Improving John Mayer Popularity with Data Analytics

Milestone 3

Group 24
Daniel Spindler - s5238197
Gabriela Almeida Monteiro - s5198626
Pedro Veloso Guimaraes – s2925534

Griffith University

7230ICT Big Data Analytics and Social Media

Lecturer: Dr Sebastian Binnewies

Assessment item: Assignment Milestone 3

Word count: 12,926

Due date: 06/10/2020

MILESTONE 1

Case Study Setting

1.1. Describe the artist/band you are managing. Make sure to reference your sources properly (don't plagiarise). Any referencing format is fine as long as you stay consistent.

Our group is currently managing John Mayer, an American singer who plays pop and rock music. Mayer started to play guitar when he was 13 years old and has been **active in the music industry since 1998**. In 1997 he started to study music at the Berklee College of Music, but abandoned his studies two semesters later with the argument that, instead of studying music, his real passions were singing and song-writing (Wikipedia, 2020). He then moved to Atlanta (Georgia) where he gained popularity by playing in bars, cafes and restaurants in the city.

John Mayer became **well-known not only for being a good compositor, but also for his abilities as a guitarist**. During his career he **had influence from Blues** and attempted to form a blues group. However, he ended up in a solo career. He has released **seven studio albums** (Wikipedia, 2020) **and has a total of 102 songs** (Song List, 2020). **Some of his songs have reached the topmost played music in the United States**. According to Billboard, Mayer "has won seven Grammy's, scored the most No. 1s on Billboard's Top Rock Albums chart of any artist, and notched 20 Hot 100 hits" (Billboard, 2018).

John Mayer **has collaborated with a number of other artists throughout his career**. For instance, he has partnered with Buddy Guy, Eric Clapton, John Scofield, BB King, Kanye West, Katy Perry, Taylor Swift (After Dark, 2020; Wikipedia, 2020). Additionally, he has toured with Maroon 5, Guster, Counting Crows, The Wallflowers, and Teitur (After Dark, 2020). According to Soundigest (2019), one of his best collaborations were "Half of My Heart", with Taylor Swift, "Go!", with Common and Kayne West and "Who You Love", with Katy Perry.

1.2. Describe the purpose of using social media analytics for your case study[1].

- *What is your hypothesis (expectation) about the analysis outcome?*

With insights from social media analytics it will be possible to confirm if certain strategies that our group was considering will be effective to improve John Mayer popularity. Therefore, before doing the social media analysis, our team had a brainstorm session to come up with some general ideas of how to improve his popularity. Some of the ideas were:

- Become more active on social media
- Take part of more marketing and ads
- Have more live shows \ tours or more online shows (streamed event), during the pandemic

[1] The revised answer can be found on page 50.

- Release more frequent albums
- Partner with other artists
- Participate on TV shows

Therefore, **these were our hypothesis (expectation) about the analysis outcome**. However, after proceeding with the analysis, it became abundantly clear that there are also some other pressing issues that were not considered in our brainstorming session and that are vital to increase John Mayer's popularity.

- *How do you want to improve the popularity?*

Although it is our interest to improve John Mayer's popularity, it is important to not lose sight of his profile as a musician. In one of his interviews, he says: "I don't make music for the club, I make music for the omelette on Sunday after the club" (Billboard, 2018). This statement makes it clear that his intention is to **play soft and calm music for a relaxing and cosy atmosphere**.

In our analysis, **topics related to women empowerment and black people** seem to have a high impact on his audience. Therefore, it comes with no surprise that John Mayer's collaboration with Taylor Swift and Kayne West showed to be highly beneficial in his career. For this reason, partnerships in this level should be part of the strategy to improve John Mayer's popularity. Another relevant insight was that the Landrover advertisement presented good results in for his reputation. This can be due to the tendency of his fans to enjoy **sports and freedom**. Finally, our team sees that there is a lot of room for enhancing John Mayer's popularity, specially by looking at figures of similar artists like Ed Sheeran, who has around 50 million followers on Spotify, whereas John Mayer has approximately 5 million. Therefore, **our team wants to improve his popularity by focusing on strategies that will consider the interests mentioned above.**

- *How can social media analytics help you achieve it?*

We will use social media analytics to improve our targeting strategies and we will look for answers to the questions below:

- Which are the brands with which that John Mayer should partner in the short term?
- Which locations is he less popular? After doing a tour in these places, is there a probability to increase his reputation?
- Which artists does he have in common with fans?
- Are there any pressing issues that people don't like about John Mayer?
- To which other genres could he branch out? (songs for games, events, movies?)
- Which kind of direct fan engagement should he promote?
- Who is his audience?

Social media analytics can help answer these questions through textual/sentiment analysis of social media users.

- *What kind of social media data do you want to analyse?*

Our team will focus the analysis on **Twitter and Spotify posts**. However, in future it is also important to consider other social networks such as Youtube, Instagram, Facebook , Amazon, Apple Music, Soundcloud and Deezer. Plus, we need to always be on the lookout for new social media to check on.

Data Selection & Exploration

1.3. Collect data about your artist/band from Twitter. Make sure to choose keywords for data retrieval that are most relevant to your artist/band. However, try not to be too narrow. As a rough guide, you should retrieve at least 1000 tweets. Explain what you have done.

When choosing the best keyword for our data retrieval, there were four options considered: the first was to use the hashtag "#johnmayer", however this option generated only 71 tweets, which is a low number for the purpose of this analysis. The second option was to use the surname "mayer", which, in contrast to the first option, generated a large number of tweets not related to the singer John Mayer. The third option was to use one of his songs title, but this led us to a very skewed data, only related to that particular song. Finally, the last option was to use "John Mayer" as a keyword. This ultimately generated the most relevant tweets, in a significant number (1000) and only related to the singer that is being analysed. A possible reason for that is the public being accustomed to using in their tweets this singer's artistic name, which is John Mayer.

1.4. List the top 5 most influential users for your artist/band. Find out what other interests/characteristics they have besides those related to your artist/band. Do these 5 have something in common? *(=> Lab 2.1)* [1 mark]

An actor graph and network was created and the top five influential users were found. Their network can be seen in figure 1. On table 1 it is summarized the five most influential twitter users for John Mayer, their respective IDs and a list of their interests. It is assumed that their interests can be extracted from the 10 most frequent words in their tweets. In order to find out their interests, 1000 tweets were retrieved from each of them, a semantic network and a semantic graph was created for each of these users and, then, a rank of the top 10 most common terms in their tweets was presented.

Between the five most influential users, it is possible to see two companies: TheAtlantic, which is a newspaper, and Youtube, which is a platform for sharing videos. The other three users are women. It is clear from the tweets that, besides John Mayer, there are some common interests between these users. **Overall, it is evident that women empowerment, (democrat) politics and issues affecting black people are dear to John Mayer's top five influential users.**

Terms like "#rebelgirls" (TheAtlantic), "#justiceforbreonnataylor" (ClaireMPLS) and "#blackgirlsruntennis" (Drinksangrhea) show the tendency to defending the women cause. Additionally, the terms "#belarusprotests", "trump" (TheAtlantic), "readthebills" (ClaireMPLS) and "#iwontvotetrump" (Youtube) indicate that these users have a political

point of view and they are, most likely, democrats. It is also interesting to observe that these users tend to have in their tweets terms related to black people, as for instance, “#justiceforbreonnataylor” (ClaireMPLS), “#blackgirlsruntennis”, “#verzuz”, “#lovecraftcounty” (Drinksangrhea), “dave” + “chapelle”, “color” (Sarah_anne9499) and “#akwanda” (Youtube). Apart from that, some users demonstrate a passion for sports and, also, for tv shows and series. For instance, terms related to sports are “game” (ClaireMPLS), “#nfl” and “#houvskc” (Drinksangrhea). On the other hand, terms related to tv shows and series are “#verzuz” and “#lovecraftcounty (Drinksangrhea).

Table 1: Rank of 10 most frequent terms amongst John Mayer’s top 5 most influential users

TheAtlantic - User ID 35773039

#metoo	#covid19	#belarusprotests	#trump
0.03385644	0.02156436	0.02148619	0.02029324
#mainstreammedia	atlantic	#rebelgirls	#facts
0.01698092	0.01479491	0.01181934	0.01136768
americas	mind		
0.01110622	0.01110622		

ClaireMPLS - User ID 117314237

#readthebills	#justiceforbreonnataylor	game
0.048900727	0.016059869	0.006325111
#blessed	hoping	relief
0.006325111	0.006123981	0.005838839
give	long	havent
0.003678944	0.003678944	0.003678944
assuming		
0.003678944		

Drinksangrhea - User ID 2561500970

#verzuz	#nfl	#savage	#blackgirlsruntennis
0.080033619	0.023154897	0.016619841	0.011238178
#clb	#houvskc	#lovecraftcountry	#lovecraftcounty
0.008547347	0.008547347	0.008547347	0.008547347
#numbers	im		
0.008547347	0.007495873		

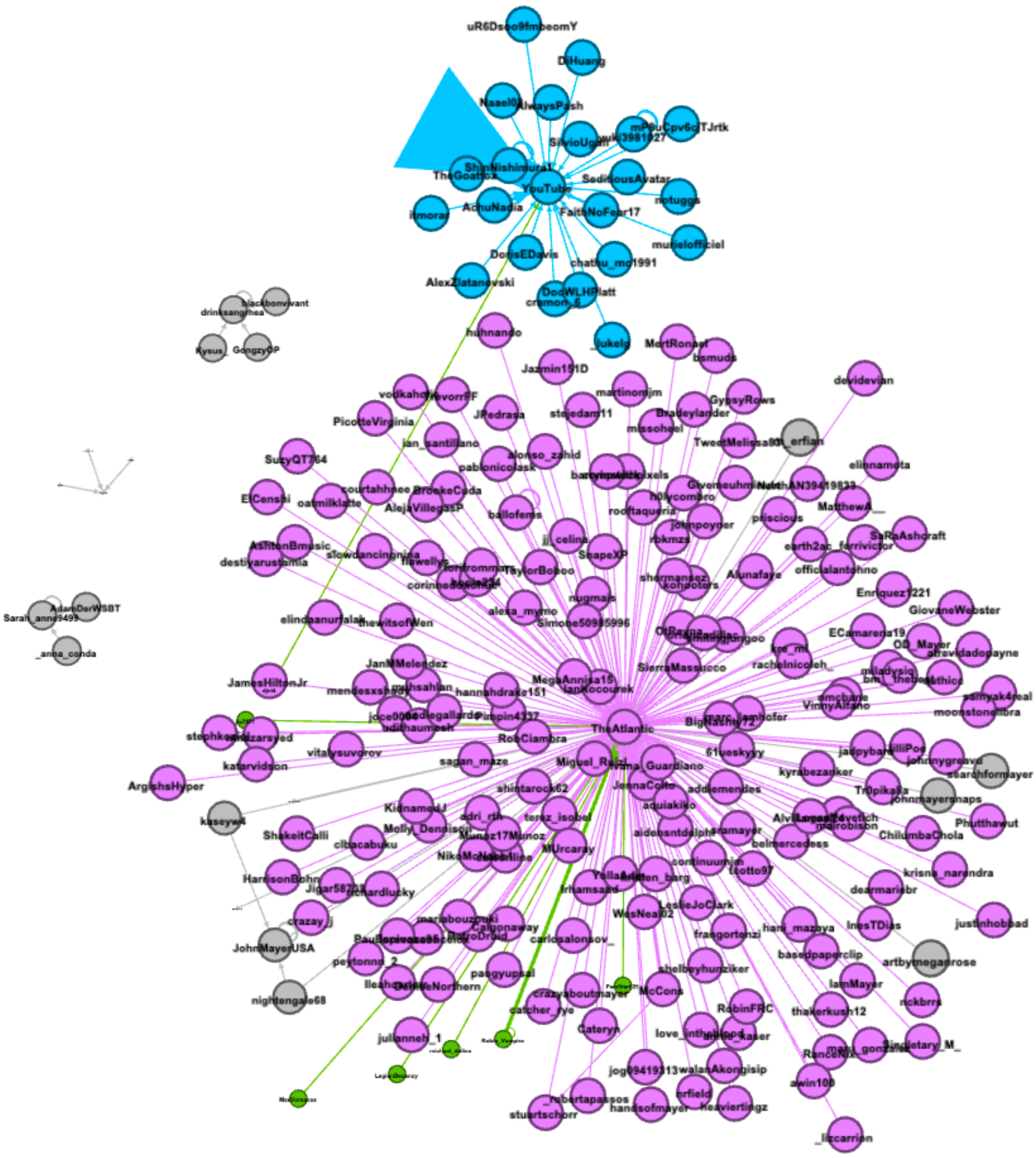
Sarah_anne9499 - User ID 2513836741

#newprofilepic	age	boys	love	perfect
0.27397260	0.27397260	0.04109589	0.04109589	0.04109589
\U0001f62d	ago	chapelle	color	dave
0.04109589	0.04109589	0.04109589	0.04109589	0.04109589

YouTube - User ID 10228272

#bts	#ohmfluke	#akwanda	#godmorningmonday
0.020242690	0.014251141	0.008970722	0.008655528
#exo	#iwontvotetrump	#drippin	#snowman
0.007850339	0.007656506	0.007303050	0.007076177
#astro	video		
0.006314063	0.006134922		

Figure 1: Influential Users Graph



1.5. List the top 10 most important terms that appear together with your keyword(s) related to your artist/band. Explain the results. *(=> Lab 2.1)*
[1 mark]

The top 10 most important terms are:

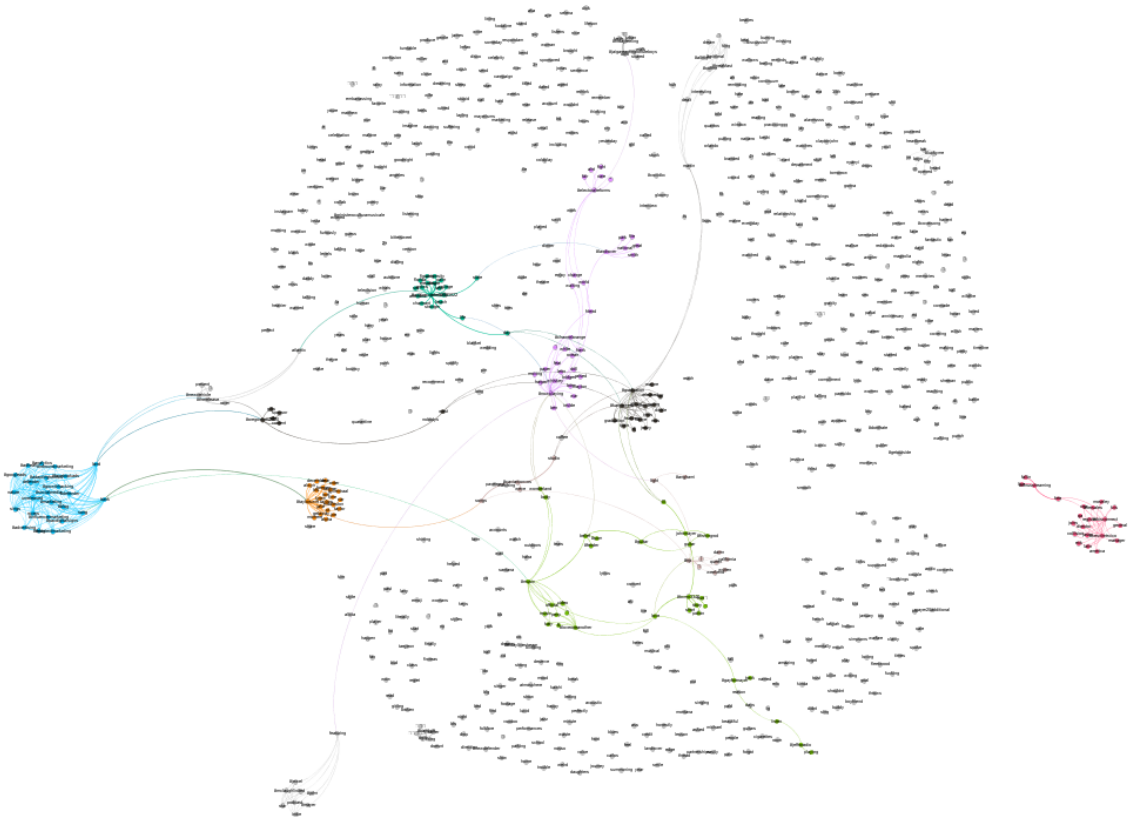
#taylorswift
#nowplaying
#writingcommunity
#graduation
#kanyewest
#np
#cesviméxico
#ibisconnect
#luckyme
#bcmn3340

This was calculated by using a semantic network graph using 25% of the most frequent terms (default is 5%), with 75% of the most frequent hashtags (default is 50%). A page rank algorithm was then used to extract the 10 most important terms.

It is interesting to note two artists that John Mayer collaborates with mentioned in our results, Kayne West and Taylor Swift. Note these two artists are much more popular than Mayer.

A visualization depicts the term centrality (page rank) within our twitter data. We can see that most terms do not have relationships(edges).

Figure 2: Term Centrality



1.6. Calculate how many of your retrieved tweets are retweets. Alternatively, if you filtered out retweets in your query, calculate how many unique user accounts there are in your dataset. What do the results tell you? *(=> Labs 1.1 – 2.1)* [1 mark].

When creating an actor network using the twitter data, the number of retweets is calculated. Within the 1000 retrieved tweets, 356 are retweets. This is a significant portion of our tweets. This tells us that around 35% of the tweet content in our dataset is based on already existing or repropagated tweets – In the form of a retweet. This would prove significant to the communication of the content across the network and the relationships formed between user's and the content.

Figure 3: Number of retweets

```
> twitter_actor_network <- twitter_data %>% Create("actor")
Generating twitter actor network...
-------------------------
collected tweets | 1000
retweets         | 356
quoting others   | 40
mentions         | 83
reply mentions   | 67
replies          | 185
self-loops       | 357
nodes            | 1148
edges            | 1088
-------------------------
```

To confirm the retweet count, a new data frame was created that contains only rows with is retweet == "TRUE". The number of rows is 356, thereby confirming the actor network calculation is correct.

1.7. Use the Spotify API to extract data about your artist/band. For example:

- **How many years have they been active?**
- **How many albums & songs have they published?**
- **With whom have they often collaborated?**
- **What are the prevalent features of their songs (e.g. valence)?**

How does the Spotify data compare to the information you collected from other sources in Step 1.1? *(=> Lab 2.2)*
[2 marks]

To calculate years active, we have taken the most recent(max) album release year and subtracted the oldest album release year to get the number of years active as 18 years (2017 – 1999).

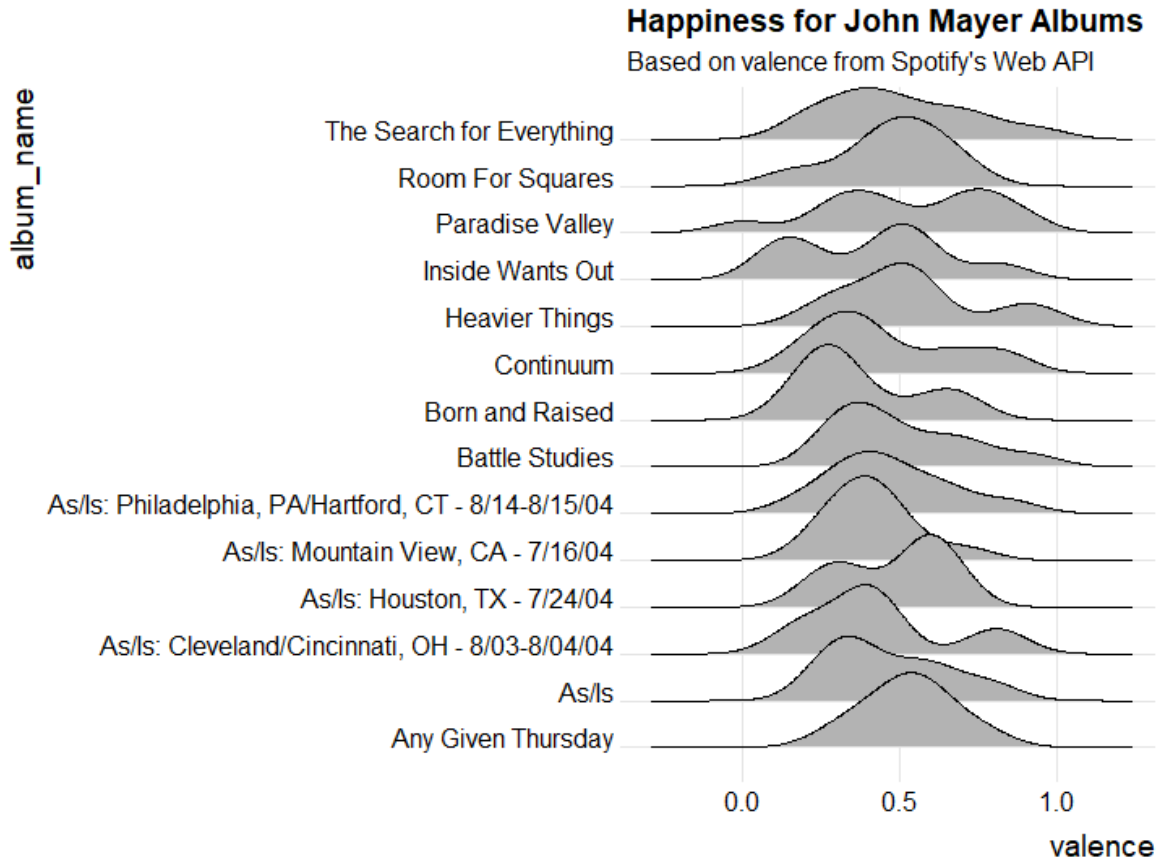
Retrieving John Mayer's albums, we note 16 albums, this includes studio and live albums.

