZA 
                       8                        8                        8                        7                        7 
```

Betweenness centrality



```
> # Display top 20 nodes from the sub-graph ordered by closeness centrality
> sort(closeness(twomode_subgraph_Momoland, mode = "in"), decreasing = TRUE)[1:20]
UCV8GYarQGPN85V-Pj2ZwJvA UClQJErPQ-uVLfajz-JCWTwg UCYdlljrtSYJ3YZcvkg42fyw UCfoMPepkQe1i_f2YvsSQgYw UCMKCn9D7K0iYVTNeMFbwvTQ 
                       1                        1                        1                        1                        1 
UCEtn0oIShpav3dAS6ZS35Lg UCgz4KdQmd4PiNzxXCCpwI-A UCCUxc1c-ER5NIyFIRbTk-VQ UCASh7WLClfhtz9Af9c_8J5g UCCq26g2ksJstUDTJ_uv-oYg 
                       1                        1                        1                        1                        1 
UC8z5TQbPStJhRNboDKms4yg UCKUvFlCnFDx4YKHVNTsA6aA UCmwhqh69LY0pkYPIWtx7RTg UCtpN6oQyL5nxfWw_s8g-TJw UC7QBtfomRk7Jt5O_mXKQFmA 
                       1                        1                        1                        1                        1 
UChfmHZXTJyppNPJTTLi01yA UCCOD08dr10V2f-ARKVUQDgA UC9b_c6YK4rK2dcao-RwesWw UCUyaYhrSgXPHiYeE-GFm6QA UC2rHWDzwAYyqQyPql0Lvw5g 
                       1                        1                        1                        1                        1 
> sort(closeness(twomode_subgraph_Momoland, mode = "out"), decreasing = TRUE)[1:20]
UCu59H8LNti6mwvcGRMge5PA UCkdCjrQNPtFUjDQ5KPyTqMg UCNmgK6FxwUPtEF1cC3HbnCA UC6dXMSZoZUTUmLDOXhP-Cfw UCCSjwg_Bl8RKzTcrZ-BJJlw 
                       1                        1                        1                        1                        1 
UCjf-W-5nlVbB3yRHTxKPqaA UCRXvn0UxoiBL43jU8rJZMKw UCJtIUdqKFl3gVbHp3ia_YCA UCV8GYarQGPN85V-Pj2ZwJvA UClpAUqRyNMpqX9BqrTAQBbw 
                       1                        1                        1                        1                        1 
UCceCxH4bru9_1s_I2xevRZg UC1kh1MA9WkgfdJbqTKjLxnQ UCh54ktWET8ojGRq6LKxrWTw UCiXiUk-aOnSj1up8h-40kyg UChdyup-HZYGDtGxdxrQqBNw 
                       1                        1                        1                        1                        1 
UC3vjF2U6ZRwRx0TeFPovGtQ UC1e678eggtgiVAXL6yfr-Hg UCTL6PRdnLnxFySX8Eo2PpVg UCRBjm22JP7LIK9xxatw4BIA UCPGBBo5k9X-uNflZMFxMFVA 
                       1                        1                        1                        1                        1 
> sort(closeness(twomode_subgraph_Momoland, mode = "total"), decreasing = TRUE)[1:20]
     VIDEOID:crUnaCpci2U UCu59H8LNti6mwvcGRMge5PA UCa9y6egAxJTObsMmHDjJ_Bg UCjsky0fxazxELN9kH7s9wYg UCEiMfZThGzfC-06zJdhUMbw 
            0.0018018018             0.0009871668             0.0009871668             0.0009746589             0.0009727626 
UCcVKMO92K8Hm5bdnhBEQ71Q UCqxZuATVXpNc8YDDGz45mkA UC6aP9UXgwLkve08qc8gKGuQ UCmdjV1IMwzFGzkb2epRgljA UC1NGHbsScN0_5tYIuq8Fk3Q 
            0.0009727626             0.0009699321             0.0009699321             0.0009699321             0.0009680542 
UCVFxx901nI7ID9pn6jmdjEQ UC3n4ivEiEvwGYMxqJn8DuJg UCa5GVXKNcZz580gXEgV0AwA UC3CKNbwKWTo-lcgmqvXqYgg UCRaamFtiG9TgemQQAslxVQA 
            0.0009680542             0.0009671180             0.0009661836             0.0009661836             0.0009652510 
UCpmNyWQqZm3wGkOCR2Qjvag UClSrzmEep-CvlfTiBjDQ89g UCHz7y0deSVFJmL4RdMvnjfA UCTL6PRdnLnxFySX8Eo2PpVg UCQAc3JNTVE7xDRzYrkjLfQQ 
            0.0009652510             0.0009652510             0.0009643202             0.0009633911             0.0009633911 
```

Closeness centrality



```
> # Display top 20 nodes from the sub-graph ordered by betweenness centrality
> sort(betweenness(twomode_subgraph_Momoland, directed = FALSE), decreasing = TRUE)[1:20]
    VIDEOID:crUnaCpci2U UCu59H8LNti6mwvcGRMge5PA UCa9y6egAxJTObsMmHDjJ_Bg UClSrzmEep-CvlfTiBjDQ89g UCHz7y0deSVFJmL4RdMvnjfA 
            119282.6540                7230.0000                2461.6694                1950.0000                1219.5000 
UCa5GVXKNcZz580gXEgV0AwA UC3n4ivEiEvwGYMxqJn8DuJg UC8pCvST0zr8VADCmBDyaHjg UC_GX8hl4DHVWU8OeaMp9j5Q UC4KggMkql-KVbCnPsvP54TA 
               979.9048                 977.0000                 977.0000                 977.0000                 977.0000 
UCVNN6zHPn5BxG_Opmv4HAeA UC1N-fPvWFxSvwEvfmMOoa2g UCTL6PRdnLnxFySX8Eo2PpVg UCjsky0fxazxELN9kH7s9wYg UCEiMfZThGzfC-06zJdhUMbw 
               977.0000                 731.5000                 653.4062                 499.2135                 493.9068 
UC6aP9UXgwLkve08qc8gKGuQ UC3CKNbwKWTo-lcgmqvXqYgg UCpmNyWQqZm3wGkOCR2Qjvag UCmJwesV5hMVkN2uoVSxL4TA UC4jOiFRTIAHajxOBzkeQYZQ 
               489.9305                 489.0000                 489.0000                 489.0000                 489.0000 
```

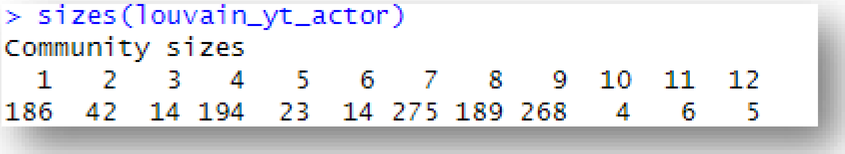
Comparation between the bands

		Twice	Blackpink	Momoland
the number of connected components		29	1	1
the highest rate of node in sub-graph ordered by degree centrality	in	58	1001	1001
	out	207	21	253
	total	207	1002	1002
the highest rate of node in sub-graph ordered by closeness centrality	in	1	1	1
	out	1	1	1
	total	0.01736111	0.0010040161	0.0018018018
the highest rate of node in sub-graph ordered by betweenness centrality		7984.2243	462918.328	119282.6540

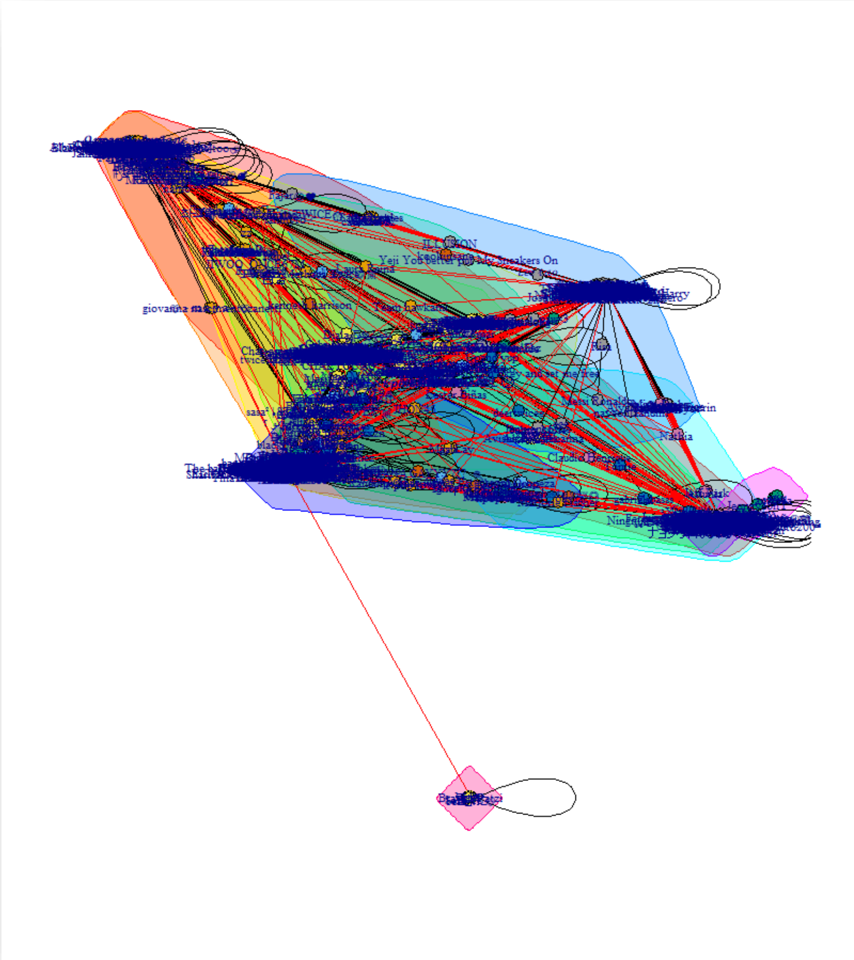
Compared to Blackpink and Momoland, the number of connected components in Twice ranked the top, indicating better fragmentation, less global connectivity, enhanced modularity. Besides, the lower proportion in degree centrality further supports the idea of reduced global connectivity in Twice. With higher rate by closeness centrality, it is also deemed higher-level efficiency [3], connectivity, and accessibility of the network [2]. Additionally, the higher figure for betweenness centrality benefits to identify the nodes that act as crucial intermediaries and potential points of vulnerability [3]. Thus, a 2-mode network is more suitable for centrality analysis than an actor network.

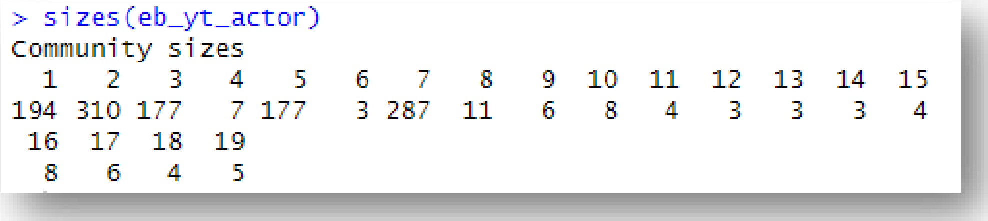
2.5. Community analysis

Twice

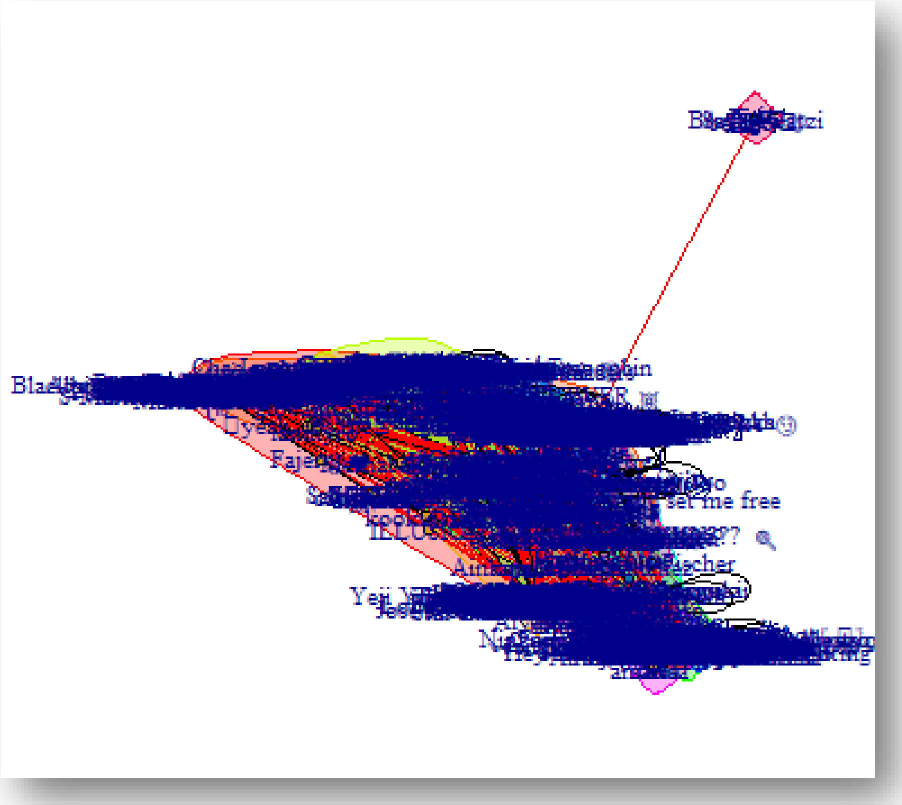


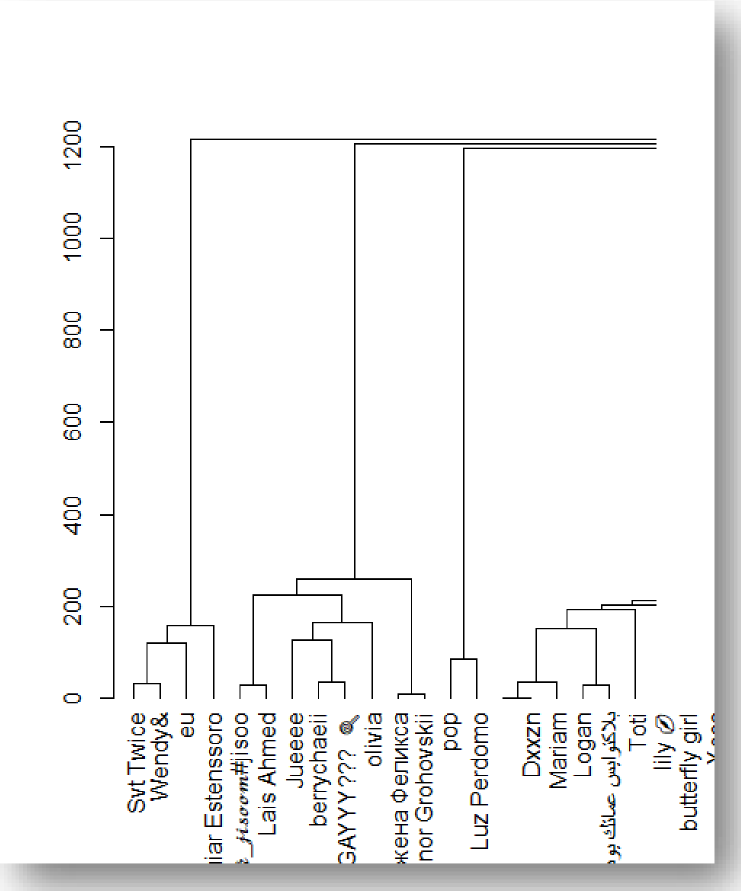
```
> sizes(louvain_yt_actor)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12 
186  42  14 194  23  14 275 189 268   4   6   5 
```



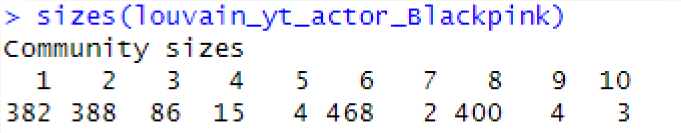