A count of John's songs, using audio_features function, yielded 145 unique song names.

To determine collaboration this has been done in two parts, firstly a search as done for the string "feat" within John Mayer's song list. This presents two John Mayer songs featuring Katy Perry and Frank Ocean. Secondly, a search of all tracks on Spotify for song names with string "feat. John Mayer" in the song name. This produced a unique list of 12 artists, taking the list of collaborative artists to 14.

Plotting John Mayer album valence, we can see that most albums are predominantly close to or just below the 0.5 mid-point. Indicating that most albums are not overly happy and would contain minor elements of negativity or sadness. Owing to John Mayer's blues type roots this is understandable. There are of course exceptions to this, such as Paradise Valley (contains a mix of positive and negative) and some of the live albums (that may be happier due to the live nature of the audio).

Figure 4: Happiness for John Mayer Albums



1.8. Find related artists/bands on Spotify and create a network graph. Did you find any interesting relationships? *(=> Lab 2.2)* [2 marks]

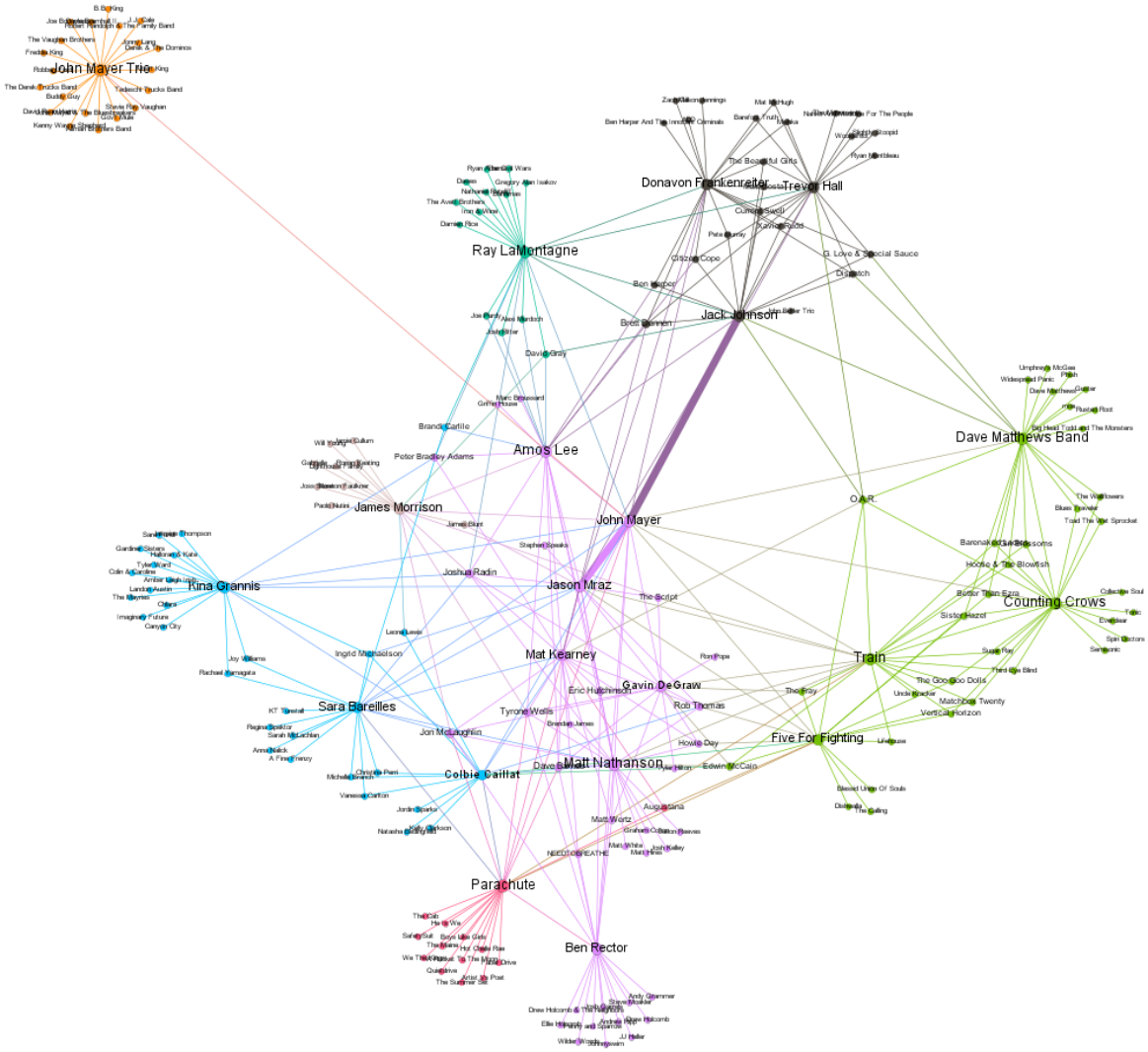
The approach taken was to first find the artists related to John Mayer. For each of these artists we then created an edge between John and that artist. For each of John's related artists, we then retrieved the list of that artist's related artist and created an edge for each of those relationships.

It is noted that only two of John's related artists have John as a related artist.

When creating the network graph in Gephi, for parallel edges, a sum merge strategy was used. The reason is to highlight nodes that have multiple relationships – in this case it will be John Mayer and his related artists. John Mayer to all his related artists is 1 edge and any of his related artists that have John Mayer listed as a related artist will have a sum of 2 edges. These have been highlighted in the graph, with thicker edges, as special relationships. John Mayer has a bidirectional relationship with Jack Johnson and Jason Mraz.

The artist John Mayer Trio is isolated – somewhat of a different genre, with more of a blues feel as seen by related artists like BB King and Stevie Ray Vaughn. This highlights that John Mayer is mainly connected to pop (his main relationships) but also to blues (John Mayer Trio).

Figure 5: Network Graph of John Mayer and Related Artists



Text Pre-Processing

1.9. Perform text pre-processing and create a Term-Document Matrix for your Twitter data. What are the 10 terms occurring with the highest frequency? How are they different to your answer for 1.5) above? *(=> Lab 2.2)* [1 mark]

A Term_document Matrix(dtm_df) has been calculated, and the top 10 terms are:

john
mayer
skeptic
time
johnmay
tongu
stick
landroverusa
remind
moment

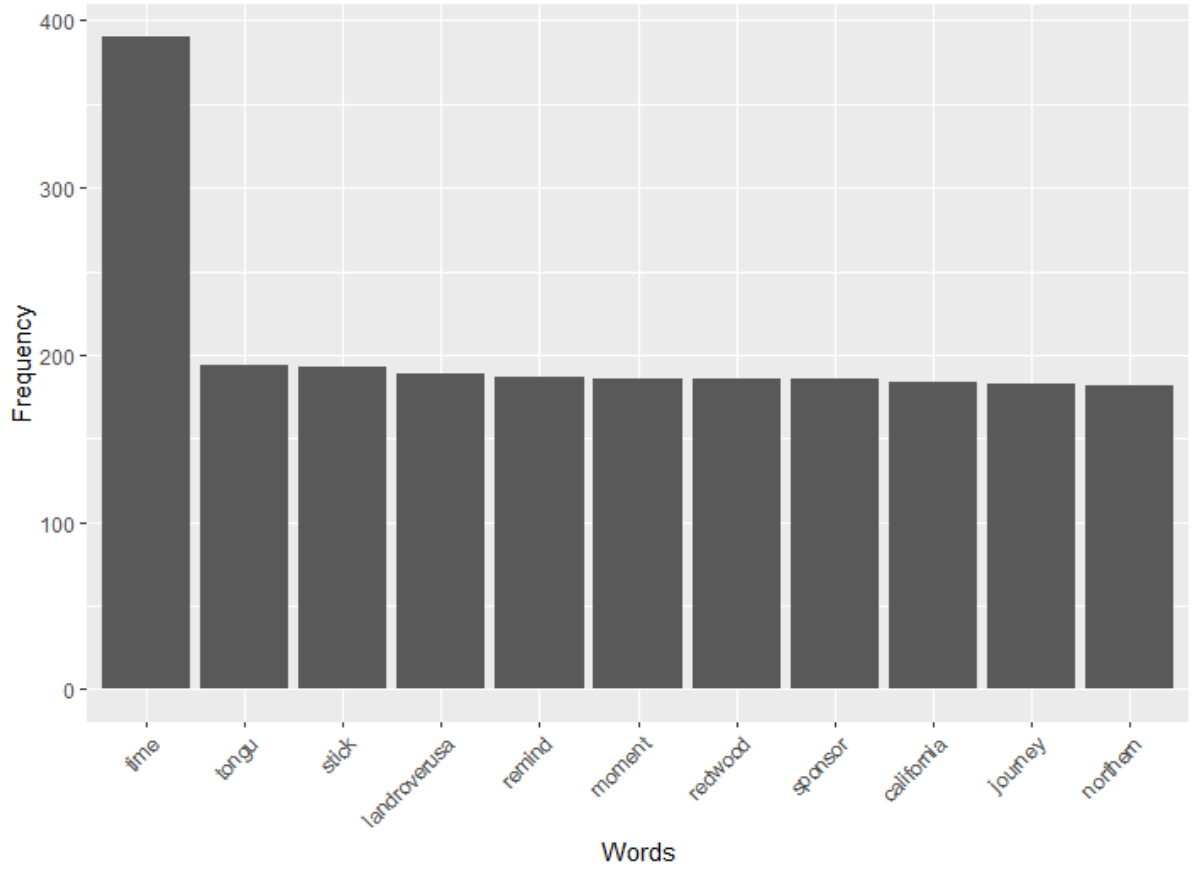
Because both “john” and “mayer” are part of our search query, they do not add any value and therefore they have been removed. We also noted that the term “skeptic” is a result of the cleaning process identifying twitter short URLs and thus we have removed it. Our new top 10 terms are:

time
johnmay
tongu
stick
landroverusa
remind
moment
redwood
sponsor
california

The top 10 highest terms as calculated by TFIDF are very different from our top 10 most important terms (as per 1.5) we found from semantic analysis with page rank. This is expected as TFIDF is focusing purely on the language term analysis, whereas our semantic with page rank analysis is looking at node centrality.

From a plot of the word frequency we can see that terms 2 through 10 have almost the same frequency. Note the term “landroverusa” that relates to advertising in which John Mayer is involved.

Figure 6: Word Frequency



MILESTONE 2

Data Selection & Exploration

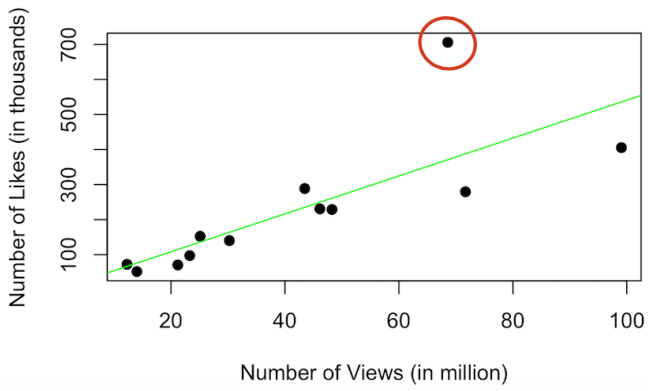
2.1. Retrieve data relevant to your artist/band from YouTube. Which videos have the highest number of views and likes? Do you see a correlation between views and likes? (Your dataset may contain hundreds of videos, so it's OK if you choose only a subset of those to get their statistics, in order to avoid hitting the rate-limit. However, you should get statistics for at least 5 videos.)

Data was retrieved from Youtube using "John Mayer" as a keyword. We used a for loop to collect statistics about number of likes and number of views for 12 videos. During the process, we got an error message indicating incorrect ID for video "uvCvocKTQ-Y" (which corresponded to the line number 7 in the "video_search" dataset). Because it was not possible to collect statistics for this video, the loop was interrupted, and no statistics were collected for none of the videos. In order to solve this problem, we deleted this video from the video_search dataset and ran the for loop again.

The statistics were saved in the new dataframe called videos_stats and the order() function was used to find the videos with the highest number of views and the videos with the highest number of likes. As a result, from the subset of 12 videos, "Free Fallin" (ID 20Ov0cDPZy8) showed the highest number of views, with approximately 99 million views, whereas "New Light" (ID mQ055hHdxbE) had the highest number of likes, with approximately 700 thousand likes.

It is clear that there is a direct relationship between number of views and number of likes, as seen in the plot below. However, an outlier (shown in the red circle) called our attention. It corresponds to the video "New Light", which despite having a high number of likes, was not as much viewed as the first and the second most viewed videos. One possible explanation for that is because this video (which was uploaded in 2018) makes people remember clothes and living style of the current quarantine period. Therefore, from the recent collected data, this video is seen a funny and prophetic.

Figure 1: Video Likes vs. Video Views



Social Network Analysis

2.2. For your Twitter actor graph, Twitter semantic graph, and YouTube actor graph, explain each of its properties and attributes.

In order to exam the graphs proposed, we reuse graphs already presented in the past. These graphs containing data related to the singer John Mayer presented the following characteristics upon utilization of the R's library function summary. This function gave much insight to graphs containing Twitter and YouTube data. The graphs are characteristics are presented in a peculiar notation (labels for you reference can be been at the end of the 2.2 question). The first part of the summary describes the graph by CAPITALIZED LETTERS. Each of these letters mean something as per labels at the end of this question's response. One can note, in the examples for each graph mentioned below, that the first column of each row relates to the ID of the source and the target node for each edge.

Twitter actor graph

Description:

IGRAPH 68cade5 *DN* -- 1148 1088 –
D= Directed Graph, *N*=Named Graph

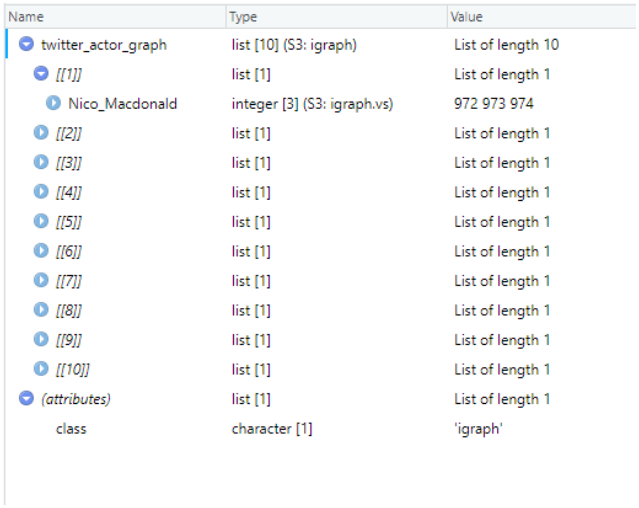
Attributes:

- type *(**g**/c)* - graph-level character attribute
- name *(**v**/c)* - vertex-level character attribute
- screen_name *(**v**/c)* - vertex-level character attribute
- label *(**v**/c)* - vertex-level character attribute
- edge_type *(**e**/c)* - edge-level character attribute
- timestamp *(**e**/c)* - edge-level character attribute
- status_id *(**e**/c)* - edge-level character attribute

Example: *[1] PopeAwesomeXIII->The_Albinoshrek carlosalonsov_ ->TheAtlantic scottybbn ->astroboi____i*

You may also find as per picture below that certain attributes are present such as the "name" which is of type class, "type" which is of type "char", as well as "value" which is of type igraph.

Figure 2: Twitter Actor Graph - View(graph_var)



Name	Type	Value
twitter_actor_graph	list [10] (S3: igraph)	List of length 10
[[1]]	list [1]	List of length 1
Nico_Macdonald	integer [3] (S3: igraph.vs)	972 973 974
[[2]]	list [1]	List of length 1
[[3]]	list [1]	List of length 1
[[4]]	list [1]	List of length 1
[[5]]	list [1]	List of length 1
[[6]]	list [1]	List of length 1
[[7]]	list [1]	List of length 1
[[8]]	list [1]	List of length 1
[[9]]	list [1]	List of length 1
[[10]]	list [1]	List of length 1
(attributes)	list [1]	List of length 1
class	character [1]	'igraph'

Twitter semantic graph

Description:

IGRAPH 7aaac73 *UNWB* 880 424 –

U= Undirected Graph, *N*=Named Graph, *W*= Weighted Graph, and *B*= Bipartite

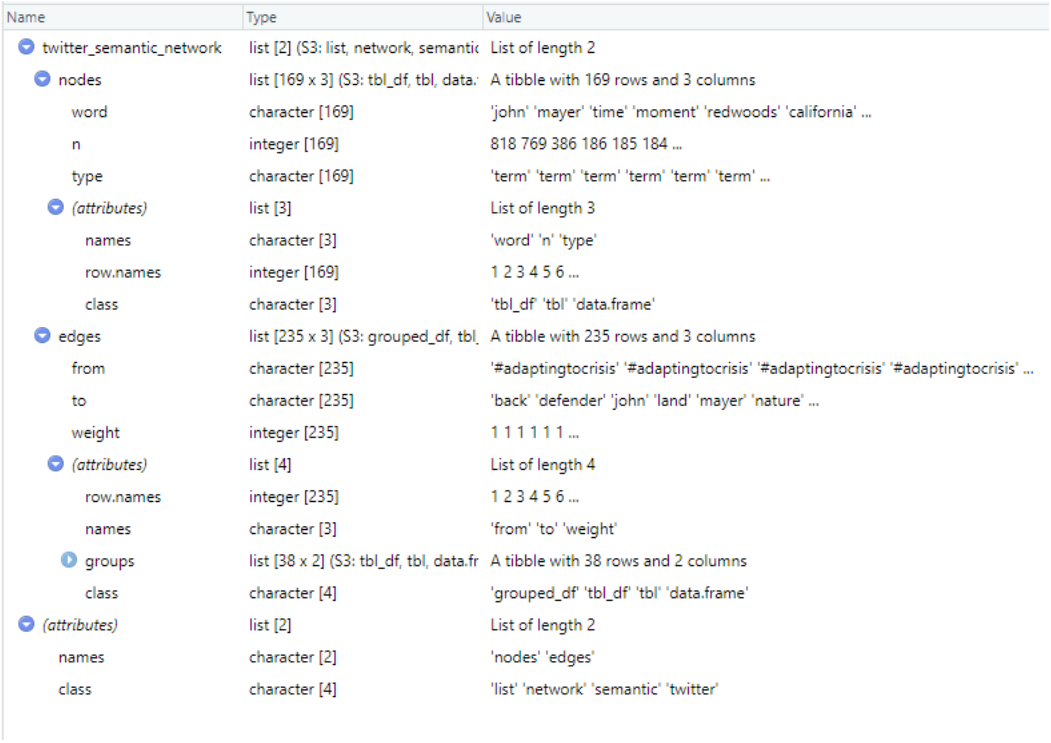
Attributes:

- type ***(g/c)*** - graph-level character attribute
- name ***(v/c)*** - vertex-level character attribute
- n ***(v/n)*** - vertex-level numeric attribute
- type ***(v/c)*** - vertex-level character attribute
- label ***(v/c)*** - vertex-level character attribute
- weight ***(e/n)*** - edge-level numeric attribute

Example: *[1] back --#adaptingtocrisis defender --#adaptingtocrisis land --#adaptingtocrisis nature --#adaptingtocrisis*

You may also find as per picture below that certain attributes are present such as the “names” and “class” which respectively of types “char”, as well as “nodes” for names and list for class

Figure 3: Twitter Semantic Network - View(graph_var)



Name	Type	Value
twitter_semantic_network	list [2] (S3: list, network, semantic	List of length 2
nodes	list [169 x 3] (S3: tbl_df, tbl, data.	A tibble with 169 rows and 3 columns
word	character [169]	'john' 'mayer' 'time' 'moment' 'redwoods' 'california' ...
n	integer [169]	818 769 386 186 185 184 ...
type	character [169]	'term' 'term' 'term' 'term' 'term' 'term' ...
(attributes)	list [3]	List of length 3
names	character [3]	'word' 'n' 'type'
row.names	integer [169]	1 2 3 4 5 6 ...
class	character [3]	'tbl_df' 'tbl' 'data.frame'
edges	list [235 x 3] (S3: grouped_df, tbl_	A tibble with 235 rows and 3 columns
from	character [235]	'#adaptingtocrisis' '#adaptingtocrisis' '#adaptingtocrisis' '#adaptingtocrisis' ...
to	character [235]	'back' 'defender' 'john' 'land' 'mayer' 'nature' ...
weight	integer [235]	1 1 1 1 1 1 ...
(attributes)	list [4]	List of length 4
row.names	integer [235]	1 2 3 4 5 6 ...
names	character [3]	'from' 'to' 'weight'
groups	list [38 x 2] (S3: tbl_df, tbl, data.fr	A tibble with 38 rows and 2 columns
class	character [4]	'grouped_df' 'tbl_df' 'tbl' 'data.frame'
(attributes)	list [2]	List of length 2
names	character [2]	'nodes' 'edges'
class	character [4]	'list' 'network' 'semantic' 'twitter'

YouTube actor graph

Description:

IGRAPH d851dd8 **DN** -- 1618 1848 –

D= Directed Graph, **N**=Named Graph

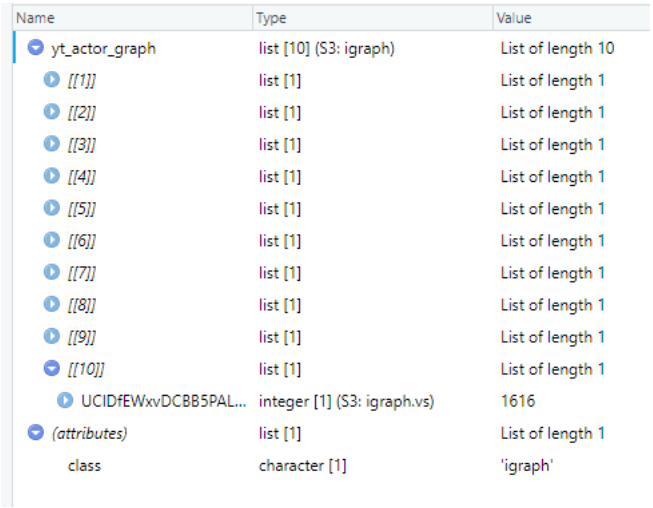
Attributes:

- type *(**g**/c)* - graph-level character attribute
- name *(**v**/c)* - vertex-level character attribute
- screen_name *(**v**/c)* - vertex-level character attribute
- node_type *(**v**/c)* - vertex-level character attribute
- label *(**v**/c)* - vertex-level character attribute
- video_id *(**e**/c)* - edge-level character attribute
- comment_id *(**e**/c)* - edge-level character attribute
- edge_type *(**e**/c)* - edge-level character attribute

Example: *[1] UCSRz9quTl2vfRuLMQFr6tVQ->VIDEOID:20Ov0cDPZy8 UCDz3vtorp77XHhWB60tLC9A->VIDEOID:20Ov0cDPZy8 Attributes: name-=class, type=character (like integer), value=igraph*

You may also find as per picture below that certain attributes are present such as the “name” which is of type class, “type” which is of type “char”, as well as “value” which is of type igraph.

Figure 4: YouTube Actor Graph - View(graph_var)



Name	Type	Value
yt_actor_graph	list [10] (S3: igraph)	List of length 10
[[1]]	list [1]	List of length 1
[[2]]	list [1]	List of length 1
[[3]]	list [1]	List of length 1
[[4]]	list [1]	List of length 1
[[5]]	list [1]	List of length 1
[[6]]	list [1]	List of length 1
[[7]]	list [1]	List of length 1
[[8]]	list [1]	List of length 1
[[9]]	list [1]	List of length 1
[[10]]	list [1]	List of length 1
UCIDfEWxvDCBB5PAL...	integer [1] (S3: igraph.vs)	1616
(attributes)	list [1]	List of length 1
class	character [1]	'igraph'

Properties and Attributes labels. The description of the igraph objects starts with up to four CAPITALIZED letters.

Description labels:

- *D or U, for a directed or undirected graph*
- *N for a named graph (where nodes have a name attribute)*
- *W for a weighted graph (where edges have a weight attribute)*
- *B for a bipartite (two-mode) graph (where nodes have a type attribute*

Attribute Labels:

- *(**g**/c) - **graph**-level character attribute*
- *(**v**/c) - **vertex**-level character attribute*
- *(**e**/n) - **edge**-level numeric attribute*
- *(**e**/c) - **edge**-level character attribute*
- *(**g**/c) - **graph**-level character attribute*

2.3. In your Twitter actor graph, how many of the edges are 'mentions', and how many are 'replies'? What do the results tell you?

Mentions and Replies have different meanings in social networks. According to Pew Research Center (2014), "a 'mentions' edge is created when one user creates a tweet that contains the name of another user" (p. 6) . On the other hand, "a 'reply' relationship is a special form of 'mention' that occurs when the user's name is at the very start of a tweet" (Pew Research Center, 2014, p. 6). Therefore, replies are an indirect type of mention.

Direct mentions are more powerful to companies and artists than when users reply to other users' tweet with the artist's name. When users mention an artist publicly, it helps to endorse the artist and future twitter recommendations will consider showing more of this artist's tweets for the user.

Our actor network graph generated from the twitter dataset from Milestone 1 has a total of 1088 edges, of which 185 are replies and 83 are mentions. In order to calculate how many of the edges are mentions and how many are replies, we used two different strategies to check our answers:

a. *Strategy 1: Through coding in* R. The function *as_edgelist()* was used to return a list of edges from our twitter actor network graph. Then, a *which*() function was used to filter the results only for "reply" and then, for "mention" from the column edge_type. The code can be verified in the figure below.

Figure 5: Code to return the number of mentions and replies.

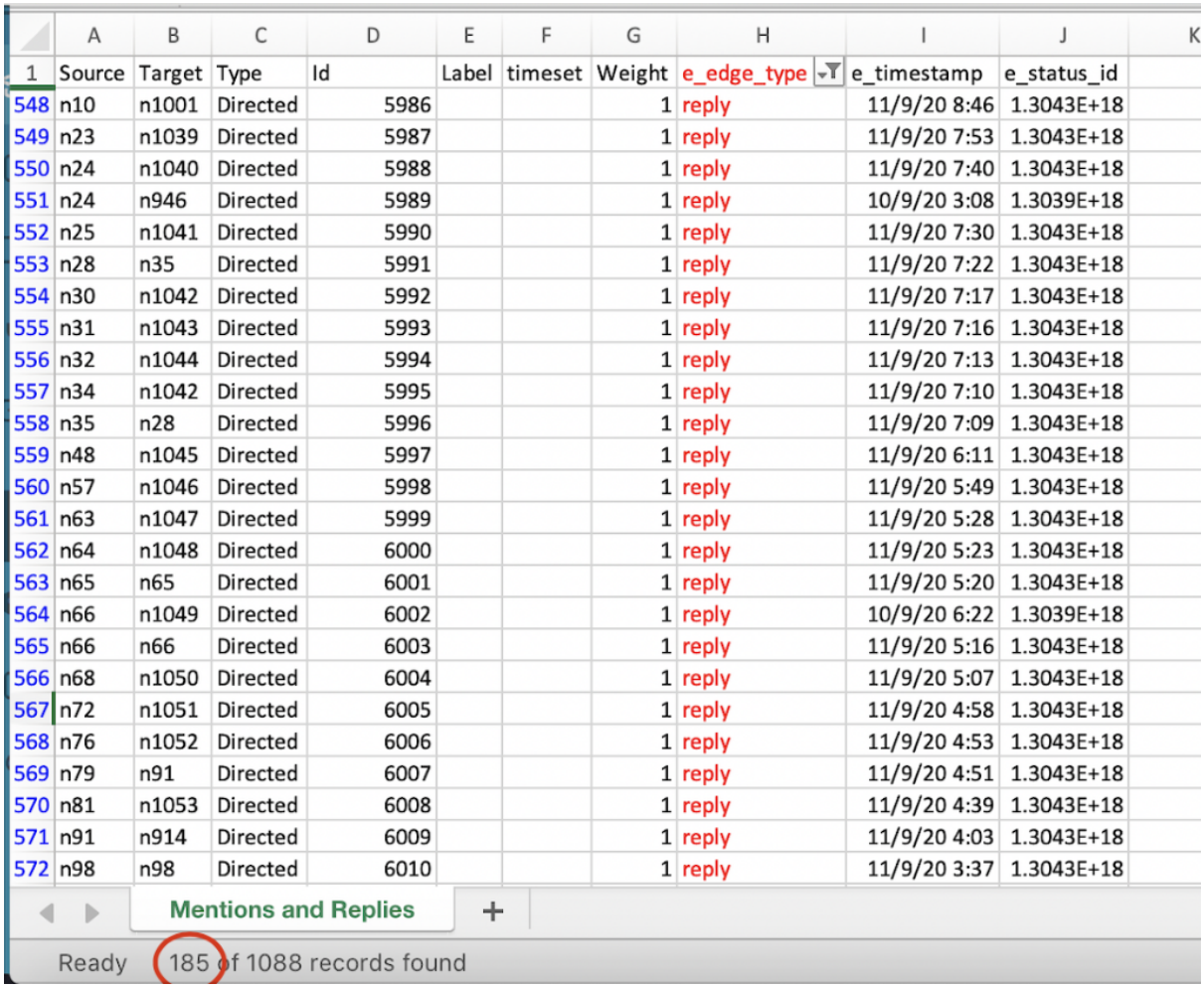
```
> # Use the as_edgelist function to return a list of edges in a graph. Then,
> # find the number of edges with mentions and replies.
>
> #as_edgelist(twitter_actor_graph, names = TRUE)
> mentions <- which(E(twitter_actor_graph)$edge_type=="mention")
> length(mentions)
[1] 83
>
> # Do the same process to find the number of replies.
>
> #as_edgelist(twitter_actor_graph, names = TRUE)
> replies <- which(E(twitter_actor_graph)$edge_type=="reply")
> length(replies)
[1] 185
```

b. *Strategy 2: Through Gephi and Excel*. First, we opened the actor network graph in Gephi. Then, we used Data Laboratory view, chose the "edge" option on the left-hand side, and exported the table to excel. There, it was possible to filter how many edges were replies and how many were mentions, using the "e_edge_type" column. A screenshot of the excel table can be seen below, with the counts for replies (Figure 6) and mentions (Figure 7).

Our result points that people tend to use more replies than mentions when it comes to John Mayer. The replies in this case happen when users reply to other users who mentioned John Mayer in their tweet. The mentions happen when users directly mention John Mayer in their tweet. Most of the mentions that were observed were associated with relevant actors in the network. For instance, 49 out of 83 mentions were done pointing at Youtube (ID n974) and John Mayer (ID n975). On the contrary, replies, which are more common in our dataset, are less concentrated in the big nodes and relates more to ordinary users. So, it is possible to conclude that users have a more passive and reactive attitude towards John Mayer (indicated by the high number of replies), instead of an active and enthusiastic approach (indicated by low mentions).

Mentions are activities which are difficult to control. However, direct replies to John Mayer's tweets (which is a type of indirect mention) can be incremented so as to improve the artist's visibility in the network. This would require the John Mayer posting more content and increasing the level of engagement with the users so as to incentivize more replies, which would ultimately improve John Mayer's popularity.

Figure 6: Number of Replies in John Mayer's Twitter Network Graph

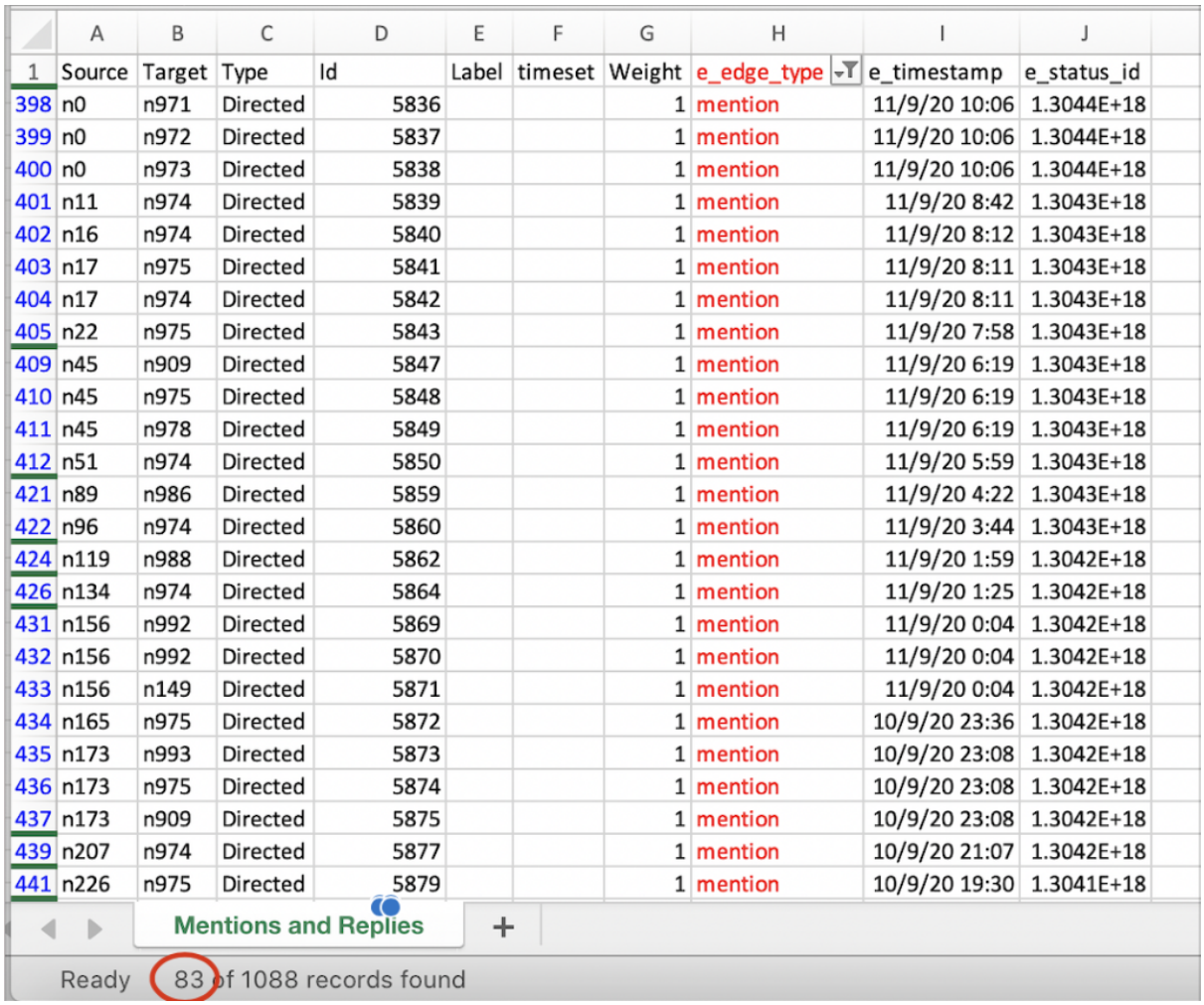


	A	B	C	D	E	F	G	H	I	J	K
1	Source	Target	Type	Id	Label	timeset	Weight	e_edge_type	e_timestamp	e_status_id	
548	n10	n1001	Directed	5986			1	reply	11/9/20 8:46	1.3043E+18	
549	n23	n1039	Directed	5987			1	reply	11/9/20 7:53	1.3043E+18	
550	n24	n1040	Directed	5988			1	reply	11/9/20 7:40	1.3043E+18	
551	n24	n946	Directed	5989			1	reply	10/9/20 3:08	1.3039E+18	
552	n25	n1041	Directed	5990			1	reply	11/9/20 7:30	1.3043E+18	
553	n28	n35	Directed	5991			1	reply	11/9/20 7:22	1.3043E+18	
554	n30	n1042	Directed	5992			1	reply	11/9/20 7:17	1.3043E+18	
555	n31	n1043	Directed	5993			1	reply	11/9/20 7:16	1.3043E+18	
556	n32	n1044	Directed	5994			1	reply	11/9/20 7:13	1.3043E+18	
557	n34	n1042	Directed	5995			1	reply	11/9/20 7:10	1.3043E+18	
558	n35	n28	Directed	5996			1	reply	11/9/20 7:09	1.3043E+18	
559	n48	n1045	Directed	5997			1	reply	11/9/20 6:11	1.3043E+18	
560	n57	n1046	Directed	5998			1	reply	11/9/20 5:49	1.3043E+18	
561	n63	n1047	Directed	5999			1	reply	11/9/20 5:28	1.3043E+18	
562	n64	n1048	Directed	6000			1	reply	11/9/20 5:23	1.3043E+18	
563	n65	n65	Directed	6001			1	reply	11/9/20 5:20	1.3043E+18	
564	n66	n1049	Directed	6002			1	reply	10/9/20 6:22	1.3039E+18	
565	n66	n66	Directed	6003			1	reply	11/9/20 5:16	1.3043E+18	
566	n68	n1050	Directed	6004			1	reply	11/9/20 5:07	1.3043E+18	
567	n72	n1051	Directed	6005			1	reply	11/9/20 4:58	1.3043E+18	
568	n76	n1052	Directed	6006			1	reply	11/9/20 4:53	1.3043E+18	
569	n79	n91	Directed	6007			1	reply	11/9/20 4:51	1.3043E+18	
570	n81	n1053	Directed	6008			1	reply	11/9/20 4:39	1.3043E+18	
571	n91	n914	Directed	6009			1	reply	11/9/20 4:03	1.3043E+18	
572	n98	n98	Directed	6010			1	reply	11/9/20 3:37	1.3043E+18	

Mentions and Replies

Ready 185 of 1088 records found

Figure 7: Number Mentions in John Mayer's Twitter Network Graph



	A	B	C	D	E	F	G	H	I	J
1	Source	Target	Type	Id	Label	timeset	Weight	e_edge_type	e_timestamp	e_status_id
398	n0	n971	Directed	5836			1	mention	11/9/20 10:06	1.3044E+18
399	n0	n972	Directed	5837			1	mention	11/9/20 10:06	1.3044E+18
400	n0	n973	Directed	5838			1	mention	11/9/20 10:06	1.3044E+18
401	n11	n974	Directed	5839			1	mention	11/9/20 8:42	1.3043E+18
402	n16	n974	Directed	5840			1	mention	11/9/20 8:12	1.3043E+18
403	n17	n975	Directed	5841			1	mention	11/9/20 8:11	1.3043E+18
404	n17	n974	Directed	5842			1	mention	11/9/20 8:11	1.3043E+18
405	n22	n975	Directed	5843			1	mention	11/9/20 7:58	1.3043E+18
409	n45	n909	Directed	5847			1	mention	11/9/20 6:19	1.3043E+18
410	n45	n975	Directed	5848			1	mention	11/9/20 6:19	1.3043E+18
411	n45	n978	Directed	5849			1	mention	11/9/20 6:19	1.3043E+18
412	n51	n974	Directed	5850			1	mention	11/9/20 5:59	1.3043E+18
421	n89	n986	Directed	5859			1	mention	11/9/20 4:22	1.3043E+18
422	n96	n974	Directed	5860			1	mention	11/9/20 3:44	1.3043E+18
424	n119	n988	Directed	5862			1	mention	11/9/20 1:59	1.3042E+18
426	n134	n974	Directed	5864			1	mention	11/9/20 1:25	1.3042E+18
431	n156	n992	Directed	5869			1	mention	11/9/20 0:04	1.3042E+18
432	n156	n992	Directed	5870			1	mention	11/9/20 0:04	1.3042E+18
433	n156	n149	Directed	5871			1	mention	11/9/20 0:04	1.3042E+18
434	n165	n975	Directed	5872			1	mention	10/9/20 23:36	1.3042E+18
435	n173	n993	Directed	5873			1	mention	10/9/20 23:08	1.3042E+18
436	n173	n975	Directed	5874			1	mention	10/9/20 23:08	1.3042E+18
437	n173	n909	Directed	5875			1	mention	10/9/20 23:08	1.3042E+18
439	n207	n974	Directed	5877			1	mention	10/9/20 21:07	1.3042E+18
441	n226	n975	Directed	5879			1	mention	10/9/20 19:30	1.3041E+18

Mentions and Replies

Ready 83 of 1088 records found

2.4. Perform centrality analysis by detecting degree centrality, betweenness centrality, and closeness centrality. Explain how relevant the results are to your artist/band. What are the actual degree, betweenness, and closeness centrality scores for your artist/band node in the network? Compare these scores to the scores for related artists.