```
> sizes(eb_yt_actor)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15 
194 310 177   7 177   3 287  11   6   8   4   3   3   3   4 
 16  17  18  19 
  8   6   4   5 
```

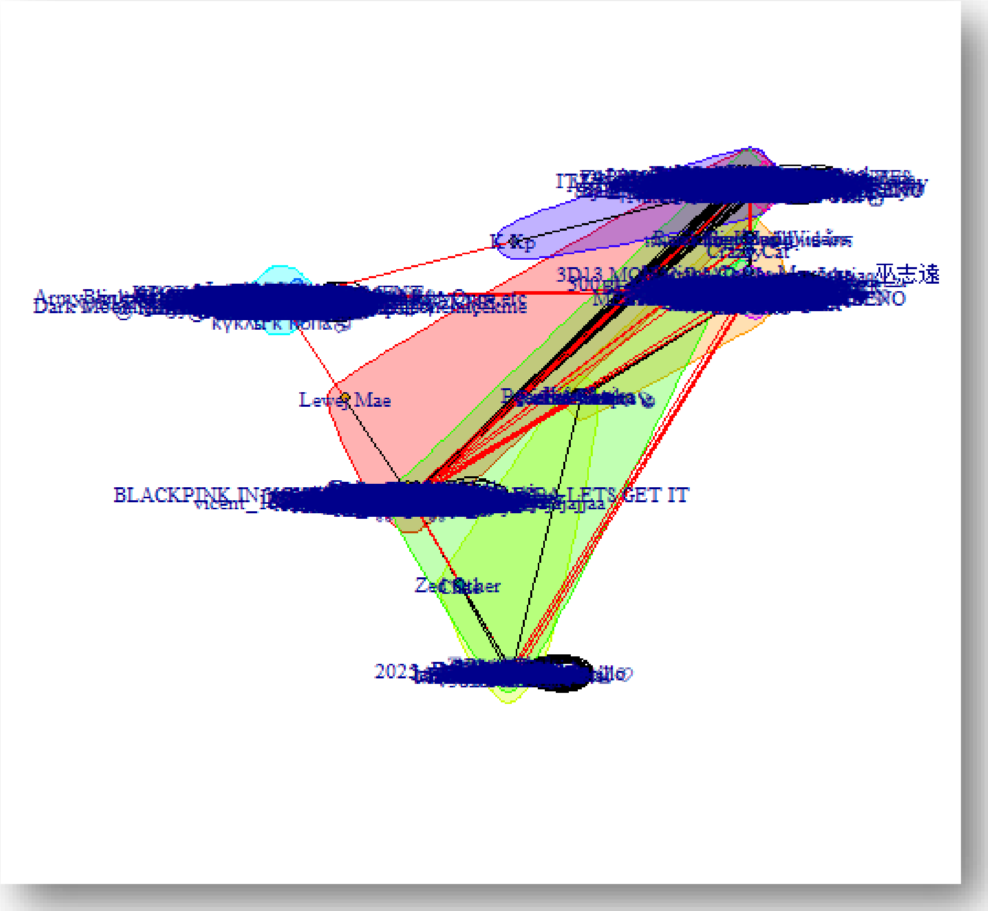


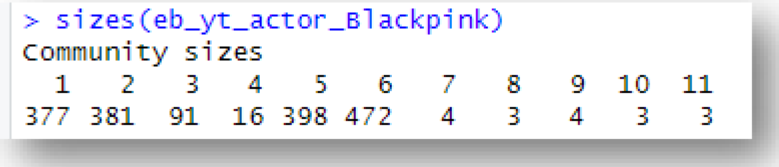

0
200
400
600
800
1000
1200
Svt Twice
Wendy&
eu
iar Estenssoro
_jisoom#jisoo
Lais Ahmed
Jueeee
berrychaeii
GAYYY???
olivia
жена Феликса
nor Grohovskii
pop
Luz Perdomo
Dxxzn
Mariam
Logan
بلاكوابس عماك بو
Toti
lily
butterfly girl

Blackpink

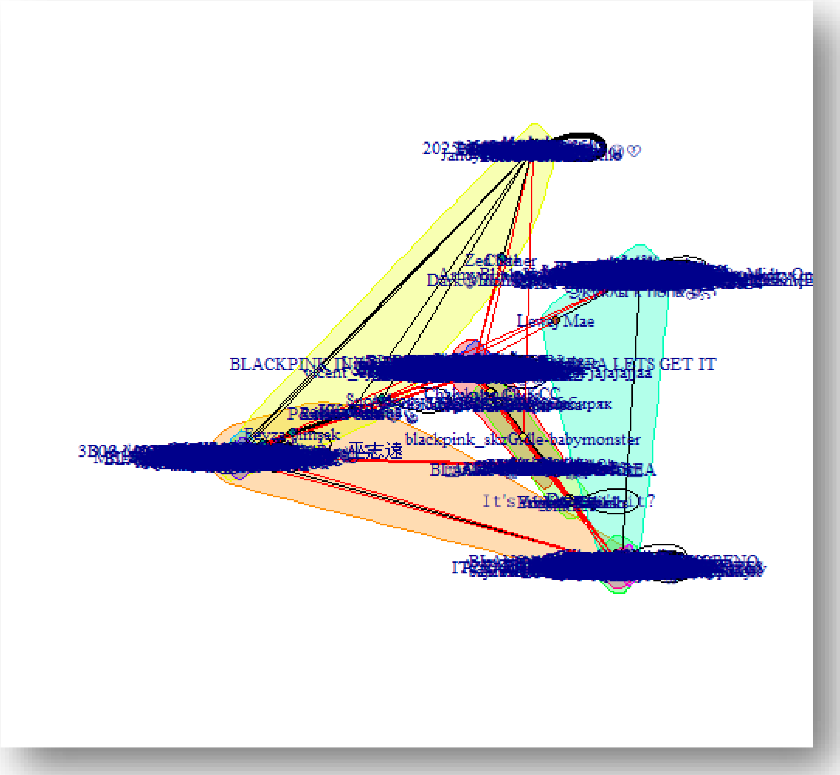


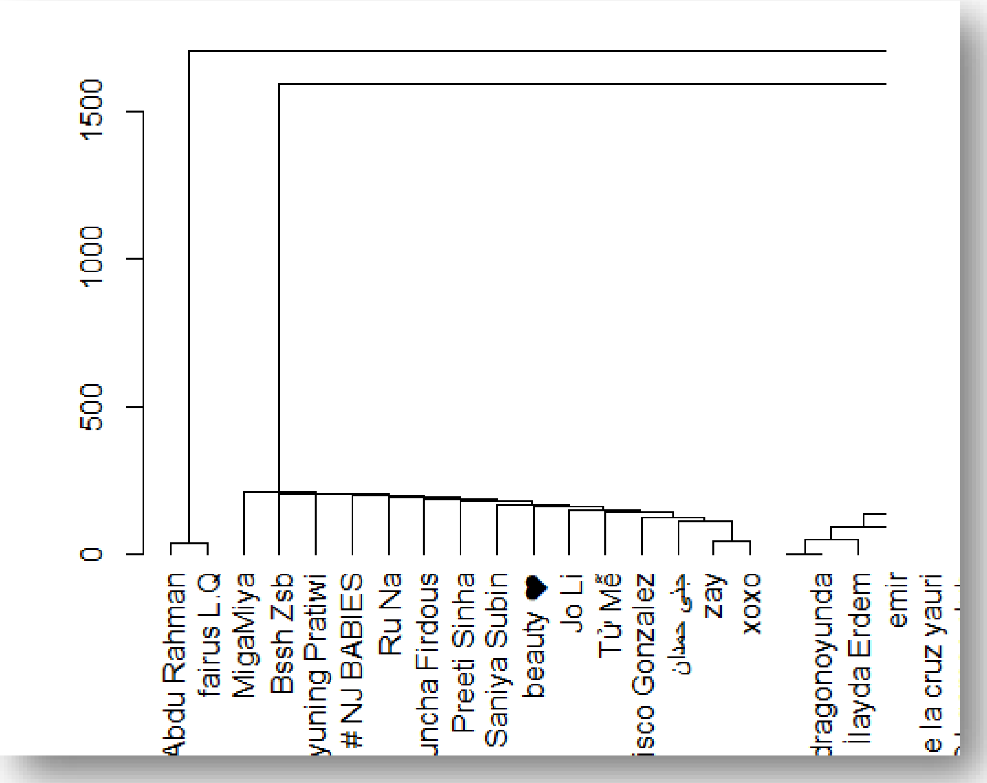
```
> sizes(louvain_yt_actor_Blackpink)
Community sizes
  1   2   3   4   5   6   7   8   9  10 
382 388  86  15   4 468   2 400   4   3 
```



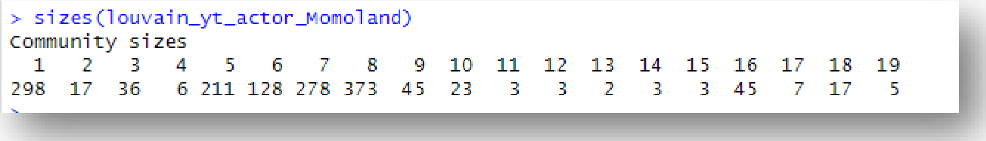


```
> sizes(eb_yt_actor_Blackpink)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11 
377 381  91  16 398 472   4   3   4   3   3 
```

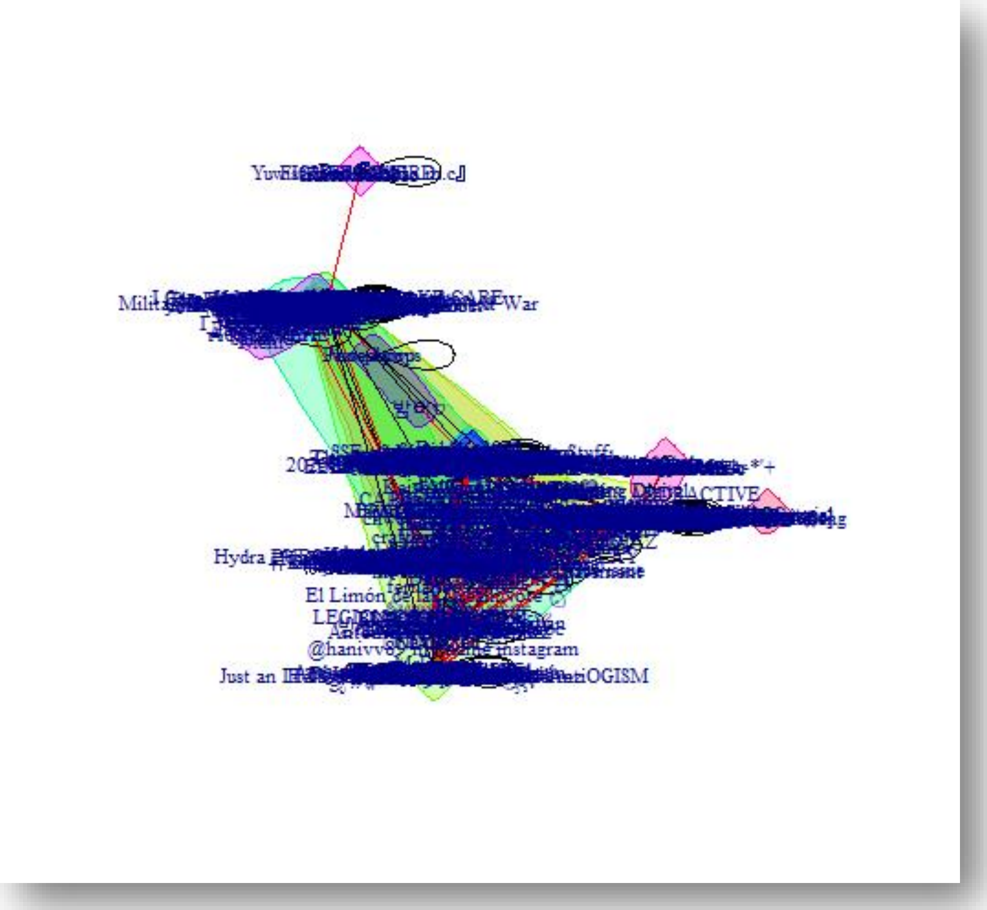


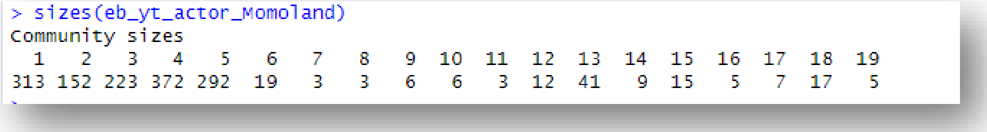

0
500
1000
1500
Abdu Rahman
fairus L.Q
MigaMiya
Bssh Zsb
yuning Pratiwi
# NJ BABIES
Ru Na
uncha Firdous
Preeti Sinha
Saniya Subin
beauty ❤
Jo Li
Tử Mễ
isco Gonzalez
جنى حمدان
zay
xoxo
dragonoyunda
İlayda Erdem
emir
e la cruz yauri

Momoland

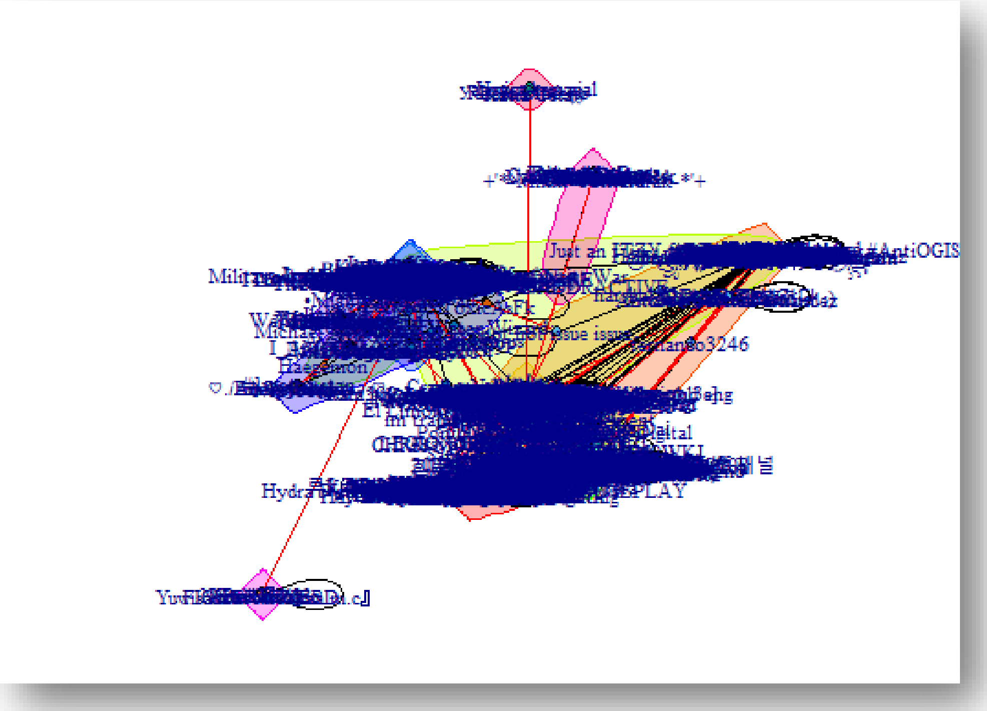


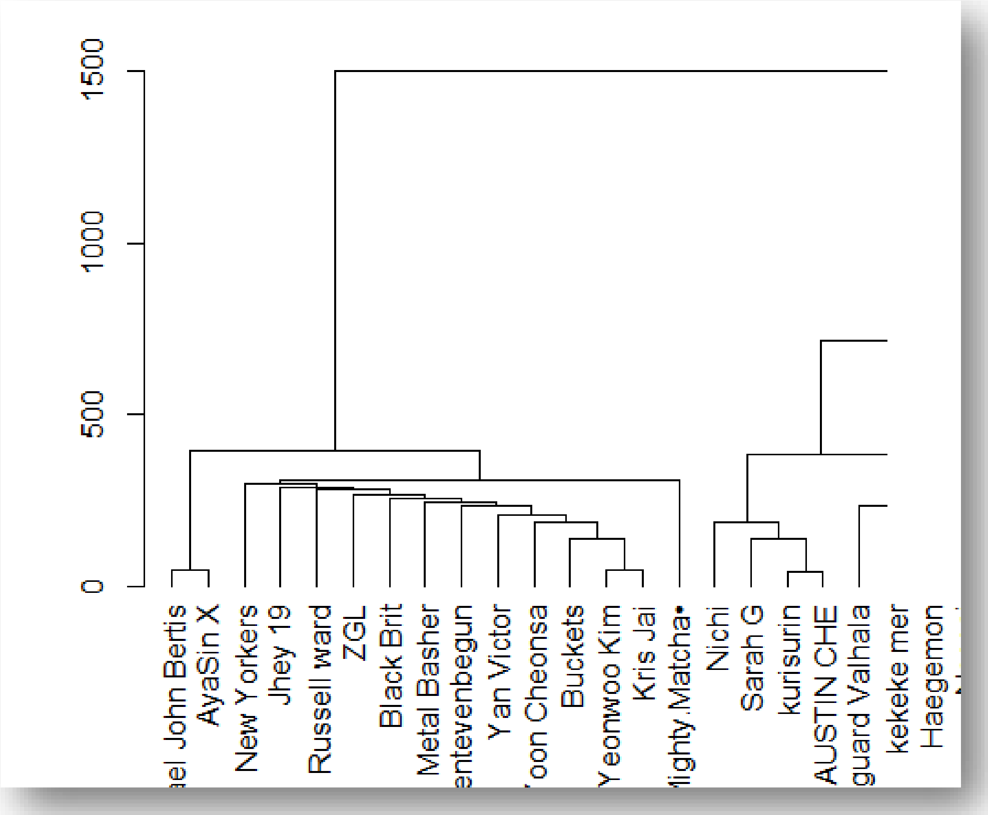
```
> sizes(louvain_yt_actor_Momoland)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
298  17  36   6 211 128 278 373  45  23   3   3   2   3   3  45   7  17   5
```