Out of 1000 collected tweets, 356 were retweets, 40 were quoting others, 83 were mentions, 67 were reply mentions, 185 were actual replies, all of this in with 357 self-loops and 1148 nodes as well as 1088 edges.

While using the data above which as related to John Mayer a two-mode network was created (named "*twomode*") as well as its graph ("*TwitterTwomode.graphml*"). The graph shows 766 of length. The first few rows are shown as per images below.

```
> length(V(twomode_graph))
[1] 766
> V(twomode_graph)$name
  [1] "@johnmayers"          "@landroverusa"        "@guitarworld"         "@youtube"             "#fridaylivestream"
  [6] "#sb19voicejustin"     "@mtv"                 "@sb19official"        "#writingcommunity"    "@johnmayer"
 [11] "@901jazz"             "@leehesng"            "@carskelly"           "@dbacks"              "@mlb"
 [16] "@atlanticrethink"     "@landrovers"          "#gayformayer"         "@lovingtay89"         "@secretmoments13"
 [21] "#newvehicle"          "#therelease"          "@declanmckenna"       "@edwardsluke04"       "@alicewegmann"
 [26] "@imbabi13"            "@khanak08"            "@velasquezjoans"      "@persecutedsgin"      "#oregonfires2020"
 [31] "#rrbc"                "#fridayfeeling"       "@johnmayer❤️"         "@keithurban"          "@hozier"
 [36] "@kingskworld"         "@rapcaviar"           "@babyyymika"          "@landrover"           "@theatlantic"
 [41] "@musicfess"           "@kathnielkhn"         "@araeviper"           "@cayennecaye"         "@jimbryson"
 [46] "@joedram"             "@mataadrian19"        "#getoutside"          "#newdefender"         "@julian"
 [51] "@bravotv"             "@pettygirleri"        "@triastik23"          "@perrytarg"           "@kenshafpf"
 [56] "@subtanyarl"          "@tyronelikeshh"       "#caviarconvos"        "@pennysoulmates"      "@jadpybara"
 [61] "@magicalgrocery"      "@danceswithtamis"     "#analytics"           "#facebookads"         "#googleads"
 [66] "#nowplaying"          "@tidal"               "@adweek"              "@seveninthetrees"     "@kolgorengaja"
 [71] "@rebelrancher"        "@islapthebasss"       "#bandgift"            "#gift"                "#donthate"
 [76] "@ellaisamazingly"     "@goontoangoblin"      "@meloncollie44"       "@thehoopcentral"      "@timmychablago"
```

To start the analysis, we have found all maximum components that are weakly connected. It has the number of 161 rows the one with the highest number of links shows 284 which is the actual “John Mayer”’s user. This puts the singer on the top of all users with the collected data. More details can be seemed below.

Filter

	V1
1	@johnmayers
2	@landroverusa
3	@guitarworld
4	@youtube
5	#fridaylivestream

```
> twomode_comps$no
[1] 161
> twomode_comps$csize
  [1] 284  19   5   3   2   2   4   2   3   4   2   3   2   2   2   2   2   3   2  12   4   5   3   3   4   2   4   4   2   2   2   2   2   2   3   4  13   2
 [38]   2   3   2   2   2   4   2   2   2   3   4   2   2   5   2   2   3   2   2   2   2   2   2   2  14   2   3   2   2   2   2   2   2   2   3   2   2
 [75]   4  11   9   3   2   6   2   2   3   2   2   2   3   2   2   2   2   2   3   2   2   2   4   2   2   5   2   3   2   2   2   4   2   2   2   2   2
[112]   2   5   2   2   3   2   2   3   2   4   3   2   2   2   2   2   2   2   6   4   5   2   2   2   6   3   2   2   2   2   2  11   3   2   2   2   2   2   3
[149]   3   2   2   2   3   2   4   3   3   2   2   2   2
```

```
$membership
          @johnmayers          @landroverusa          @guitarworld               @youtube     #fridaylivestream
                    1                      1                      2                      1                      3
    #sb19voicejustin                  @mtv            @sb19official     #writingcommunity            @johnmayer
                    3                      3                      3                      4                      1
             @901jazz              @leehesng              @carskelly               @dbacks                  @mlb
                    5                      6                      7                      7                      7
      @atlanticrethink          @landrovers           #gayformayer          @lovingtay89       @secretmoments13
                    1                      1                      8                      9                      9
          #newvehicle           #therelease         @declanmckenna          @edwardsluke04         @alicewegmann
                   10                     10                      1                     11                     12
            @imbabi13              @khanak08         @velasquezjoans          @persecutedsgin     #oregonfires2020
                   13                     14                     15                     16                     17
                #rrbc        #fridayfeeling          @johnmayer❤️           @keithurban               @hozier
                   18                     19                     19                     19                      1
         @kingskworld           @rapcaviar          @babyyymika           @landrover            @theatlantic
```

We create subgraphs for better in dept analysis. Since the artist John Mayer was already in the top users list, and in fact the most top one (we isolate that component per size). The ins and outs of the subgraph were collected.

```
> twomode_subgraph
IGRAPH fa2f5cf DNW- 284 521 --
+ attr: type (g/c), name (v/c), user_id (v/c), label (v/c), weight (e/n)
+ edges from fa2f5cf (vertex names):
 [1] @rapcaviar     ->#caviarconvos    @rapcaviar     ->@pennysoulmates  @pennysoulmates->@rapcaviar      @pennysoulmates->#caviarconvos
 [5] @jadpybara     ->@johnmayers      @jadpybara     ->@landroverusa    @adweek         ->@johnmayer     @adweek         ->@atlanticrethink
 [9] @adweek        ->@landrovers      @stuartschorr  ->@johnmayers      @stuartschorr   ->@landroverusa  @stuartschorr   ->@johnmayer
[13] @stuartschorr  ->@atlanticrethink @stuartschorr  ->@landrovers      @_ferrivictor   ->@johnmayers    @_ferrivictor   ->@landroverusa
[17] @_lizcarrion   ->@johnmayers      @_lizcarrion   ->@landroverusa    @_lukelg        ->@youtube       @_robertapassos->@johnmayers
[21] @_robertapassos->@landroverusa    @4ndrewjb      ->@johnmayer       @achunadia      ->@youtube       @addiemendes    ->@johnmayers
[25] @addiemendes   ->@landroverusa    @adri_rth      ->@johnmayers      @adri_rth       ->@landroverusa  @aidensntdelphi->@johnmayers
[29] @aidensntdelphi->@landroverusa    @alejavillegasp->@johnmayers      @alejavillegasp->@landroverusa  @alexa_mymo     ->@johnmayers
+ ... omitted several edges
```

Ins and outs degrees of 284 length subgraph were gathered.

While gathering the top 20 of the items listed you can see in the list below the top inner strength *degree* centrality. @johnmayers (John Mayer) has the second highest *degree* centrality in this case. He is in the top hashtags and top users who tweet the most.

@landroverusa	@johnmayers	@johnmayer	@youtube	@theatlantic	@landrover	#bandgift
187	182	39	22	20	15	4
#gift	#music	@atlanticrethink	@rapcaviar	#caviarconvos	#guitar	@landrovers
4	4	3	3	3	3	2
#getoutside	#newdefender	@pennysoulmates	@adweek	#fender	#som	
2	2	2	2	2	2	

While gathering the top 200 of the items listed you can see in the list below the "outs" *degree* centrality. @stuartschorr has the highest *degree* centrality in this case

@stuartschorr	@centresteer	@mccons	@riccispeckels	@tlstanleyla	@brandonabueg	@gutobordin_
5	5	5	5	5	4	4
@siriusoficial_	@thegoatfox	@adweek	@buddygoat612	@chrismjones05	@crystalpistol18	@dihuang
4	4	3	3	3	3	3
@duke_strad	@ejlazar	@followmelila	@jbace22	@jle7571	@kaseyw4	
3	3	3	3	3	3	

While gathering the top 20 of the items listed you can see in the list below the "in" and "out" totals by *degree* centrality @landroverusa has the highest *degree* centrality within "ins" collected as well as in the overall total of degrees identified with *degree* of 187.

@landroverusa	@johnmayers	@johnmayer	@youtube	@theatlantic	@landrover	@stuartschorr	@rapcaviar
187	182	39	22	20	15	6	5
@adweek	@centresteer	@mccons	@riccispeckels	@tlstanleyla	@pennysoulmates	#bandgift	#gift
5	5	5	5	5	4	4	4
#music	@brandonabueg	@gutobordin_	@siriusoficial_				
4	4	4	4				

Much like the *degree* centrality results above is the *Closeness* centrality results found. It puts the same users @landroverusa at top of the list with higher *Closeness* centrality level of 3.680395e-05 and @johnmayers in second with 3.498461e-05. Having said that, the results for *Closeness* centrality shows something different where calculating the "outs" which marks @mccons with highest *Closeness* centrality.

In the total of ins and outs shown below you see the most influential users.

@landroverusa	@johnmayers	@stuartschorr	@centresteer	@mccons
0.001564945	0.001533742	0.001353180	0.001351351	0.001351351

The *betweenness* centrality calculation can be seen in the below image. It points out that @johnmayers has this type of centrality counted as 18166.911 and top one is @ @landroverusa with count 20712.608

@landroverusa	@johnmayer	@johnmayers	@youtube	@ejlazar	@djid4	@stuartschorr	@thegoatfox
20712.608	18166.911	13981.026	10393.000	5202.000	4920.000	3910.803	3533.000
@jle7571	#music	@centresteer	@riccispeckels	@mccons	@legiondecency	@pwnstar629	@brandonabueg
3277.860	2746.250	2175.929	2175.929	1918.881	1895.429	1895.429	1671.000
@dihuang	@theatlantic	@rapcaviar	#guitar				
1390.250	1285.050	840.500	840.250				

The artist appears in the top 10 of *degree* centrality calculation results. As a matter of fact, he is in the top 3 listed with score 182 just behind top user counted at 187. Although having a high *degree* centrality does not guarantee a high *Closeness* centrality, the artist does have a high *Closeness* centrality of score 3.498461e-05, which is far ahead the third listed user in this *degree* list, however not too far behind the top user with score of 3.680395e-05.

The calculation of the *betweenness* centrality puts the artist now in third with score of as per shown table below.

@landroverusa@johnmayer@johnmayers@youtube@ejlazar@djid4
20712.60818166.91113981.02610393.0005202.0004920.000

These results shown that @johnmayers is considerably with a very low *betweenness* centrality in comparation to @landroverusa and @johnmayer

Two other artists, Jason Mraz and James Morrison, were chosen to do a comparation against John Mayer's results. When using the same dataset used for John Mayer's above analysis the results for Jason Mraz give somewhat different results. For instance, the *betweenness* centrality of John Mayer is still the second highest, however, it is higher (19843.9440) than previously mentioned (18166.911). Similarly, John Mayer's *closeness* centrality changes from 3.498461e-05 to 3.192746e-05 and *degree* centrality count did not change.

Degree centrality where the same for Jason Mraz and James Morrison in the same dataset.

@landroverusa@johnmayers@johnmayer@youtube@theatlantic@landrover
18718239222015

Closeness centrality where the same for Jason Mraz and James Morrison in the same dataset.

@landroverusa@johnmayers@johnmayer@youtube@theatlantic
3.347504e-053.192746e-051.373740e-051.281673e-051.272410e-05

Betweenness centrality where the same for Jason Mraz and James Morrison in the same dataset.

@landroverusa@johnmayer@johnmayers@youtube@ejlazar@stuartschorr
21561.941119843.944014405.69238382.90008133.78334203.2477

Jason Mraz and James Morrison results accumulated in 767 of the two-mode network and has 159 of no count and its user with the highest count accumulated to 291.

As one can see in the below graph, some of the influential users of John Mayer also appears to be influential users for Jason Mraz. Likely due to similarities in the artist music style.

```
      @johnmayers       @landroverusa       @guitarworld            @youtube    #fridaylivestream
                1                   1                  2                   1                    3
#sb19voicejustin             @mtv          @sb19official #writingcommunity             @johnmayer
                3                   3                  3                   4                    1
         @901jazz          @leehesng           @carskelly            @dbacks              @mlb
                5                   6                  7                   7                    7
```

Another data collection with instead of 1000 tweets but 5000 tweets was gathered. 4529 of those tweets end up in our two-mode network. All search terms that were used for all three artist were moved in order to create the two-mode graph. After use the mode weak the no resulted in 790 count.

```
> length(V(twomode_graph))
[1] 4529
> twomode_comps$no
[1] 790
> twomode_comps$csize
  [1]  27  25   2 540   6   2   2   3 245   2   3   2   2   2   2  38  11 130   2   3   2   2   4  13   7   2  81  16   3  14   6
 [32]  23   6   3   3   2   2   3   4   2  16   7  15   6   2   5   5   2   2   3   9   2   9   5   2   2   2  26   2  11   3   2
 [63]   3   2   2   2  11   2   4   2   5   5   4   2   2   2   3  16   4   4   9  13   5  12   2  45   4   2  22   2   2   2   3
 [94]   3   3   3   2   2   3   2   3   2   2   9   4   4   4   5   4   2  14   6   4  41   2  99  17   3   5   7  12   5   5   8
[125]  30   4   2   8   2   2   2   2   2   2  38   4   2   4   9   2  19   2   4 177   5  14   7   2   4   8   2   3   2   6   2
[156]   4   2   2   3   3   3   2   4   3   2  16   4   2   3   8   4   7   2   7   2   9   4   2   4   2   4   4   3   2   2   3
```

540 ins and outs were found. When comparing the artist in the larger dataset two-mode network the following results were found. Jason Mraz appears to have a greater degree centrality compared to the other two artists.

```
> V(twomode_subgraph)[order(in_degrees, decreasing = TRUE)[1:100]]
+ 100/540 vertices, named, from b3f7a3a:
  [1] @youtube          @jasonmraz        #nowplaying       @johnmayer        @bobmayer
  [6] #live             #events           #fierceentertainment #oifw          #orlando
 [11] #podcast          #talkshow         @askmonicamay     @bcheroes         @kogsonwax
 [16] @topps            @williamshatner   @combkex          @waxio            @mimimala1
 [21] @benandbenmusic   @jacobcollier     @mayerpm          #askbenandbenjam  @theatlantic
 [26] @giseleofficial   #music            #kisstheground    #nationaldaughtersday @michaelmayer
 [31] #crypto           @iansomerhalder   @woodyharrelson   #history          #lookforthegood
 [36] #mindthegap       @datatransradio   @thatericalper    @kissthegroundca  #nfts
 [41] @netflix          #nft              @landrover        #headlinerusa     @tidal
 [46] @jzed74           #quote            @kompaktrec       #throwbackthursday @davidarquette
```

@johnmayers has a lower degree of 3 than @jasonmraz with 53 while taking into consideration the top 100 with highest degree centrality. Having said that, James Morrison does not even appear in the top 400 listed which make one assume that he has the degree of ZERO in this dataset as most of the other users in the top 500 shows degree of ZERO.

@youtube	@jasonmraz	#nowplaying	@johnmayer	@bobmayer
67	53	40	34	22
@jacobcollier	@theatlantic	@giseleofficial	#kisstheground	@bcheroes
8	7	7	6	6
@combkex	@kogsonwax	@mayerpm	@topps	@williamshatner
6	6	6	6	6
ationaldaughtersday	@iansomerhalder	@woodyharrelson	#music	@michaelmayer
6	6	6	6	5
@thatericalper	@kisstheground ca	@netflix	#history	@landrover
5	5	5	4	4
#lookforthegood	#mindthegap	@datatransradio	#quote	@kompaktrec
4	4	4	4	4
#throwbackthursday	@davidarquette	@rosariodawson	@joshtickell	@rebeccatickell
4	4	4	4	4
@waxio	#breakfastshow	@delphiband	@dermotkennedy	@dualipa
4	3	3	3	3
@kylieminogue	@nikkershaw	@owenpaulreal	@theleerocker	@thestraycats
3	3	3	3	3
@taylorwane69	#headlinerusa	@jzed74	#crypto	#daughters
3	3	3	3	3
#friday	#dtradio	#live	#askbenandbenjam	@benandbenmusic
3	3	3	3	3
@kisstheground oc	@johnmayers	@landroverusa	#deephouse	#housemusic
3	3	3	3	3
@keithurban	#hope	#darknetflix	#doctorwho	#homeschooling
2	2	2	2	2
#timetravel	@logansroadhouse	@tidal	@cosmintrg	@dynamitemc
2	2	2	2	2
@iambop	@manifestodj	@mcdrs	@yaksound	#army
2	2	2	2	2
#fridayreads	#military	@tonimoons	#nfts	#fridaythoughts
2	2	2	2	2
#coversong	#events	#fierceentertainment	#oifw	#orlando
2	2	2	2	2
#podcast	#talkshow	@askmonicamay	@allanrsavory	@drmarkhyman
2	2	2	2	2
@tombrady	@understandingag	@enhypenmembers	@littlemix	@michaelbuble
2	2	2	2	2
@parsonjames	#listenlive	#capitolstudios	#fbf	#flashbackfriday
2	2	2	2	2
#recordingsession	@capitolstudios	#regeneration	#dance	@100fables
2	2	2	2	2
#bbradio	#berlin	#nellyfurtado	#rockt	@dayvidbillini
1	1	1	1	1
@batman0kan	@fatmakracaa	@tahinlikahin	@coolguspub	@emilahp
1	1	1	1	1
@jimoneillnc	@kimmiosborne	#gabebrown	#johnwick	#kisstheground movie

On the other hand, the results for John Mayer related to *closeness* centrality are better. The artists do how within the top 60 listed in the below image. John Mayer has a lower *closeness* centrality of 3.454912e-06 when comparing with Jason Mraz with 3.809524e-06.