```
> sizes(eb_yt_actor_Momoland)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19 
313 152 223 372 292  19   3   3   6   6   3  12  41   9  15   5   7  17   5 
```





Comparation between the bands

	The number of communities	
	Louvain analysis	Girvan-Newman
Twice	12	19
Blackpink	10	11
Momoland	19	19

Communities by Louvain analysis are less than or equal to those of Girvan-Newman in all musical bands. This means that Louvain analysis is producing a coarser or more generalized division of the network compared to Girvan-Newman.

Machine Learning Models

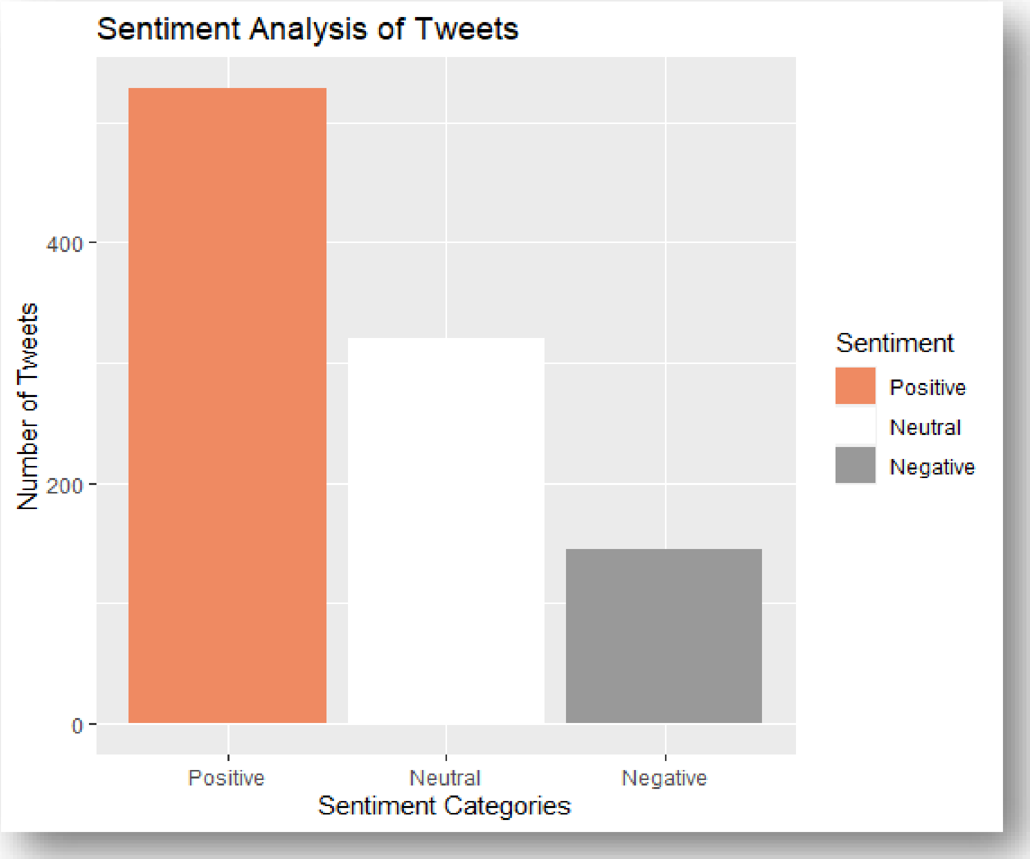
2.6. Sentiment analysis

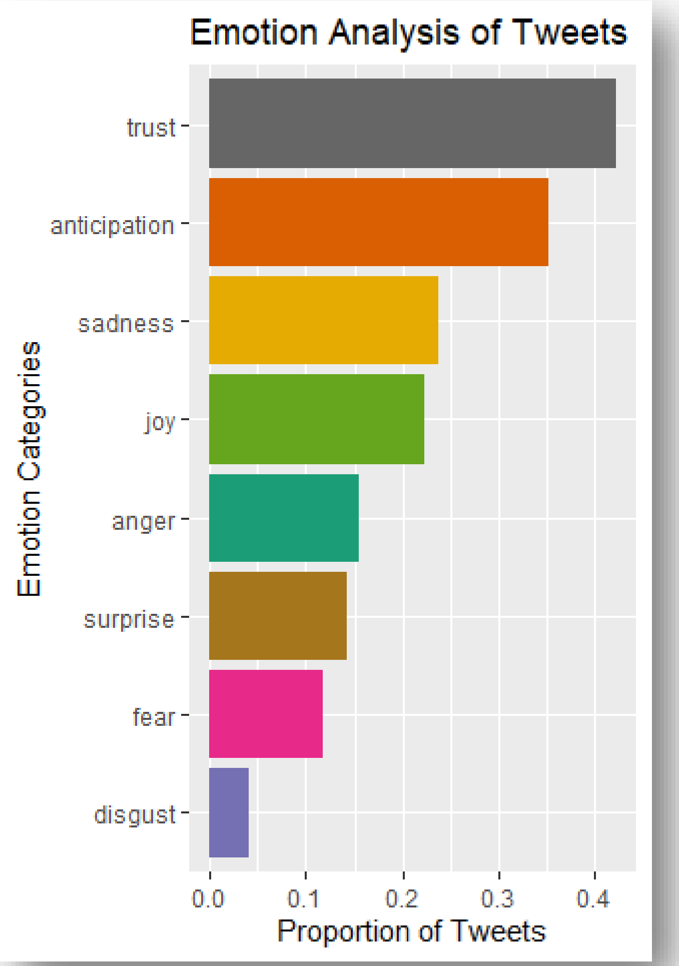
In the sentiment analysis, after cleaning tweet text, *get_sentiment()* is applied to match each word to its sentiment score in a sentiment lexicon. In there, the scores include 3 different types: -1(negative), 0 (neutral), 1 (positive).

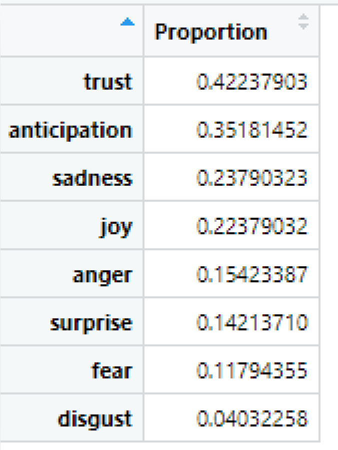
In the Emotion analysis, *get_nrc_sentiment()* function is used to get emotion with emotion scores as well as sentiment score.

Below are the plots of sentiment and emotion analysis in the three bands.

Twice

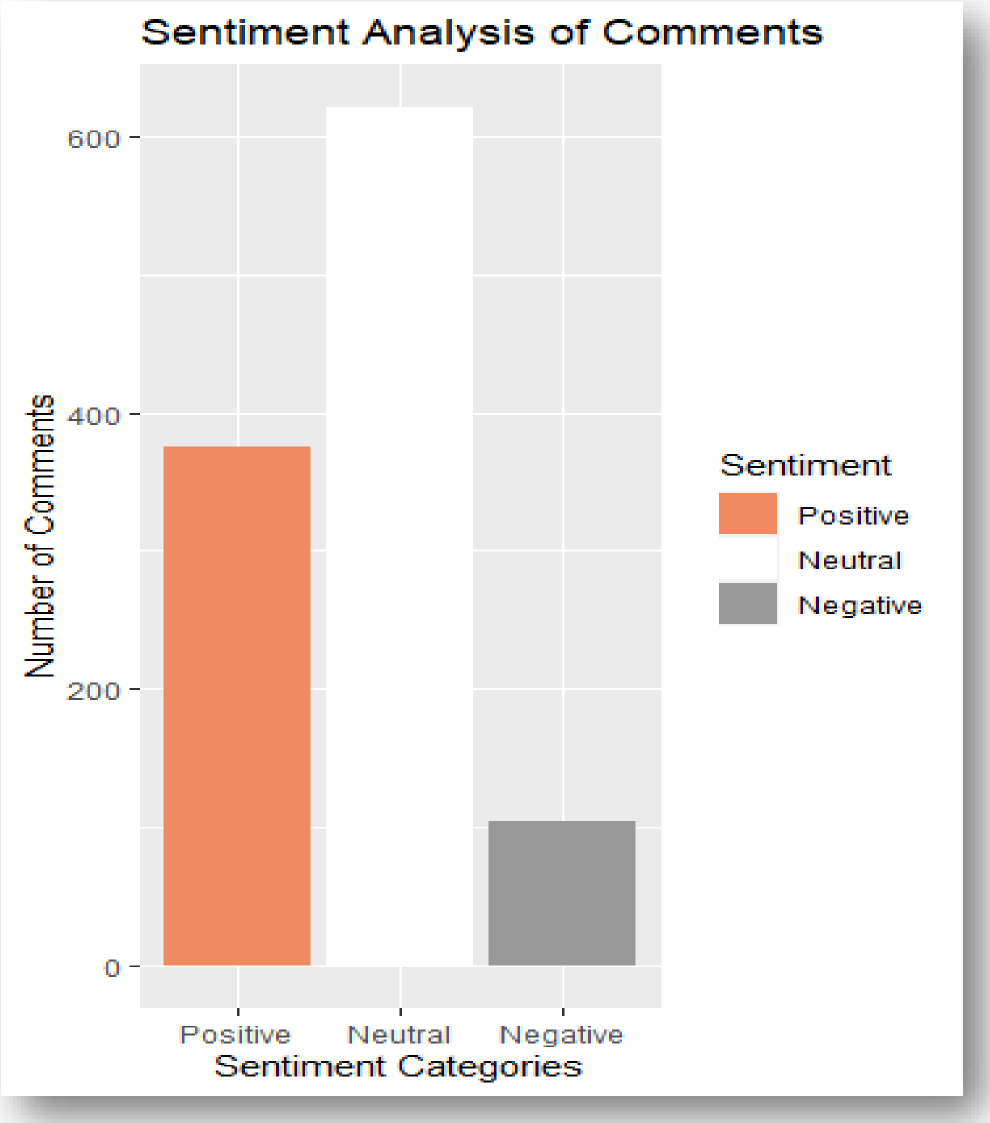


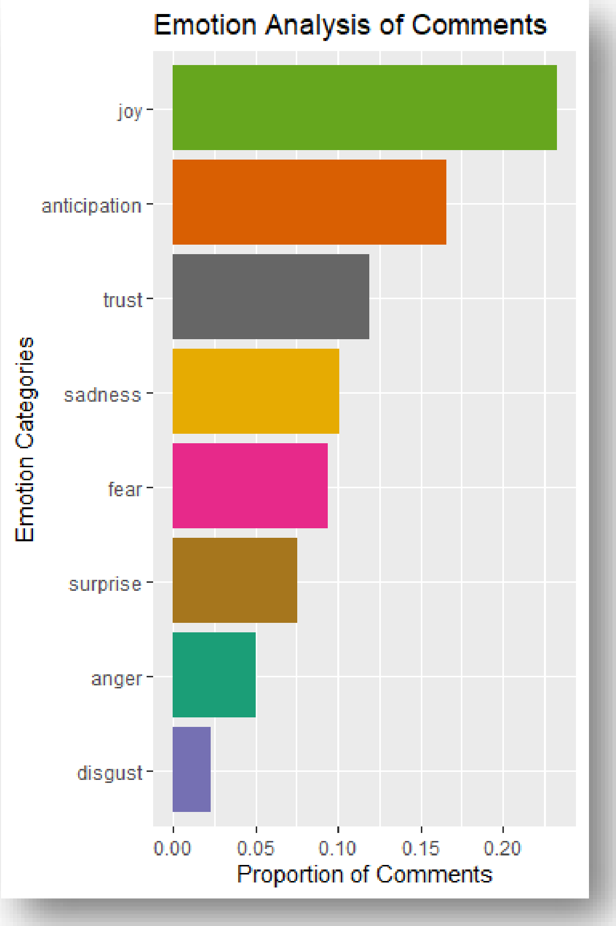


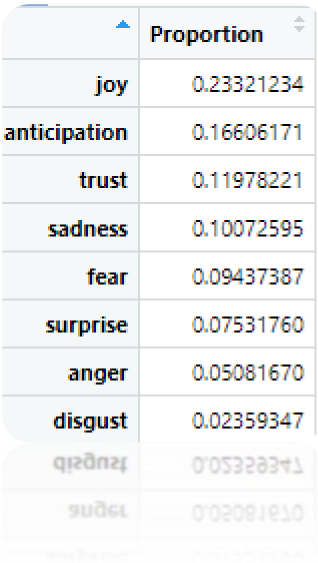


	Proportion
trust	0.42237903
anticipation	0.35181452
sadness	0.23790323
joy	0.22379032
anger	0.15423387
surprise	0.14213710
fear	0.11794355
disgust	0.04032258

Blackpink

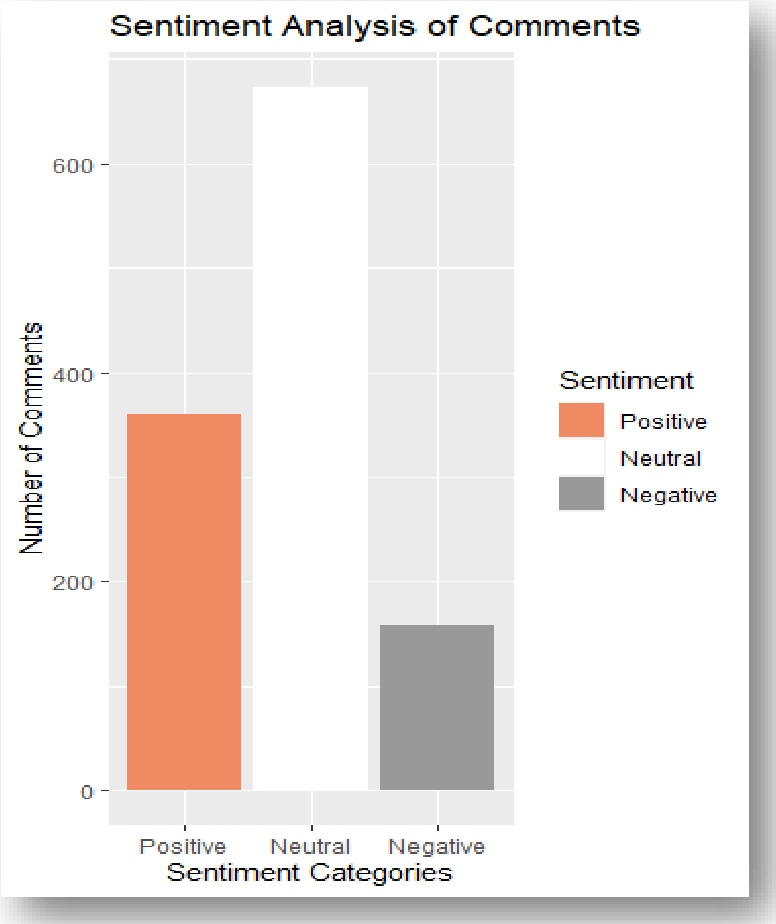

Sentiment Analysis of Comments
Number of Comments
600
400
200
0
Positive
Neutral
Negative
Sentiment Categories
Sentiment
Positive
Neutral
Negative

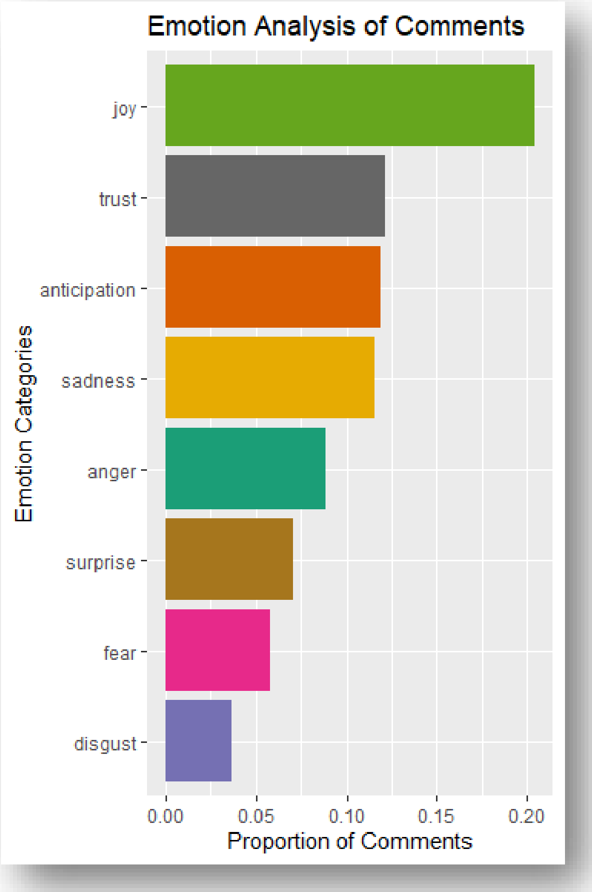


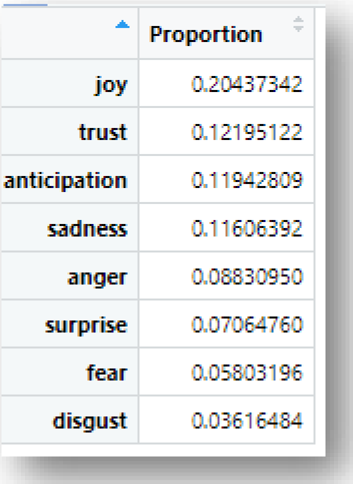


	Proportion
joy	0.23321234
anticipation	0.16606171
trust	0.11978221
sadness	0.10072595
fear	0.09437387
surprise	0.07531760
anger	0.05081670
disgust	0.02359347

Momoland


Sentiment Analysis of Comments
Number of Comments
0
200
400
600
Positive
Neutral
Negative
Sentiment Categories
Sentiment
Positive
Neutral
Negative





	Proportion
joy	0.20437342
trust	0.12195122
anticipation	0.11942809
sadness	0.11606392
anger	0.08830950
surprise	0.07064760
fear	0.05803196
disgust	0.03616484

Comparation between the bands

In sentiment analysis graphs, the positive rate highest in the Twice plot, indicating a greater expression of positive emotions, opinions, or attitudes. Meanwhile, the graphs for Blackpink and Momoland show more prominent figures for neutral sentiment, suggesting that the text in those cases is neither strongly positive nor strongly negative. This is noteworthy that a similarity among all three graphs is that the rate of negative sentiment is lower compared to other categories.

Regarding emotion analysis, the graph for Twice stands out in terms of the emotion of "trust." This suggests that the individuals or sources mentioned in the text related to Twice are perceived as trustworthy, dependable, or credible. Additionally, the graphs for Blackpink and Momoland show higher figures for the emotion of "joy" compared to other categories. This is also good point that the proportion of "disgust" across all charts is consistently the lowest.

2.7. Decision tree

A decision tree model can classify whether it is a song of the artist/band.

Twice

To construct a decision tree model for Twice, `get_artist_audio_features` function is used to retrieve data on the audio feature of their songs. Subsequently, certain columns from the database are eliminated, as the focus is solely on audio features and track ID.