@youtube	@jasonmraz	#nowplaying	@johnmayer	@bobmayer
3.922245e-06	3.809524e-06	3.717887e-06	3.688404e-06	3.581611e-06
@theatlantic	@jacobcollier	@giseleofficial	#kisstheground	#nationaldaughtersday
3.500506e-06	3.487370e-06	3.480840e-06	3.474321e-06	3.474321e-06
@iansomerhalder	@woodyharrelson	#music	@mayerpm	@combkex
3.474321e-06	3.474321e-06	3.474309e-06	3.474297e-06	3.474249e-06
@kogsonwax	@topps	@williamshatner	@bcheroes	@thatericalper
3.474237e-06	3.474237e-06	3.474237e-06	3.474225e-06	3.467827e-06
@kisstheground ca	@netflix	@michaelmayer	@kompaktrec	@landrover
3.467827e-06	3.467827e-06	3.467815e-06	3.467815e-06	3.461357e-06
#quote	#throwbackthursday	@davidarquette	@rosariodawson	@joshtickell
3.461357e-06	3.461357e-06	3.461357e-06	3.461357e-06	3.461357e-06
@rebeccatickell	#history	#lookforthegood	#mindthegap	@datatransradio
3.461357e-06	3.461345e-06	3.461345e-06	3.461345e-06	3.461345e-06
#dtradio	@waxio	#breakfastshow	@delphiband	@dermotkennedy
3.461345e-06	3.461262e-06	3.454912e-06	3.454912e-06	3.454912e-06
@dualipa	@kylieminogue	@nikkershaw	@owenpaulreal	@theleerocker
3.454912e-06	3.454912e-06	3.454912e-06	3.454912e-06	3.454912e-06
@thestraycats	@taylorwane69	#daughters	#friday	@kisstheground oc
3.454912e-06	3.454912e-06	3.454912e-06	3.454912e-06	3.454912e-06
@johnmayers	@landroverusa	#deephouse	#housemusic	#headlinerusa
3.454912e-06	3.454912e-06	3.454912e-06	3.454912e-06	3.454900e-06
@jzed74	#crypto	#askbenandbenjam	@benandbenmusic	#live
3.454900e-06	3.454876e-06	3.454852e-06	3.454828e-06	3.454745e-06

Now, considering the *betweenness* centrality of all three artists as show below, John Mayer has the centrality of 1.500000 (at the end of the top 180 listed) when and Jason Mraz of 88790.219526 (the second on the top 5 listed).

#nowplaying	@jasonmraz	#music	@zozozo333	@johnmayer
97818.640476	88790.219526	68478.750000	67515.000000	66923.625000
@youtube	@dms_ambon	@hothitsuk247	@ifm1005ibadan	#throwbackthursday
42523.208333	35567.500000	35567.500000	34765.000000	32284.500000
#history	@mayer_pm	@grannjarmo	#crypto	@bob_mayer
29158.500000	28979.000000	28740.333333	28378.000000	27074.500000
@publishingfire	@shellidiego	@thesound_box	@bobmayer	@chalkz0ne_
22258.333333	22258.333333	22258.333333	21107.000000	20875.500000
@kompaktrec	#kisstheground	@ajdcbarnett	@evankirstel	@jobinindia
17374.190476	17217.524293	16003.000000	14756.000000	14756.000000

…continue…

@metalgrounds1	@pepperjanemusic	@rodger_lydia	@anonymouslyobv3	@cheesmanart
178.666667	178.666667	178.666667	139.833333	139.833333
@badailalu	@d_berberi	@jamesnord	@nicksloggett	@delcasi1
133.750000	133.750000	133.750000	133.750000	95.468074
@giseleofficial	@alligatortearsq	@barberpicards	@hey_quackidee	@nhgirlusa
79.475959	75.714286	75.714286	75.714286	75.714286
@sergiobaca_	@drift4king1	@dutch_mandel	@edem_k	@matt_j_campbell
75.714286	62.946429	62.946429	62.946429	62.946429
@mehmeto33440789	@kissthegroundca	@woodyharrelson	@iansomerhalder	@davidarquette
62.946429	49.684639	44.021286	42.591884	38.672327
@rosariodawson	@meawwofficial	@mayerpm	@kissthegroundoc	@joshtickell
38.672327	25.496992	23.500000	17.714618	15.814827
@rebeccatickell	@jacobcollier	#darknetflix	#doctorwho	#homeschooling
15.814827	14.833333	14.000000	14.000000	14.000000
#timetravel	#lookforthegood	@landrover	@tonimoons	@waxio
14.000000	5.000000	3.000000	3.000000	3.000000
@bcheroes	@kogsonwax	@topps	@williamshatner	@johnmayers
1.500000	1.500000	1.500000	1.500000	1.500000
#dtradio	#mindthegap	@datatransradio	#headlinerusa	@georgenexblanco
1.390476	1.107143	1.107143	1.000000	1.000000

When considering the last largest dataset one can presume that Jason Mraz is a more popular artists than John Mayer and both are much popular than James Morrison. Two Twitter datasets that were obtained using different search terms to gather these results.

When using Gephi the confirmation of information can be seen while pointing out the edges and nodes.

# of Nodes:	**4529**
# of Edges:	**5553**

Figure 8: Degree centrality check.

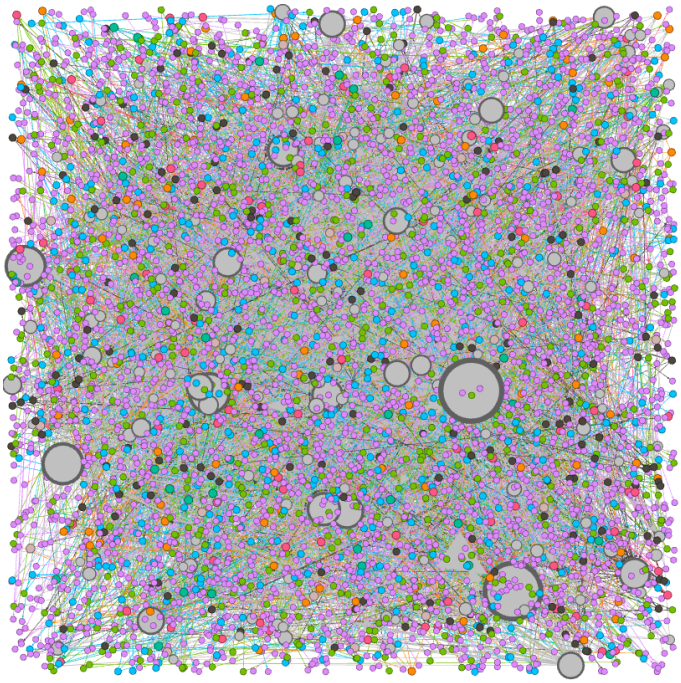


Figure 9: Closeness Centrality Check

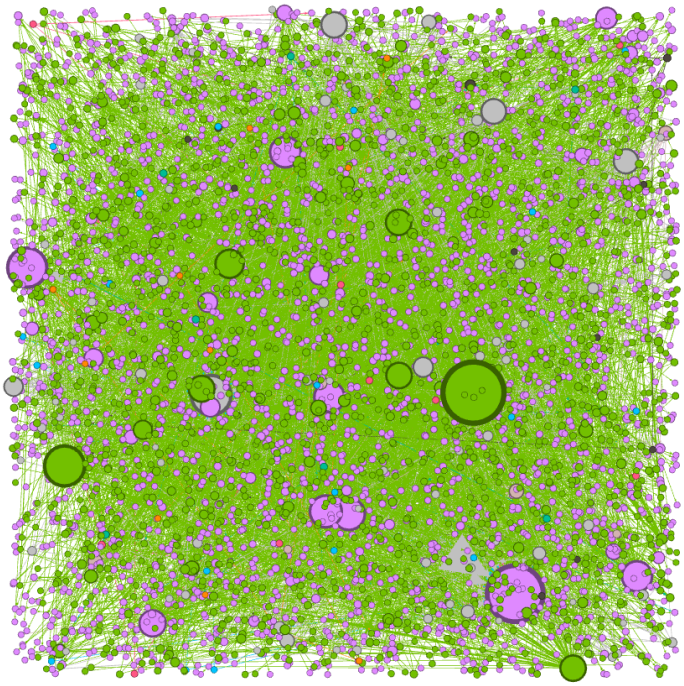


Figure 10: Betweeness Centrality Check

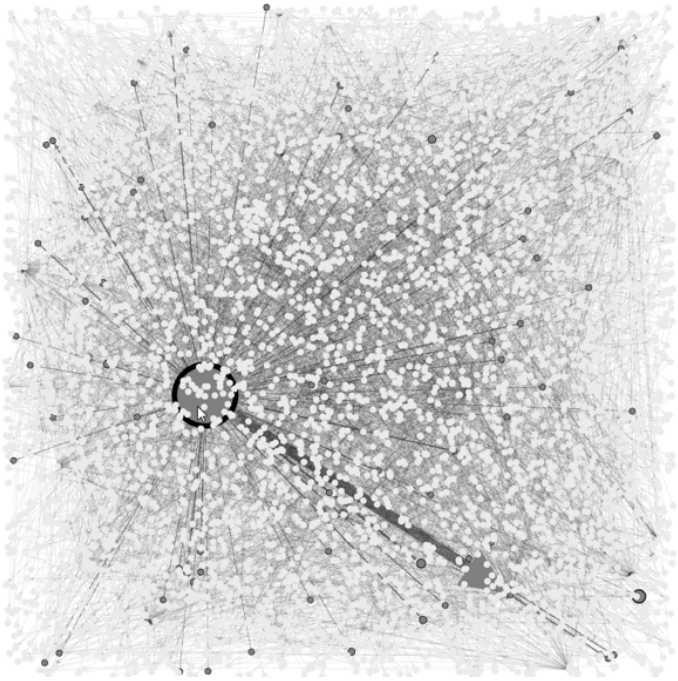
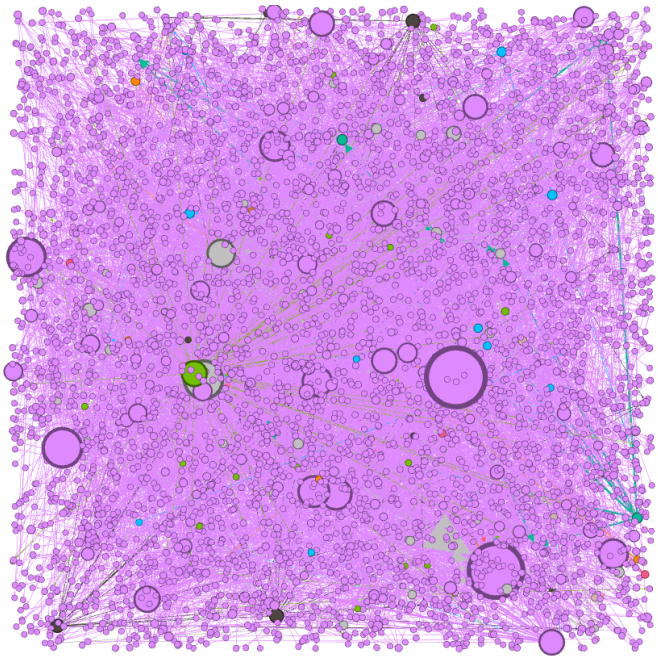


Figure 11: Betweeness Centrality Check



2.5. Perform community analysis with the Girvan-Newman (edge betweenness) and Louvain methods. Explain how relevant the results are to your artist/band. Perform the community analysis also for related artists. Is their community structure similar?

Comments from three popular John Mayer YouTube videos have been collated, with the user from comments collected into an actor network graph files. The actor network files has been exported to a graphml file format for analysis in Gephi.

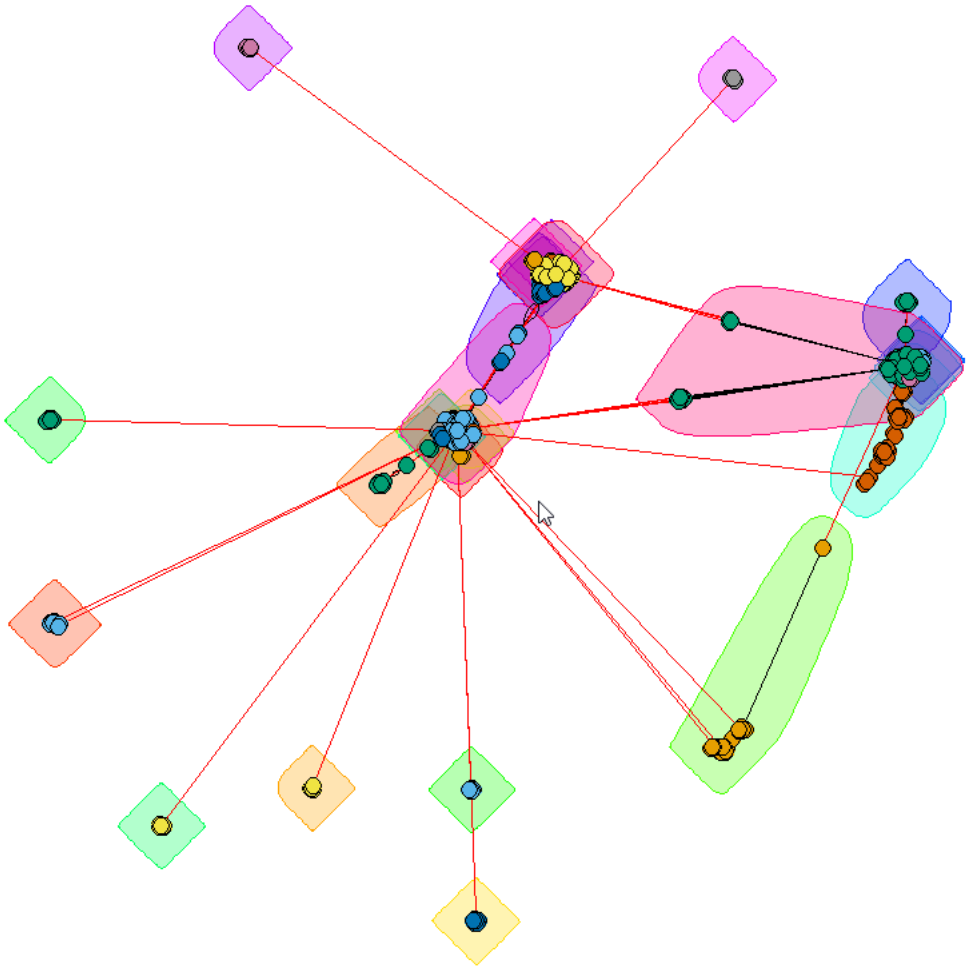
Running the Louvain algorithm across our actor network, we have detected 39 distinct communities of varying size:

Figure 12: Distinct Communities Detected

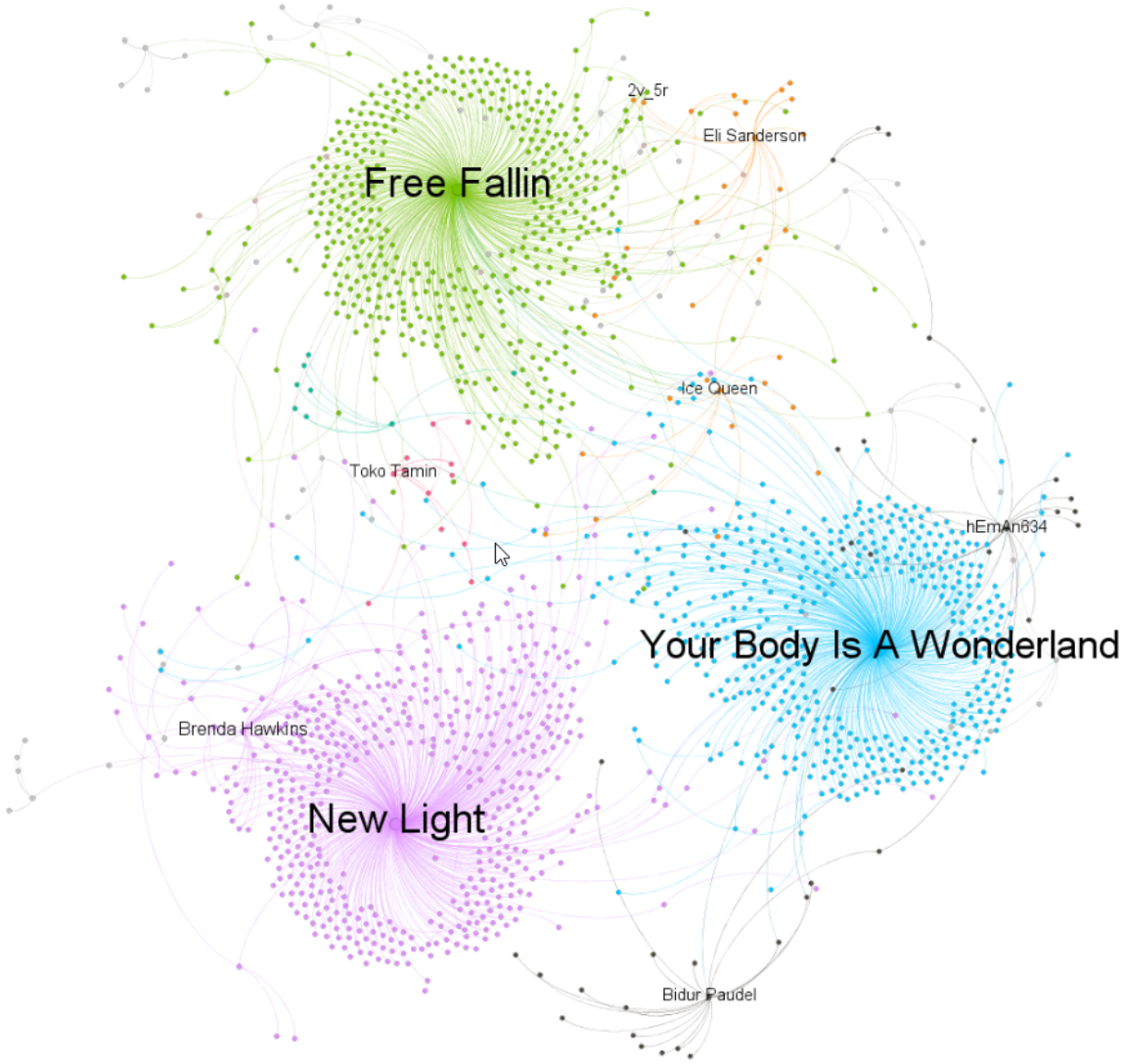
	Community.sizes	Freq
1	1	5
2	2	2
3	3	2
4	4	2
5	5	12
6	6	8
7	7	5
8	8	7
9	9	3
10	10	3
11	11	3
12	12	32
13	13	2
14	14	6
15	15	8
16	16	6
17	17	4
18	18	40
19	19	2
20	20	2
21	21	2
22	22	3
23	23	3
24	24	3
25	25	6
26	26	5
27	27	9
28	28	3
29	29	2
30	30	46
31	31	3
32	32	2
33	33	5
34	34	4
35	35	5
36	36	2
37	37	467
38	38	448
39	39	446

Analysis by community size, we have three communities with a large number of nodes\actors, namely 446,448 and 467, a further three with a small to moderate amount, 32,40,46 and lastly a majority of 33 that consist of a single digit number of nodes\actors.

Firstly, we plot the actor graph and communities from within R and removed the labels to make the plot less cluttered. The result can be seen in the figure below.



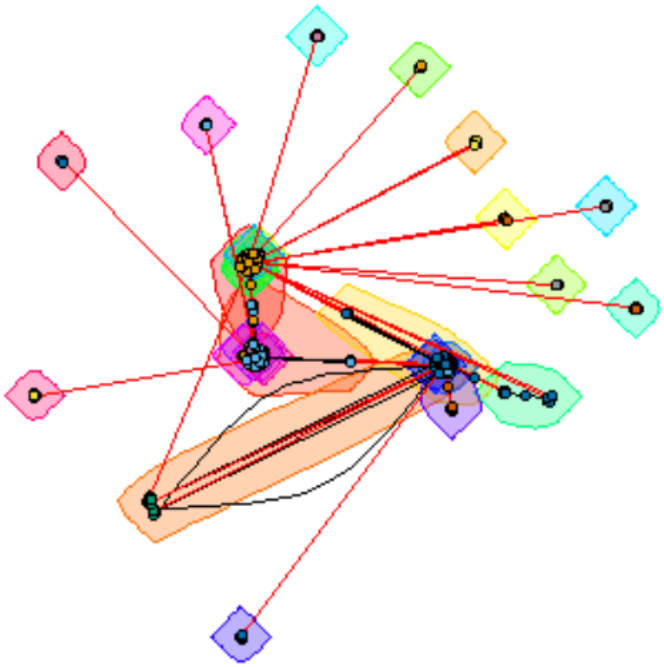
We then import the actor graph into Gephi, as an undirected graph with an edges merge strategy of last.