```
Twice_features <- get_artist_audio_features("Twice")

data.frame(colnames(Twice_features))

Twice_features_subset <- Twice_features[ , 9:20]
```

In order to train the model on non-Twice songs, Spotify 100 playlist is obtained with only those columns containing the audio features and track ID.

```
top100_features <- get_playlist_audio_features("spotify",
"4hOKQuZbraPDIfaGbM3lKI")

data.frame(colnames(top100_features))

top100_features_subset <- top100_features[ , 6:17]

top100_features_subset <- top100_features_subset %>% rename(track_id =
track.id)
```

To indicate 1 for songs owned by Twice and 0 for which are not:

```
top100_features_subset["isTwice"] <- 0

Twice_features_subset["isTwice"] <- 1
```

After removing the band's songs in the top 100, it combines two data frames into one dataset.

```
top100_features_noTwice <- anti_join(top100_features_subset,
                               Twice_features_subset,
                                     by = "track_id")
comb_data <- rbind(top100_features_noTwice, Twice_features_subset)
```

Change the ‘isTwice’ column into a factor and remove the ‘track_id’ column. Then, the dataset is randomised for the purpose of split the dataset into training and testing set:

```
# Randomise the dataset (shuffle the rows)
comb_data <- comb_data[sample(1:nrow(comb_data)), ]

# Split the dataset into training and testing sets (80% training, 20%
testing)
split_point <- as.integer(nrow(comb_data)*0.8)
training_set <- comb_data[1:split_point, ]
testing_set <- comb_data[(split_point + 1):nrow(comb_data), ]
```

Start training the decision tree model

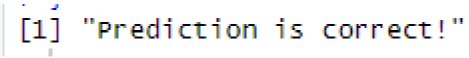
```
dt_model <- train(isTwice~ ., data = training_set, method = "C5.0")
```

Finally, test the model for a particular song from the testing dataset

Predict in top 8, to check prediction:

```
prediction_row <- 8
```

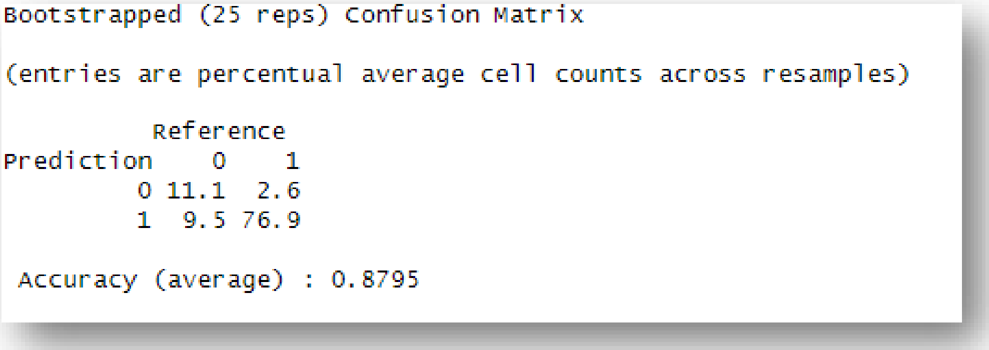
```
if (tibble(predict(dt_model, testing_set[prediction_row, ])) ==
testing_set[prediction_row, 12]){
      print("Prediction is correct!")
} else {
("Prediction is wrong")
}
```



```
[1] "Prediction is correct!"
```

Analyse the model accuracy with a confusion matrix:

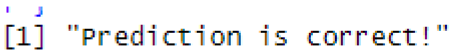
```
confusionMatrix(dt_model, reference = testing_set$isTwice)
```



```
Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 11.1  2.6
         1  9.5 76.9

 Accuracy (average) : 0.8795
```

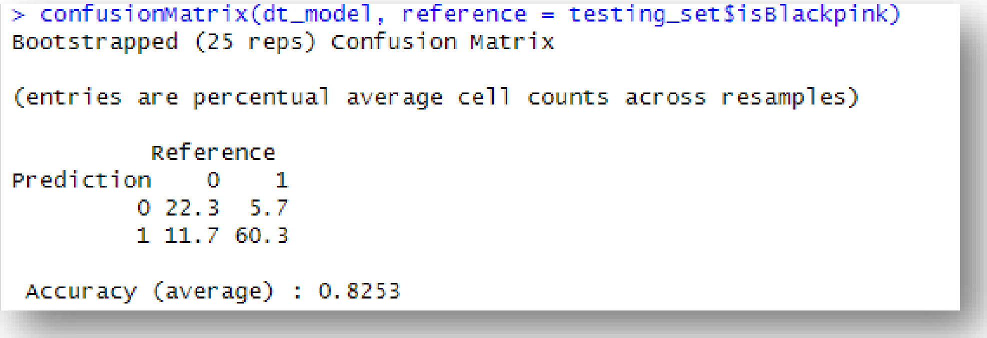
Blackpink

With similar construction of tree model to Twice, I predict their songs at top 1.



```
[1] "Prediction is correct!"
```

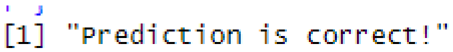
Analyse the model accuracy with a confusion matrix:



```
> confusionMatrix(dt_model, reference = testing_set$isBlackpink)
Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 22.3  5.7
         1 11.7 60.3

 Accuracy (average) : 0.8253
```

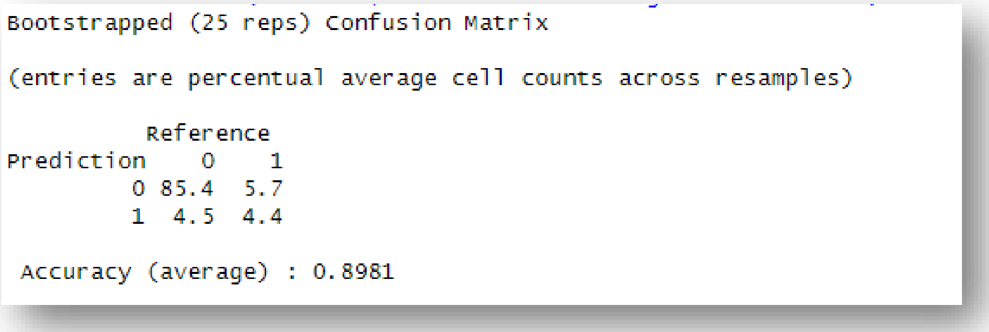
Momoland

With similar construction of tree model to Twice, I predict their songs at top 1.