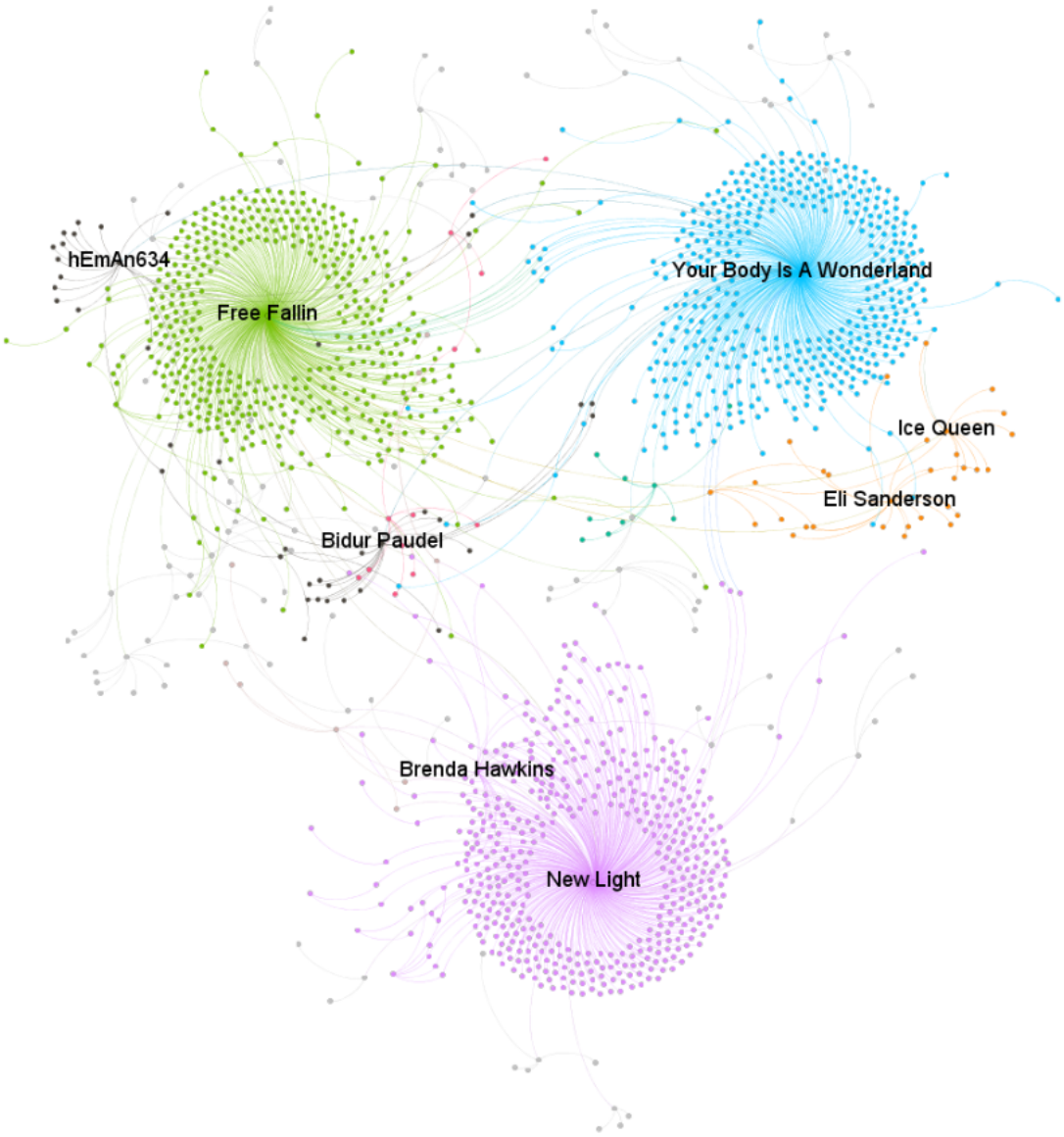
The modularity in Gephi produced 18 modularity classes(communities),as opposed to R's implementation of Louvain producing 39, however we did note a very similar structure to the community node membership, namely three very large communities with a few 10-40 size and mostly very small communities.

Running Girvan-Newman (edge betweenness) across the same dataset within R, we arrive at 29 communities, with similar results in terms of community node member numbers as within Louvain. Three very large communities with 2 much smaller moderate size ones and many very small.

	Community.sizes	Freq
1	1	474
2	2	489
3	3	40
4	4	12
5	5	457
6	6	8
7	7	5
8	8	5
9	9	7
10	10	3
11	11	3
12	12	3
13	13	31
14	14	6
15	15	8
16	16	6
17	17	4
18	18	3
19	19	3
20	20	3
21	21	6
22	22	9
23	23	9
24	24	3
25	25	5
26	26	4
27	27	3
28	28	5
29	29	4



Running Girvan-Newman within Gephi yields 31 communities.



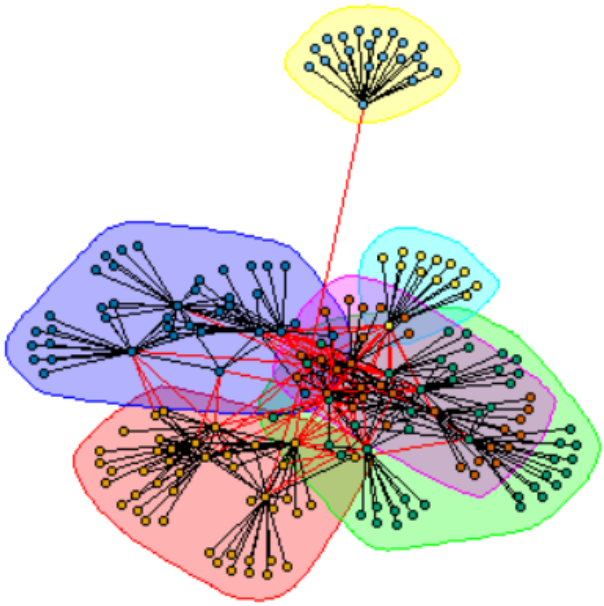
It is noted that both Louvain and Girvan-Newman have produced very similar results for our John Mayer dataset.

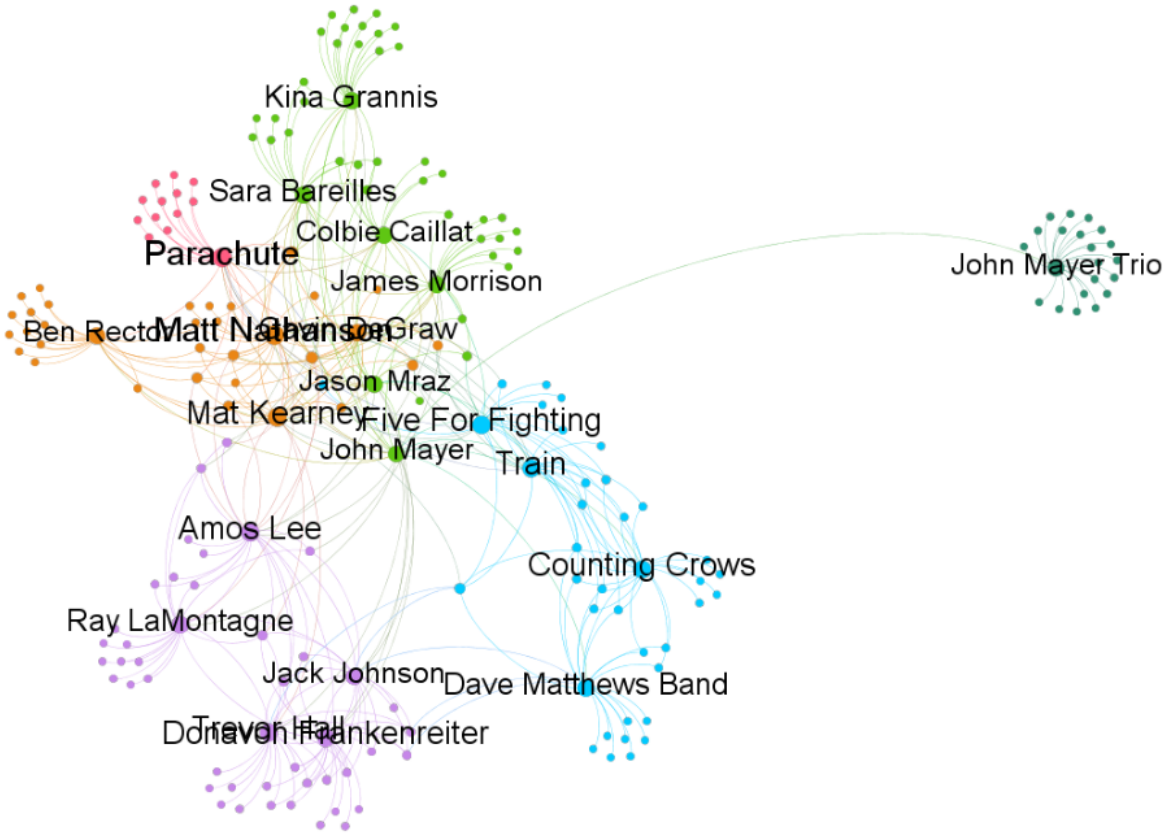
The results show actors commenting on three of John Mayer's most popular YouTube videos and from the community analysis we can see how some of the actors relate across comments on videos. The three main communities are structured around the videos with many smaller communities forming around them.

Next, we move on to examine John Mayer related artists as per our data from Milestone 1 Q1.8.

Louvain analysis with R shows 6 communities with more similar node numbers:

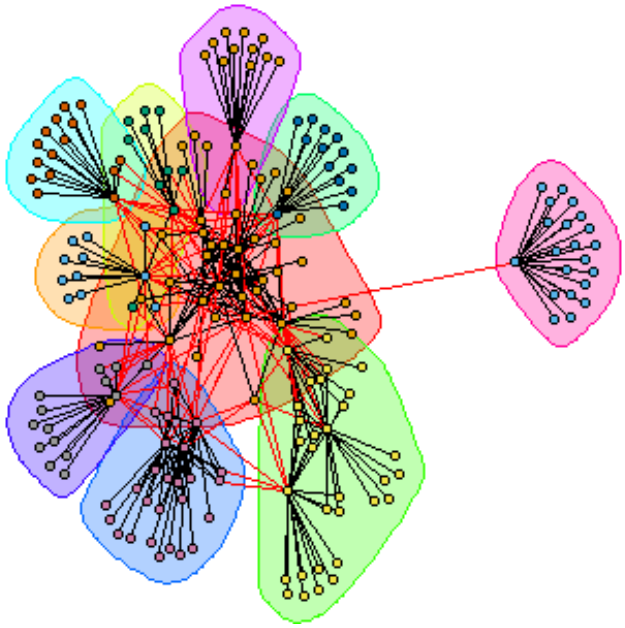
	Community.sizes	Freq
1	1	44
2	2	21
3	3	46
4	4	13
5	5	39
6	6	31





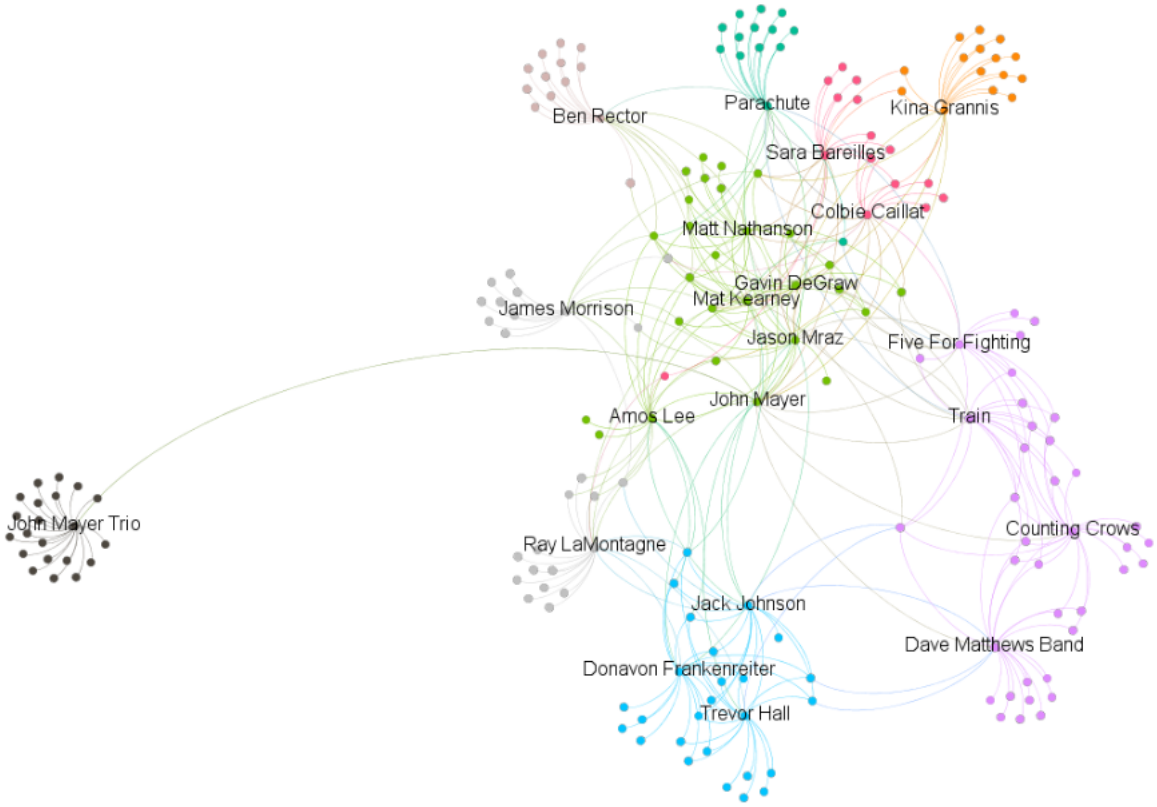
Running Girvan-Newman (edge betweenness) across the same related artists dataset within R, we arrive at 10 communities:

	Community.sizes	Freq
1	1	40
2	2	11
3	3	10
4	4	31
5	5	13
6	6	15
7	7	26
8	8	14
9	9	13
10	10	21



We can clearly see a central community in red – Which is John Mayer at the heart of his related artists, as an important edge betweenness community.

For related artists, running Girvan-Newman within Gephi yields 10 communities:



Again, we note the similarities in the community output of both algorithms. The edge betweenness R plot really shows the connected nature of the related artists back to John Mayer.

In conclusion, we have analysed community structures using both the Girvan-Newman (edge betweenness) and Louvain methods visually across both R and Gephi.

The relevant results have showed similar structures across both methods. We have noted the connected nature of YouTube actors commenting on John Mayer's videos.

Related artists Twitter actor network has demonstrated the community nature of edge betweenness to show how different communities are connected through a central cluster.

Machine Learning Models

2.6. Use sentiment analysis to identify how the public reacts to events and/or topics related to your artist/band. Provide a summary of public opinions (emotions, reactions).

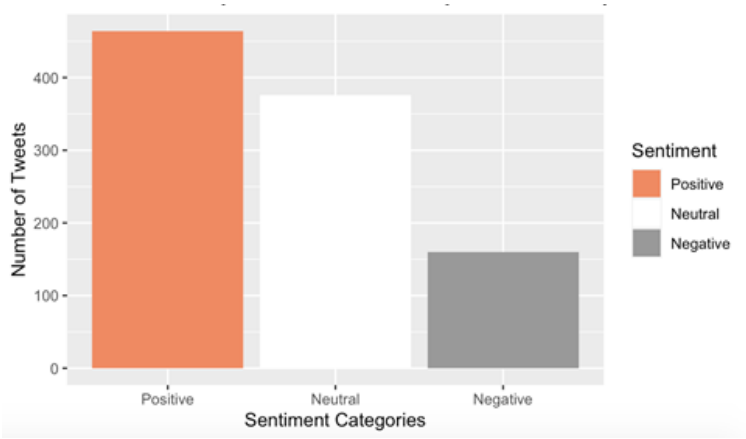
In order to perform sentiment analysis that could provide relevant insights to improving John Mayer's popularity, our team has used

Our team has provided sentiment analysis on twitter and youtube data. For twitter, we have used the same dataset collected for Milestone 1, whereas for youtube we have retrieved comments from two videos. One video is the Land Rover advertisement with John Mayer and the other video is from the song "Who you love", which he sings with Katy Perry.

Analysis 1: Twitter Sentiment Analysis

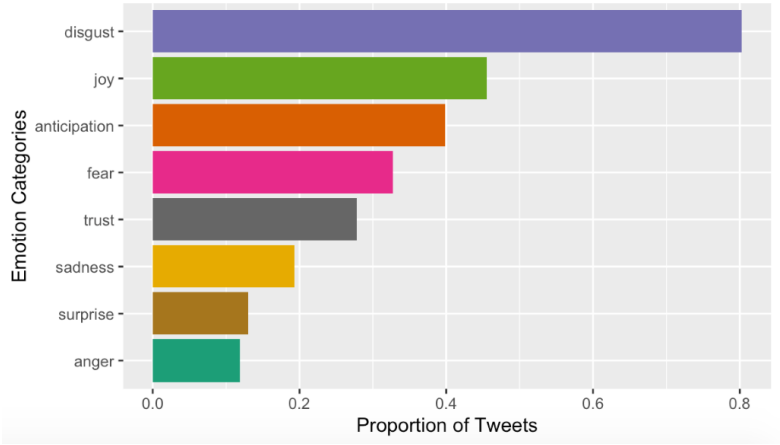
Twitter data collected for Milestone 1 was used to identify how the public reacts to topics related to John Mayer. After cleaning the data, we used the get_sentiment() function to assign sentiment scores to each tweet, and then, we converted the numbers to labels, where –1 became "Negative", 0 became "Neutral" and +1 became "Positive". The result of this first part of the analysis can be seen in the figure below. Overall, the sentiment is mostly positive, but with a significant number of neutral sentiment, which shows a lot of room for improving John Mayer's popularity and acceptance by the public.

Figure 13: Sentiment Analysis of Tweets with Keyword 'John Mayer'



In the sequence, we used get_nrc_sentiment() function to get a more detailed view of the sentiments, where the presence of anger, anticipation, disgust, fear, joy, sadness, surprise and trust can be detected. Furthermore, the proportion of these emotions is calculated, and the result can be seen in the following figure:

Figure 14: Emotion Analysis of Tweets with Keyword 'John Mayer'



As we can see, the neutral feelings from the previous plot, now tend to be more distributed towards the negative feelings. Overall, the sentiment analysis of these tweets points that disgust is the predominant emotion among users, followed by joy in the second place and anticipation, in the third place. Interestingly almost all tweets were classified with a certain level of disgust.

Analysis 2: Youtube Sentiment Analysis (Land Rover Advertisement)

Our team has also conducted a Youtube sentiment analysisfor the Land Rover advertisement (video ID xRvsC0sqguc) where John Mayer features. The results show that there is lower number of users with neutral sentiments (figure 15) and in particular when looking at the different individual emotions that the add evokes on people, more positive emotions are on the top. Unlike the previous twitter analysis, the analysis of Youtube comments on this add shows that people have very positive feelings about this partnership between John Mayer and Land Rover. Between the top three emotions, two of are positive (figure 16): joy (in the first place) and trust (in the third place).

One possible reason for the add having such a high acceptance among the public is because it invites people to go outside, disconnect from their devices and connect with the nature. During COVID, this appeals to people very much and it was a smart marketing catch. Therefore, a relevant insight is that it is important not to lose the right momentum for partnering for advertisement. For instance, in this case, it is the COVID that provides a good base for marketing with nature and freedom.

Figure 15: Sentiment Analysis of Land Rover Advertisement John Mayer.

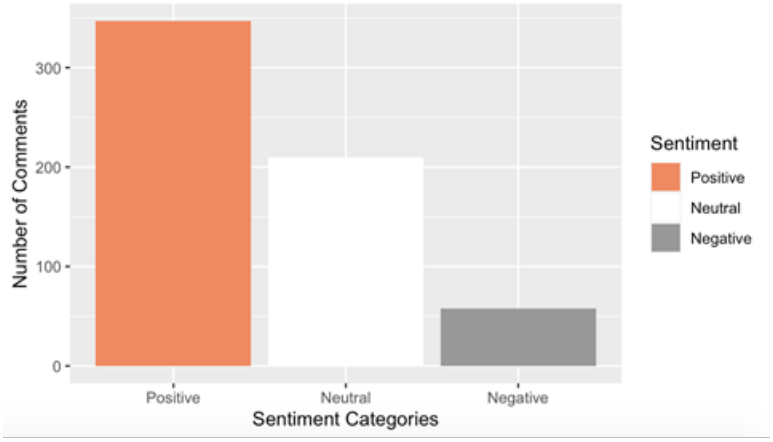
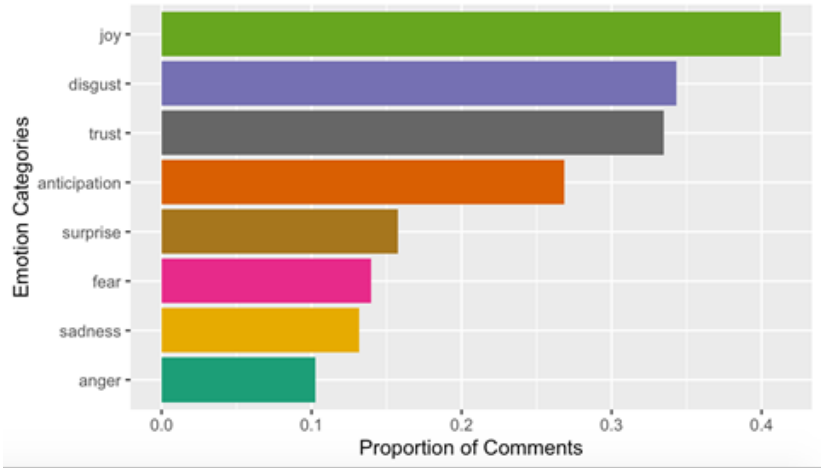


Figure 16: Emotions Analysis of Land Rover Advertisement John Mayer.



Analysis 3: Youtube Sentiment Analysis (song "Who you Love" – feat. Katy Perry)

The last sentiment analysis was on comments about the video clip "Who you love" where John Mayer and Katy Perry sing together. When looking at figure 17 it's possible to see that neutral sentiments are higher than positive and negative. However, a more in-depth analysis (figure 18) shows that these neutral sentiments tend to be more distributed towards positive emotions. This video clearly shows that this partnership had a high acceptance by the public and, therefore, was successful. As a recommendation to improve John Mayer's popularity we would suggest partnering with more female singers with profiles similar to Katy Perry.

Figure 17: Sentiment analysis of the song "Who you love"

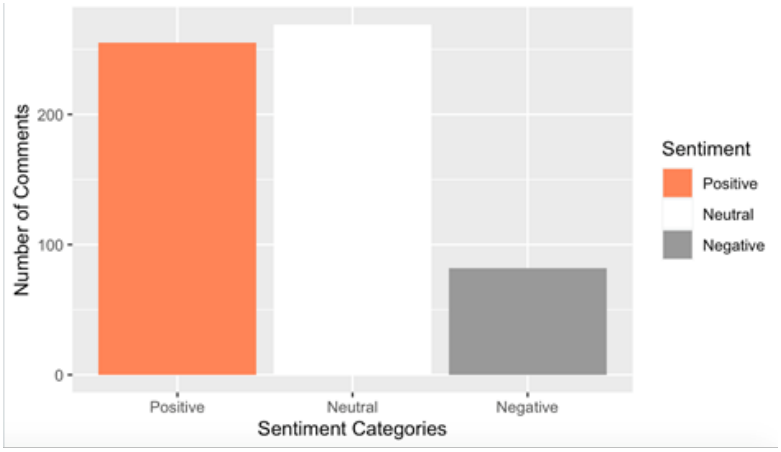
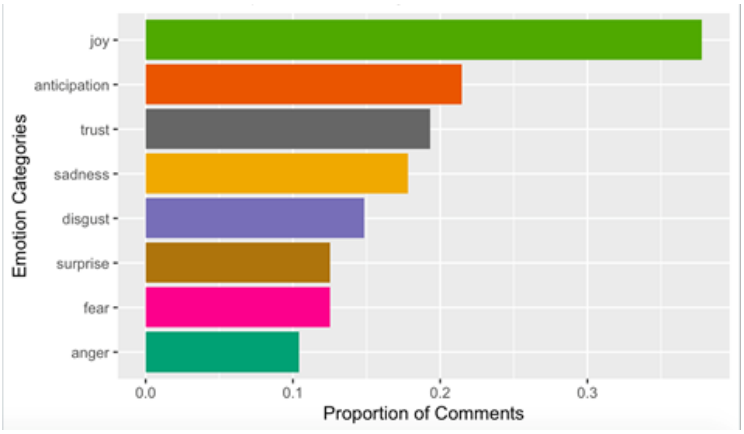


Figure 18: Emotion Analysis of the song "Who you love"



2.7. Build a decision tree and evaluate its performance in predicting whether a song is by your artist/band.

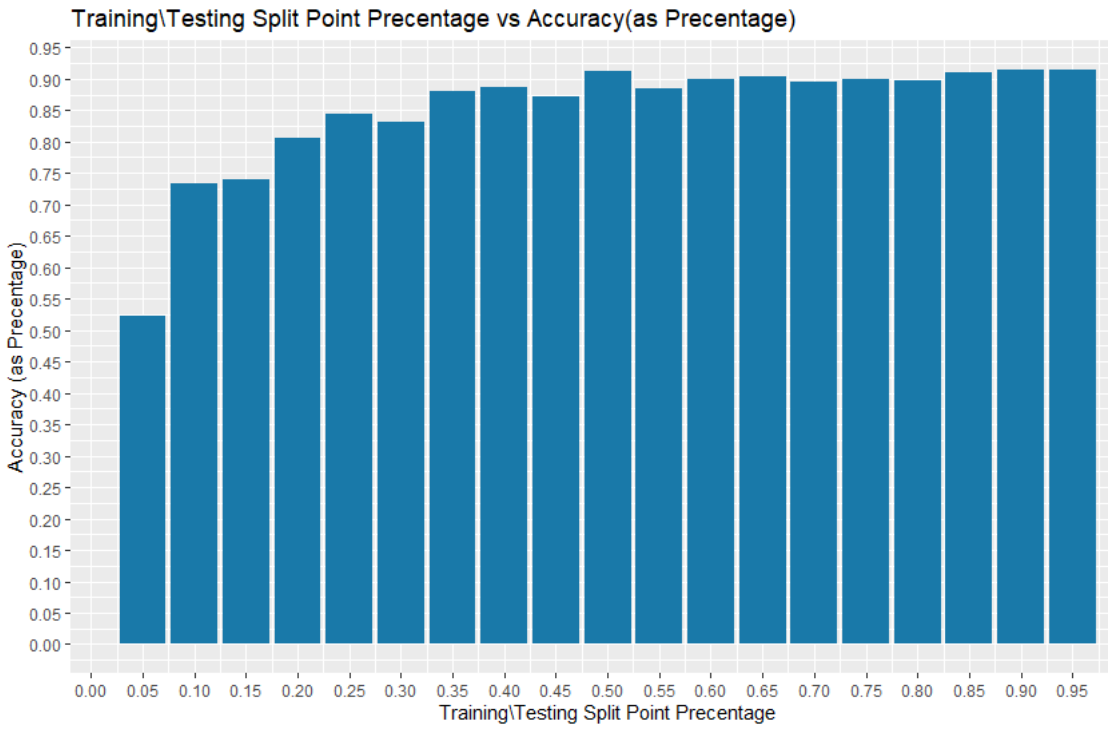
A decision tree model has been trained to predict if a song is a John Mayer song or not. The initial training and testing dataset used comprises of all John Mayer's songs coupled with songs from Spotify's top100 playlist. The full dataset contains approx. 169 John Mayer songs and 100 songs from Spotify's top 100 songs playlist.

The approach taken was to evaluate the accuracy of predictions for many different training and testing data split points. Our test will evaluate a split of 0.05 of the data to be used as training data (remaining to be used as testing data) through to 0.95 of the data to be used as training data by 0.05 increments. We then view and plot the split point percentage value against the accuracy of the model as a whole.

Figure 19: Decision Tree Model Accuracy Against Training\Testing Split Percentage Table

	SplitPointPrecent	Accuracy
1	0.05	0.5229358
2	0.10	0.7333333
3	0.15	0.7406340
4	0.20	0.8055556
5	0.25	0.8436975
6	0.30	0.8319783
7	0.35	0.8804220
8	0.40	0.8863179
9	0.45	0.8716094
10	0.50	0.9120521
11	0.55	0.8846431
12	0.60	0.8999324
13	0.65	0.9042945
14	0.70	0.8956772
15	0.75	0.8997290
16	0.80	0.8973055
17	0.85	0.9103707
18	0.90	0.9153005
19	0.95	0.9144681

Figure 20: Decision Tree Model Accuracy Against Training\Testing Split Percentage Plot



From a training\testing split of 50\50 we note very little change in accuracy, with the greatest accuracy at 0.90 split.

We then re-run our model with a 90\10 split and observed the full confusion matrix output:

```
Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 32.2  3.8
         1  5.0 59.0

 Accuracy (average) : 0.9118
```

We see a relatively positive result of 91.2% accuracy(True Positive + True Negative). Our decision tree model will correctly predict if a song from the dataset is a John Mayer song 91.2% of the time. Drilling into these results we observe:

- True Positive result of 59%, i.e. the song is a John Mayer song and the model predicted it was a John Mayer song.
- True Negative result of 32.2%, i.e. the song is not a John Mayer song and the model predicted correctly that is was not a John Mayer Song.
- False Positive result of 5%, i.e. the song is not a John Mayer song, but the model incorrectly predicted it was a John Mayer song
- False Negative of 3.8%, i.e. the song is a John Mayer song, but the model incorrectly predicted it was not a John Mayer song.

Further analysis was conducted with a different dataset, namely combining a playlist of songs from a vastly different genre. The genre of metal was chosen as it will have very different values for the song feature properties that should provide for a larger contrast between metal songs and John Mayer songs.

We then re-run our model with a 90\10 split on the new dataset and observed the full confusion matrix output:

```
Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 33.8  0.4
         1  2.7 63.1

 Accuracy (average) : 0.969
```

We can see a significant improvement in accuracy, as expected with a more contrasting song list.

In conclusion, the relatively lower accuracy on the first model with the top 100 song list is due to the similar nature of John Mayer's songs to songs in the top 100, both being of similar genres, namely pop. The second result shows how we can achieve a better accuracy against data that is very different from John Mayer. This highlights a limitation of the decision tree model for prediction and should be considered accordingly.

2.8. Use k-means clustering to classify a user's friends (following) and followers. You have to identify one influential user related to your artist/band and analyse his/her friends and followers. Justify why he/she is an influential user. Explain the results.

Twitter user "TheAtlantic" is a very influential user related to John Mayer. From Milestone 1 the page rank algorithm was run over the twitter actor graph with "TheAtlantic" found to be the most influential. "TheAtlantic" is also influential as John is promoting Land Rover in conjunction with Atlantic's marketing team (The Atlantic, 2020). We also know that The Atlantic has many followers on Twitter(approx. 2 Million) and is an influential media company. Note that TheAtlantic has only 1,056 friends(users TheAtlantic follows).

After collecting Twitter friend and follower information for TheAtlantic, a logarithm was applied to make the scale more meaningful when plotting. The "elbow method" (Julianhi, 2013) was applied to our data to gauge the best number of clusters(as k). A value of 4 or 5 was found to be optimal.

Figure 21: Elbow Method Results

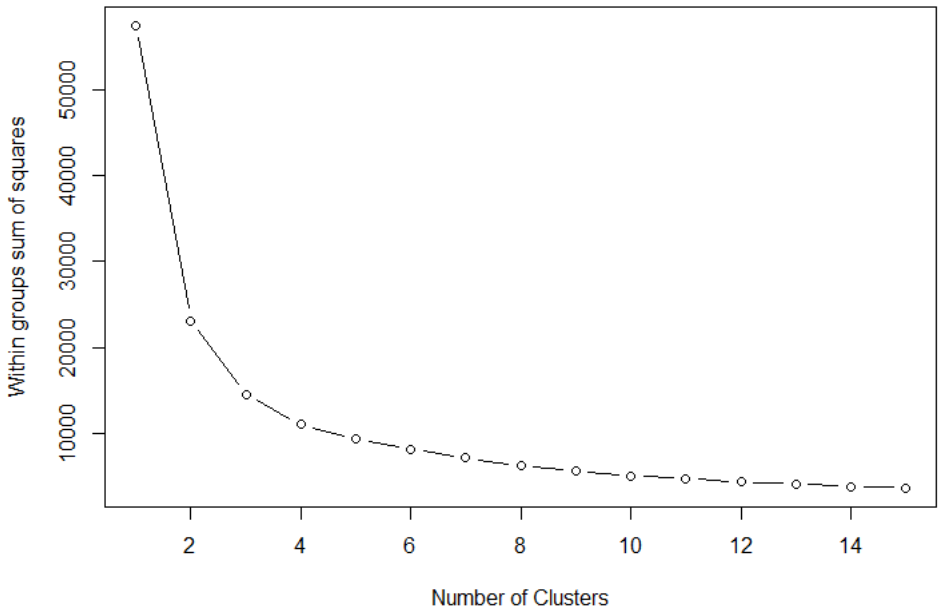


Figure 22: TheAtlantic Follower Users: Friends and Followers with 4 Clusters

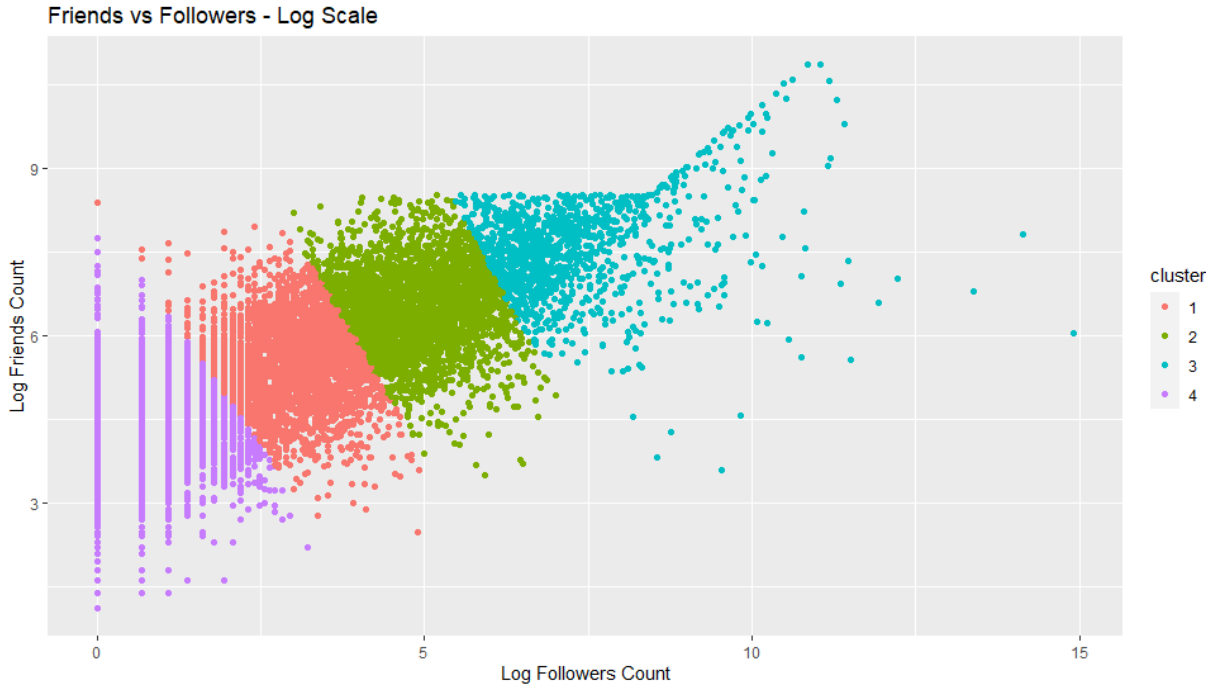
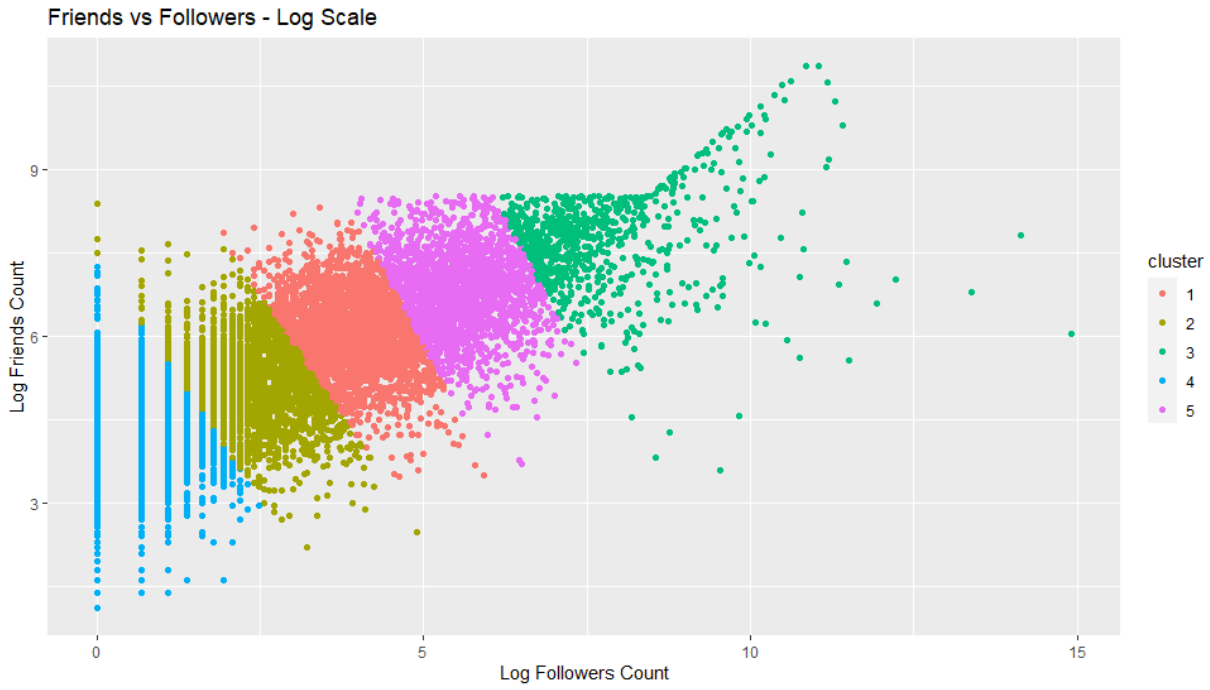


Figure 23: TheAtlantic Follower Users: Friends and Followers with 5 Clusters



Using 5 clusters was found to better capture the lower and higher end of the dataset. We note in the upper right above log friend count of approx. 8.5 and log follower count of approx. 8.25, we see a sharp and well-defined linear increase for a smaller subset of users,

indicating the smaller subset of users that have both a large number of followers and friends. Beyond this we note the outliers that have a very large number of followers but not a large number of friends(people they follow) comparatively. TheAtlantic itself falls into this category as the user has 2 million followers but only has approx. 1000 friends.

We could infer that users in the green cluster 3, are likely more influential as many more users will receive their tweets, increasing the visibility of those tweets. Overall, the clusters demonstrate the clear relationship between a user's friends and their follower counts. As users become more engage in the platform, they would follow more users(friend them) and be followed by more users – possibly with the latter occurring in greater numbers as a user becomes more popular.

Two lessons we can learn for John Mayer, is to look to increase his follower count to improve his popularity – likely by posting and replying to more content. Also, to engage more with users from the green cluster 3 in the top right corner to try and make content about John Mayer "go viral" to also increase his popularity. An assumption would be made that the more content about John Mayer we can get out into the Twittersphere , the more revenue can be obtained through music streaming and ultimately users attending live shows.

2.9. Use LDA topic modelling to identify some terms that are closely related to your artist/band. Find at least 3 significant groups of words that can be meaningful to your analysis. Explain your findings.

Topic modelling was conducted on our original twitter data from milestone one to find terms closely related to John Mayer. During the data preparation phase, experimentation was done with the sparsity level to achieve the best mix of meaningful terms into topics. Reducing the sparsity to 0.99 provided for 12 topics. Reducing the sparsity level any lower than 0.99 provided too few terms, with repeated terms throughout all topics, making it difficult to derive any insight.