```
[1] "Prediction is correct!"
```

Analyze the model accuracy with a confusion matrix:



```
Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 85.4  5.7
         1  4.5  4.4

 Accuracy (average) : 0.8981
```

Comparation between the bands

	Twice	Blackpink	Momoland
true positive	11.1	22.3	85.4
false positive	2.6	5.7	5.7
false negative	9.5	11.7	4,5
true negative	76.9	60.3	4.4

average accuracy	0.8795	0.8253	0.8981

In the case of Twice band, the true negative rate ranks at the top, followed by the true positive rate, false negative rate, and positive rate. A similar pattern can be observed for Blackpink as well. However, when it comes to Momoland, the true positive rate stands out, surpassing the rates of the other bands. This indicates that the model correctly identifies a significant number of positive instances as positive, as well as negative instances as negative [1].

Overall, the average accuracy across all three bands exceeds 0.82, which demonstrates a high level of accuracy for the model.

2.8. Topic modelling

Twice

LDA is a good and popular technique in text mining.

First, transform the cleaned tweets into vector corpus, remove stop words from each of the tweets. To optimize the running process, remove objects and run garbage collection. The document-term matrix is used to create the LDA model. I choose 6 topics (k =6):

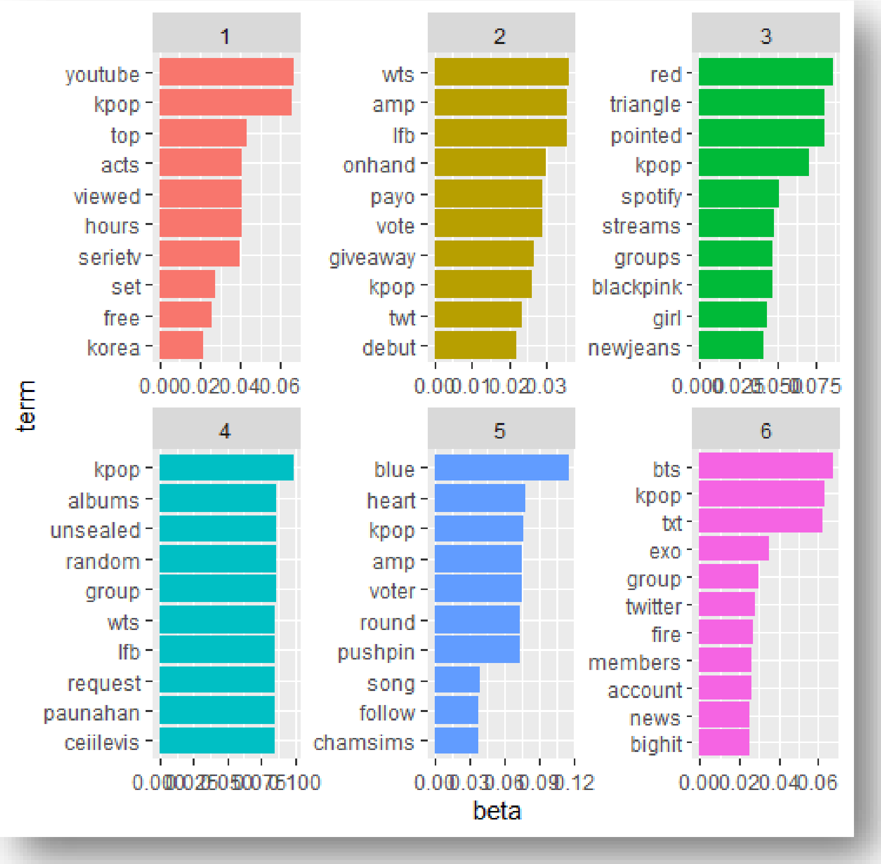
```
lda_model <- LDA(dtm, k = 6)
```

It, next, create the topic porbilities for each words belong to that topic. Use tidy function:

```
tweet_topics <- tidy(lda_model, matrix = "beta")
```

Finally, I selected 10 words which have the highest probabilities for each topic.

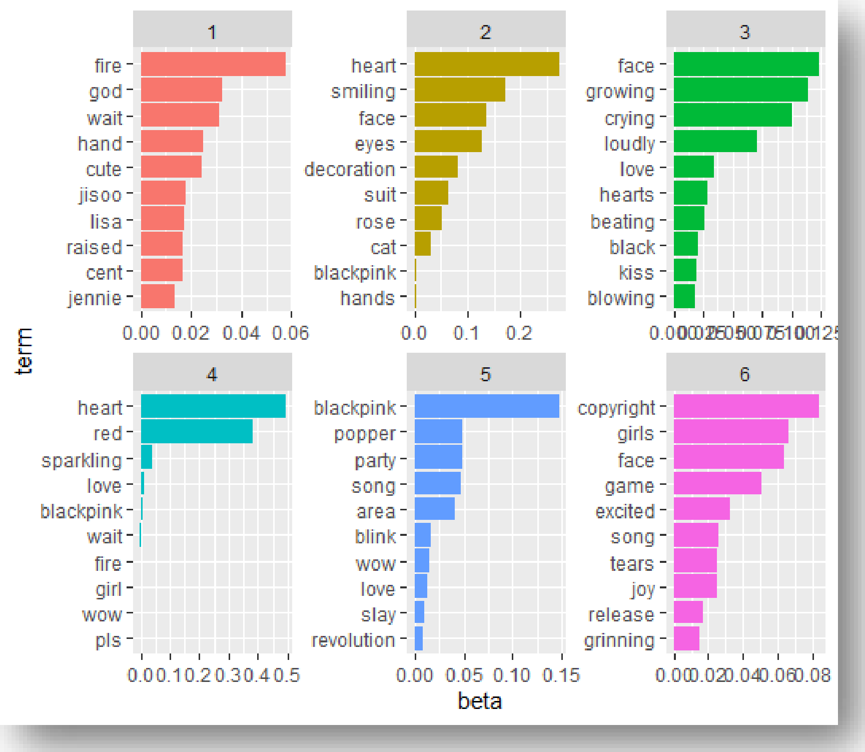
```
top_terms <- tweet_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 10) %>%
  ungroup() %>%
  arrange(topic, -beta)
top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(beta, term, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered()
```



topic	aspects should be focused on
1	youtube, kpop, top, acts, viewed, hours, serietv, set, free, korea
2	wts, amp, lfb, onhand, payo, vote, giveaway, kpop, twt, debut
3	red, triangle, pointed, kpop, sportify, streams, groups, blackpink, girl, newjeans
4	kpop, albums, unsealed, random, group, wts, lfb, request, paunahan, ceilievis
5	blue, heart, kpop, amp, voter, round, pushpin, song, follow, chamsims
6	bts, kpop, txt, exo, group, twitter, fire, members, account, news, bighit

Blackpink

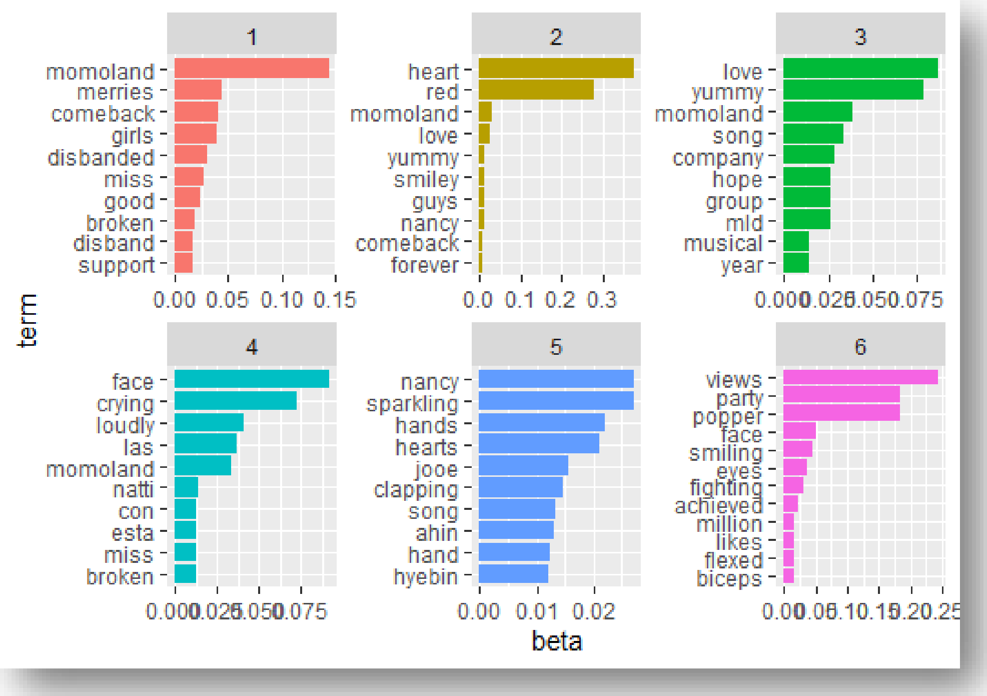
Depend on the LDA model above, the result for Blackpink band is:



topic	aspects should be focused on
1	fire, god, wait, hand, cute, jisoo, lisa, raised, cent, jennie
2	heart, smiling, face, eyes, decoration, suit, rose, cat, blackpink, hands
3	face, growing, crying, loundly, love, hearts, beating, black, kiss, blowing
4	heart, red, sparking, love, blackpink, wait, fire, girl, wow, pls
5	blackpink, popper, patty, song, area, blink, wow, love, stay, revolution
6	coppyright, girls, face, game, exicited, song, tears, joy, relase, grinning

Momoland

Depend on the LDA model above, the result for Momoland band is:



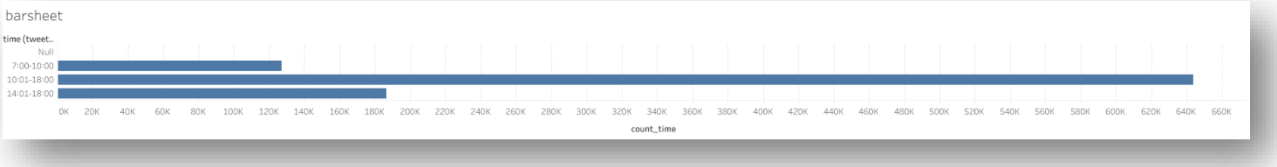
topic	aspects should be focused on
1	momoland, merries, comeback, girls, disbanded, miss, good, broken, disband, support
2	heart, red, momoland, love, yummy, smiley, guys, nancy, comeback, forever
3	love, yummy, momoland, song, company, hope, group, mid, musical, year
4	face, crying, loundly, las, momoland, natti, con, esta, miss, broken
5	nancy, sparking, hands, hearts, jooe, clapping, song, ahin, hand, hyebin
6	views, party, popper, face, smilling, eyes, fighting, achieved million, likes flexed, biceps

Visualization

2.10. Dashboard

From Tweet dataset involving *Twice* collected in Milestone 1, I visualize some insights detail in Tableau Dashboard

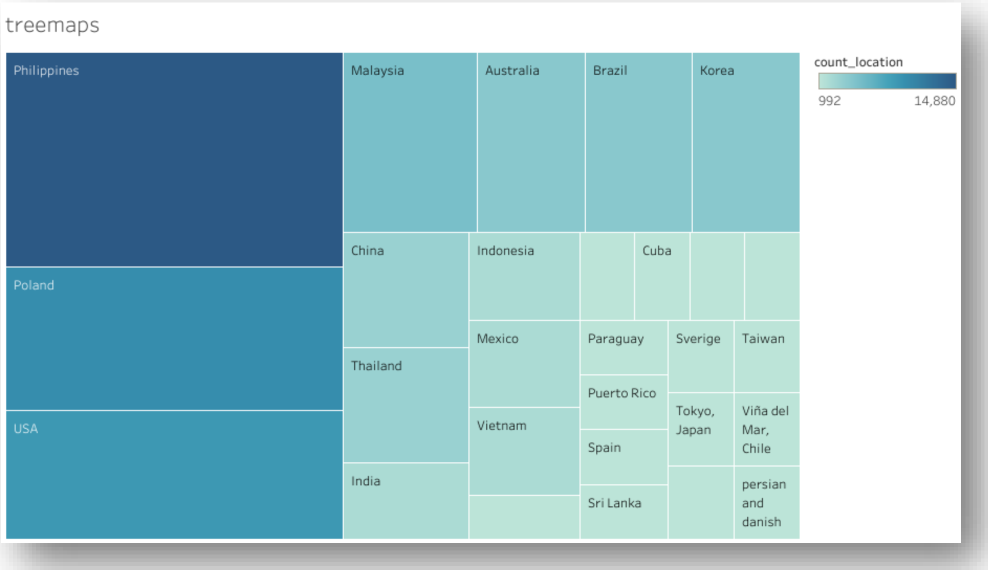
Graph 1: The numbers of tweets published in three period time of day



A bar chart is a commonly employed tool for visualizing and comparing data across various categories or groups. When it comes to counting the number of tweets posted during different periods of the day, a bar chart can effectively present the information in a clear and easily comprehensible manner.

By utilizing the count of tweets posted in different time, managers can gain an understanding of user behavior. This enables insights into the most active periods of Twitter usage, facilitating the identification of peak usage times, user preferences for specific periods, and potential patterns in user behavior.

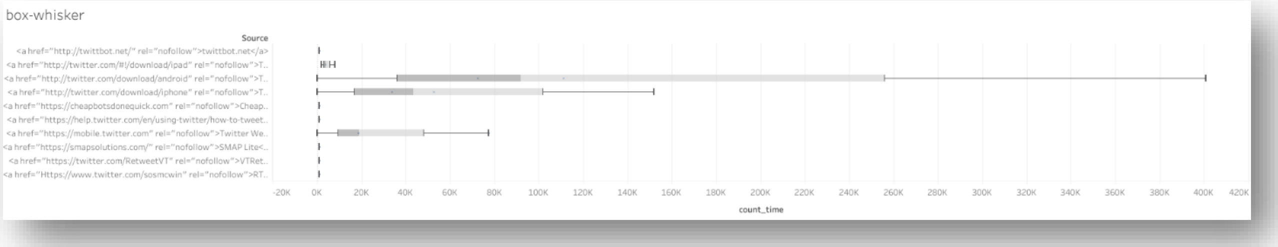
Graph 2: The number of tweets from different areas around the World



Treemap is deemed a useful visualization technique for representing hierarchical data in a rectangular form. When counting how many tweets are posted in different area around the world, the graph can provide a hierarchical representation, proportional display, efficient space utilization, and opportunities for interactive exploration, enabling a comprehensive understanding of the distribution of users across different countries.

By the figures how many tweets are posted in different locations, it will present the geographic distribution of users.

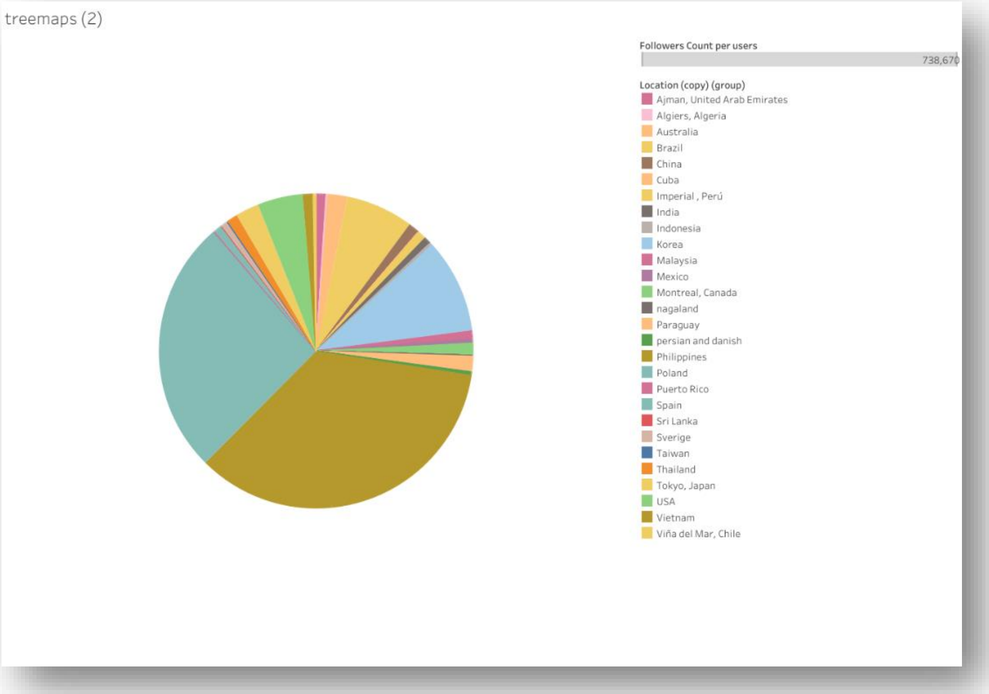
Graph 3: The graph of how many tweets were posted in different period time of day crossing sources



Box-and-whisker plots are effective for displaying and analyzing the distribution and variability of continuous variables. They offer a visual summary of key statistical measures, including the minimum, first quartile, median, third quartile, and maximum values within a dataset.

When examining the number of tweets posted throughout a day across different sources, box-and-whisker plots can help identify peak activity times. These plots enable the determination of periods when user engagement is at its highest. By comparing the data between sources, it becomes possible to evaluate the effectiveness of marketing or content distribution efforts across different platforms. This analysis aids in understanding which sources yield the most favorable outcomes and assists in making informed decisions regarding resource allocation and strategy adjustments.

Graph 4: The pie chart about country-wise distribution of user followers with source details



A pie chart is a popular choice when comparing different components as it provides a clear visual representation of their relative sizes. It allows for a quick assessment of the proportions each category holds within the whole, enabling us to identify which categories have a larger or smaller share.

In the context of determining the number of followers for each user, a pie chart can help showcase the influence and reach across different countries or regions. This can be particularly useful in identifying strong markets and understanding the distribution of followers geographically. By combining this information with the details of the sources, we can determine which sources are prominent in specific areas or regions, providing valuable insights into localized impact and popularity.

Conclusion

The report has indeed proven that social media analytics through given methods in multiple datasets to aid Twice band improve their popularity. This details some key information of Twice and analysis of 2 prevalent features of their songs throughout Spotify data sources. Besides, YouTube dataset gives the numbers like and views of 5-top videos involving into Twice with valuable insights and feedback. Thanks to the data source in Milestone 1, term-document matrix and top 10 terms in Twitter with some comparation with the last information are also figured out. In terms of Social Network Analysis about Twice and 2 more related bands, Centrality analysis presents the degree, betweenness, and closeness analysis, while Community analysis are depicted. Through the use of multiple datasets, Machine Learning models figure out and explain Sentiment analysis, decision tree, and topic modeling about all three bands. Additionally, there are four different charts detailing with description and reasons for inclusion. As a result, Twice can leverage this information to develop their popularity in Twitter, YouTube, and Spotify.

References

1) Callon, M. (2001 - p.62-66). Actor Network Theory. *International Encyclopedia of the Social & Behavioral Sciences*

2) Golbeck, J. (2013- p.25-44). Network Structure and Measures. *Analyzing the Social Web*

3) Golbeck, J. (2015- p.221-235). *Analyzing networks. Introduction to Social Media Investigation*.

4) Spotify. (n.d). *Twice*. https://open.spotify.com/artist/7n2Ycct7Beij7Dj7meI4X0