Figure 24: John Mayer LDA Topic Modelling Maximum Harmonic Mean (K as optimal number of topics)

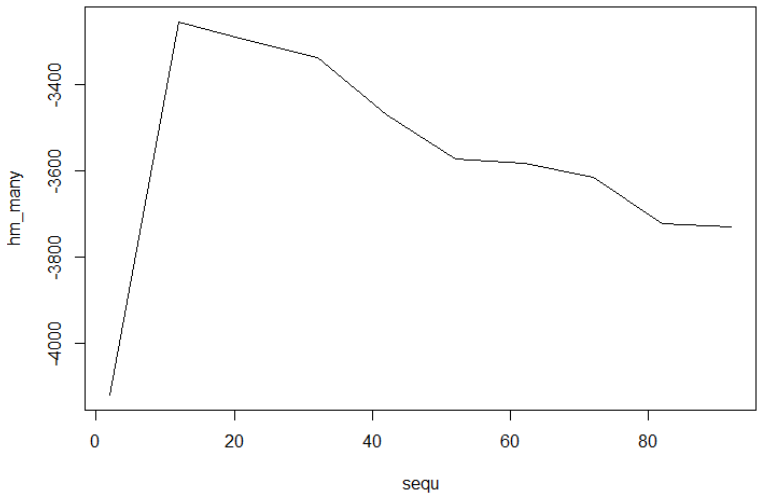


Figure 25: John Mayer LDA Topic Modelling Results – Topics and Terms

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	listen	cri	year	tear	bodi	smile	cover	amp	taylor	wonderland	day	thing
2	world	loud	live	today	red	eye	room	live	make	light	play	feel
3	wait	play	life	hous	hand	swift	burn	start	fuck	stop	atlant	slow
4	dream	chang	light	wait	back	back	featur	listen	night	gonna	rais	danc
5	boy	boy	good	hand	wonderland	perfect	content	car	releas	studio	urban	album
6	life	album	day	graviti	version	amp	good	chang	cover	featur	graviti	releas
7	amp	featur	gonna	free	perfect	atlant	day	cover	back	ass	version	life
8	album	make	wonderland	danc	car	light	ass	make	album	free	studio	today
9	good	today	tear	red	dream	ass	light	urban	loud	life	featur	make
10	hous	danc	version	gonna	cover	make	thing	bodi	content	fuck	feel	free
11	swift	room	play	ass	swift	wonderland	cri	burn	thing	version	boy	hand
12	bodi	slow	thing	thing	taylor	wait	boy	ass	rais	day	free	bodi
13	wonderland	bodi	album	car	releas	live	danc	light	atlant	chang	smile	world
14	eye	hous	content	play	hous	loud	eye	content	stop	boy	wonderland	amp
15	ass	wonderland	free	chang	eye	fuck	album	year	urban	bodi	chang	gonna

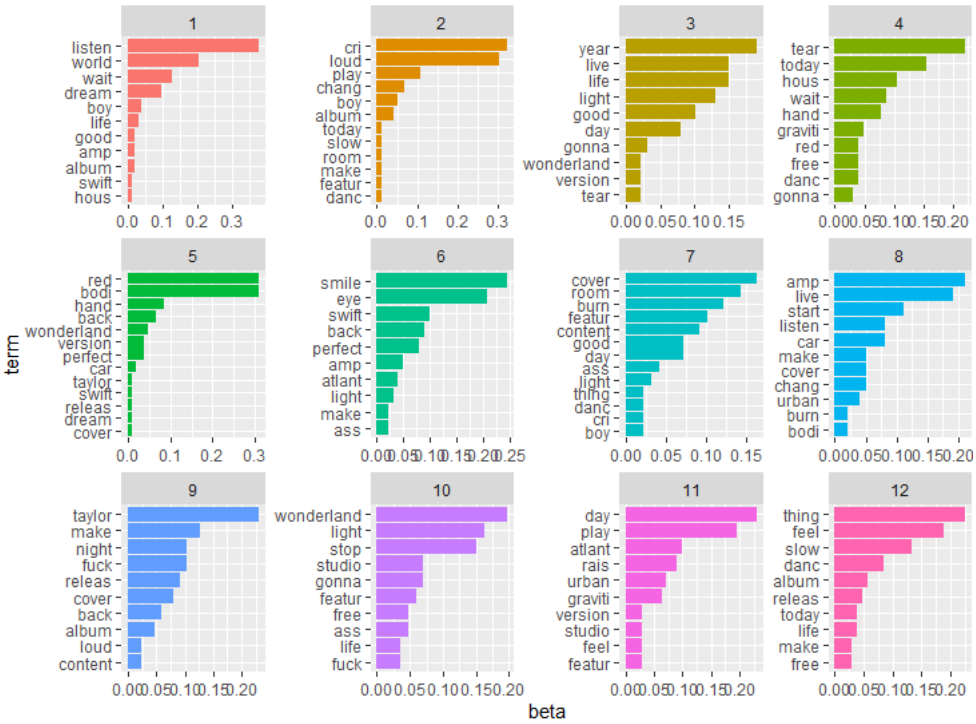
Firstly, we note many references in each topic to John's songs, for example, in topic 5 the terms "wonderland" and "bodi" are references to one of this most popular tracks "Your Body Is A Wonderland". Another example is the term "gravity" from Topic 4 referencing John Mayer's top track "Gravity" and the term "light" from topic 6 referring the track "New Light". It is clear that these tracks are quite popular as they are referenced across many tweets, often by user's praising the track or sharing as part of a suggested playlist.

Topics 4,5 and 6 denote terms referencing popular tracks along with comments describing those tracks, for example, "danc(e)", "car", "play", "perfect", "live" and "loud". This can provide insight into how users consume and interact with John's music. This could be used to focus marketing better.

John has collaborated with, and shares fans with, Taylor Swift as seen by the terms "taylor" and "swift" through topics 5 and 6. With Taylor Swift being much more popular that John, one strategy to propagate John's tweets and increase his popularity is to engage more with Taylor's fans. This could help to expose John's music to a larger audience.

Lastly, one important term is "releas(e)" as this pertains to calls for John to release more music, also to posts relating to when John released certain tracks and albums, reminding people about these events and also the release of a song for John's new Land Rover commercial. The commercial will raise John's profile and provides good publicity. The media partner for this advertisement is www.theatlantic.com – referenced by term "atlanti(c)" in Topic 6.

Figure 26: John Mayer LDA Topic Modelling - Probability of that term being generated from that (code extracted from Tidy Text Mining)



Based on the analysis of our original data above, it was decided to make another search specific to three of John Mayer's top songs and analyse using topic modelling. The twitter search used was:

(Mayer Wonderland) OR (Mayer Gravity) OR (Mayer World Change)

This is in reference to three popular songs:

- Your Body is a Wonderland
- Gravity
- Waiting on the World to Change

Please note that this search only returned 299 tweets.

Figure 27: John Mayer LDA Topic Modelling Results – Top 3 Songs - Topics and Terms

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	made	red	laugh	listen	bring	angel	inform	hous	guitar	johnmay	day	bounci
2	rememb	addit	fess	write	red	decemb	stupid	favorit	sing	obsess	trust	whoa
3	note	honest	dont	lennon	sing	qualiti	beatl	gonna	televis	start	fall	today
4	addit	rememb	theatr	solo	album	note	hurt	ain	andamp	hard	take	show
5	wed	top	gonna	shadowtodd	nowplay	nokia	stop	listen	make	los	check	los
6	solo	shadowtodd	make	nokia	stop	wed	bend	peopl	tym	yebba	hous	beckham
7	peopl	take	real	hard	top	version	rain	spicefmk	theatr	forev	real	make
8	word	theatr	week	ain	check	johnmay	guitar	forev	qualiti	home	favorit	top
9	spicefmk	ain	talk	gonna	talk	clark	andamp	guitar	clark	red	guitar	stay
10	yebba	bend	stay	hous	mine	imagin	mine	johnmay	today	fall	johnmay	beatl
11	week	wed	favorit	bend	beckham	made	write	solo	bounci	hurt	solo	lennon
12	day	tym	guitar	rain	clark	favorit	cri	check	invent	talk	fess	shadowtodd
13	imagin	cri	johnmay	cri	bounci	guitar	favorit	fess	favorit	imagin	red	take
14	hous	home	solo	invent	version	solo	johnmay	red	johnmay	word	bring	listen
15	stay	favorit	check	favorit	invent	check	solo	bring	solo	wed	whoa	week

The terms “guitar” and “solo” appear in several topics and centre around discussions of John Mayer’s guitar ability and talk of John Mayer cover songs.

Topic 12 contains discussions of John Mayer covers and song reviews, with a note that John Mayer is no Lennon. ShadowTodd is a prominent music review user on Twitter, that contrasts John Mayer and The Beatles\John Lennon, in a negative way. Clark Beckham has done a seemingly popular cover of Gravity. The leading term of topic 12 – “bounci”, references a very popular comedic retweet, appearing 11 times in this dataset:

“John Mayer: Your body is a wonderland. Me: My body is a bouncy house.”

In conclusion, it is noted that raising John Mayer’s profile and therefore his popularity can be improved and analysed by looking at discussions about his most popular songs.

MILESTONE 3

Case Study Setting

Revisit Question 1.2) from Milestone 1 and improve your answer based on your new knowledge that you gained throughout the last weeks. [2-3 paragraphs, 0.7 mark]

- How do you want to improve the popularity? How can social media analytics help you achieve it?

In order to improve John Mayer popularity, our team has come up with some recommendations and explanation on how social media can help us achieve each of them:

- Participation in more advertisements.

With data retrieved from social media, it is possible to understand which types of advertisements are likely to give John Mayer more publicity. This can be done by finding related artists and which kind of advertisements these artists have done. Using sentiment analysis, we can see which adds evoked positive feelings and, with TF-IDF we can understand the features the users find more interesting in each add. Then, John Mayer can engage on advertisements that would bring him positive reactions from the public.

- Increase in the number of interactions with its users on social networks.

Social media analysis can also help understand how users interact with content related to John Mayer and where the influential users are. This knowledge can be used to focus the interaction with the right people and boost the way information about the singer is spread in the network.

- Collaboration with artists with whom John Mayer shares a similar fan base.

Social media analytics can help to identify those artists which are related to John Mayer. These can be considered for future collaboration in order to increase his visibility inside a network that, due to the identified similarities, has a high probability of providing new fans.

- Creating songs with content that is appealing to his audience.

Again, sentiment analysis can help us to identify opportunities to improve his content as well and TF-IDF can be used to have a better glance at which issues should be covered in the music/video clips in order to have a high acceptance by the public. Plus, LDA topic modelling can also help to identify terms that are closely related to John Mayer.

- What kind of social media data do you want to analyse?

Our team will focus the analysis on **Twitter, Spotify and Youtube content**. However, in future it is also important to consider other social networks such as Instagram, Facebook , Amazon, Apple Music, Soundcloud and Deezer. Plus, we need to always be vigilant regarding new social medias that may arise.

- *What is your hypothesis (expectation) about the analysis outcome?*

Our hypotheses are:

- Advertisements help increasing John Mayer visibility.
- More interaction with social media users provides good publicity.
- Collaboration with artists increases John Mayer popularity.
- Covers are a good option to rise John Mayer's reputation.

With social media analytics it was possible to test these hypotheses and it became evident that advertisements such as the Land Rover had a very positive impact on John Mayer's visibility. Furthermore, it was clear that John Mayer's fans use more replies and retweets than mentions, which shows that more posts would give them content on which to retweet and reply and as a result, John Mayer would have more publicity. Collaboration with artists was also confirmed as a good strategy since some of the analysis pointed that names such as Taylor Swift and Kanye West were still positively related to John Mayer, even years after the collaboration. Finally, covers such as "Free Fallin" have become hits and demonstrate to have positive results on the singer's reputation.

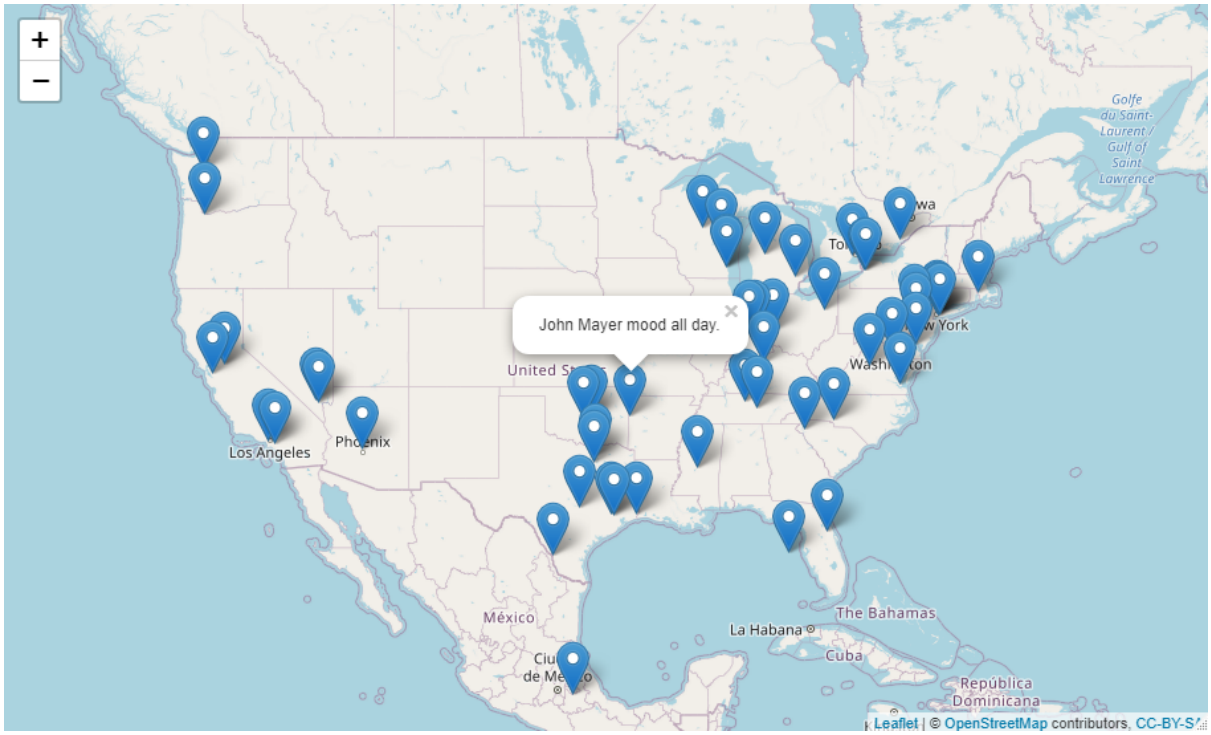
Visualisation

3.1 Plot the location of tweets from your Twitter dataset on a map using RStudio. Explain your findings. (If you do not have enough tweets with location data in your Twitter dataset that you have been using for the previous questions, you can run a new search to get a dataset with more location data.)

John Mayer is an American(USA) artist, with much of his fanbase residing in North America. John was born on the east coast of the USA, lived for a time in Boston, before moving to Atlanta Georgia and furthermore spending a lot of time in Los Angles and the state of Montana. For this reason, the continental USA was chosen as the bounding box for our geo tweet analysis.

Of 10,000 tweets requested, 1102 were turned and of that 53 contained geo tag data. These tweets have been collected, plotted, and visualized on a map of the USA.

Figure 1: John Mayer Tweets with Geo Tag Data Plotted on Map of USA



The first point that should be noted is that Twitter users must explicitly opt in to add geo tags to their tweets (Twitter, 2020). Based on this we cannot make specific representations on the number or percentage of tweets from different areas – as data is likely skewed by differing numbers of users with geo tags enabled. However, the north eastern part of the USA is more densely populated, and we do see more tweets in that region. Outside of the USA, we have one tweet in Canada and one tweet in Mexico. Notably there is not many tweets from Los Angles and no tweets from both Georgia and Montana – areas John has spent significant time. The mid and mid-north region of the USA, that is more sparsely populated, has no tweets.

An interactive html version of the map can be found accompanying this report:
Filename = map_geo_twitter_data_John-Mayer.html
Within the interactive map we can hover over tweet pins to reveal the text of that tweet.

3.2 Visualise your Twitter actor network in Gephi, with the node size determined by the number of followers for that actor. What insights can you extract from the visualisation?

The current dataset and therefore network and actor graphs do not include the attribute for the number of followers. Research was conducted on how we could get the follower attribute available in R, passed to a network a graph object for consumption in Gephi. A careful review of the vosonSML documentation demonstrates the use of the AddUserData function with the lookupUsers parameter (R Project, 2020). Using these we can query the Twitter API using our original actor network and obtain the follower information for each actor, updating our actor network object ready to export to graphml.

Figure 2: **R Code to Obtain Followers Count**

```
> # https://cran.r-project.org/web/packages/vosonSML/vignettes/Intro-to-vosonSML.html
> # Use AddUserData with lookupUsers to obtain further attributes for our actors(inc. followers)
> actorGraphWithUserAttr <- twitter_actor_network %>%
+   AddUserData(twitter_data,
+               lookupUsers = TRUE,
+               twitterAuth = twitterAuth) %>% Graph(writeToFile = TRUE)
Adding user profile data to network...
Fetching user information for 240 users.
User information collected for 237 users.
Collected user records does not match the number requested. Adding incomplete records back in.
Done.
Creating igraph network graph...
GRAPHML file written: C:/Users/daniel/Griffith University/7230ICT Group sign-up sheet 4 - General/Milestone
 3/Q3.2/2020-10-02_171432-TwitterActor.graphml
Done.
>
> # write network graph to file ready to open in Gephi
> write.graph(actorGraphWithUserAttr, file = "TwitterActorWithUserAttr.graphml", format = "graphml")
>
```

We now proceed to load our graphml into Gephi to visualize followers count using node size. Colour denotes modularity.

Figure 3: **Gephi Node Size Attribute**

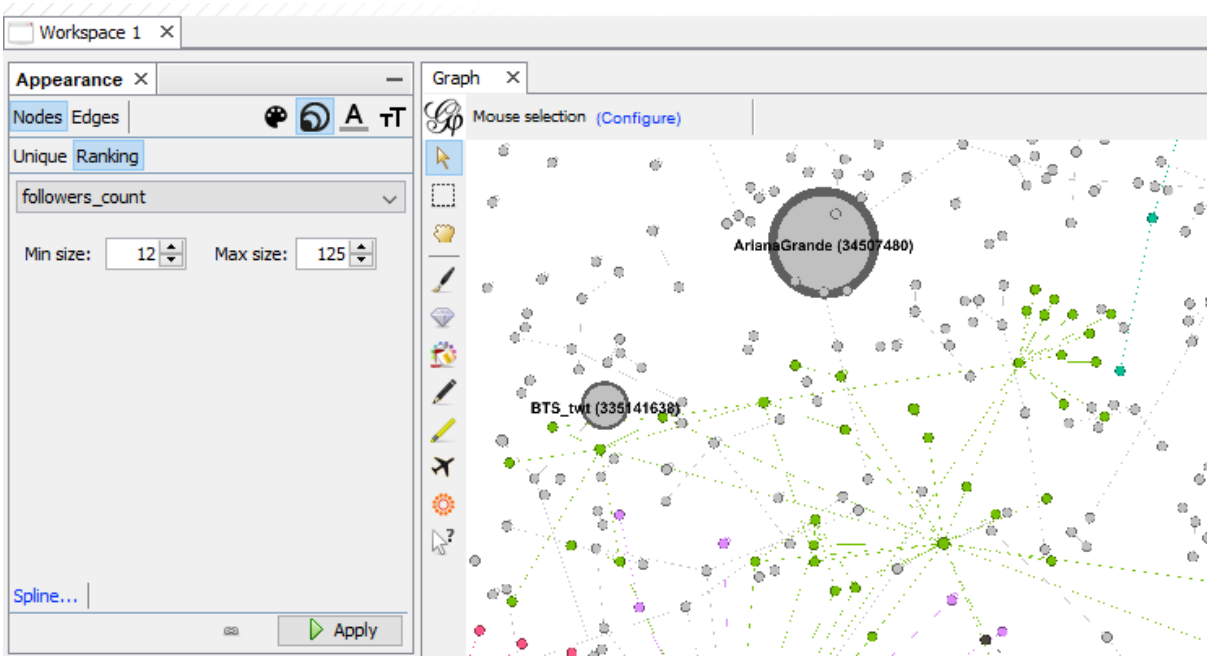
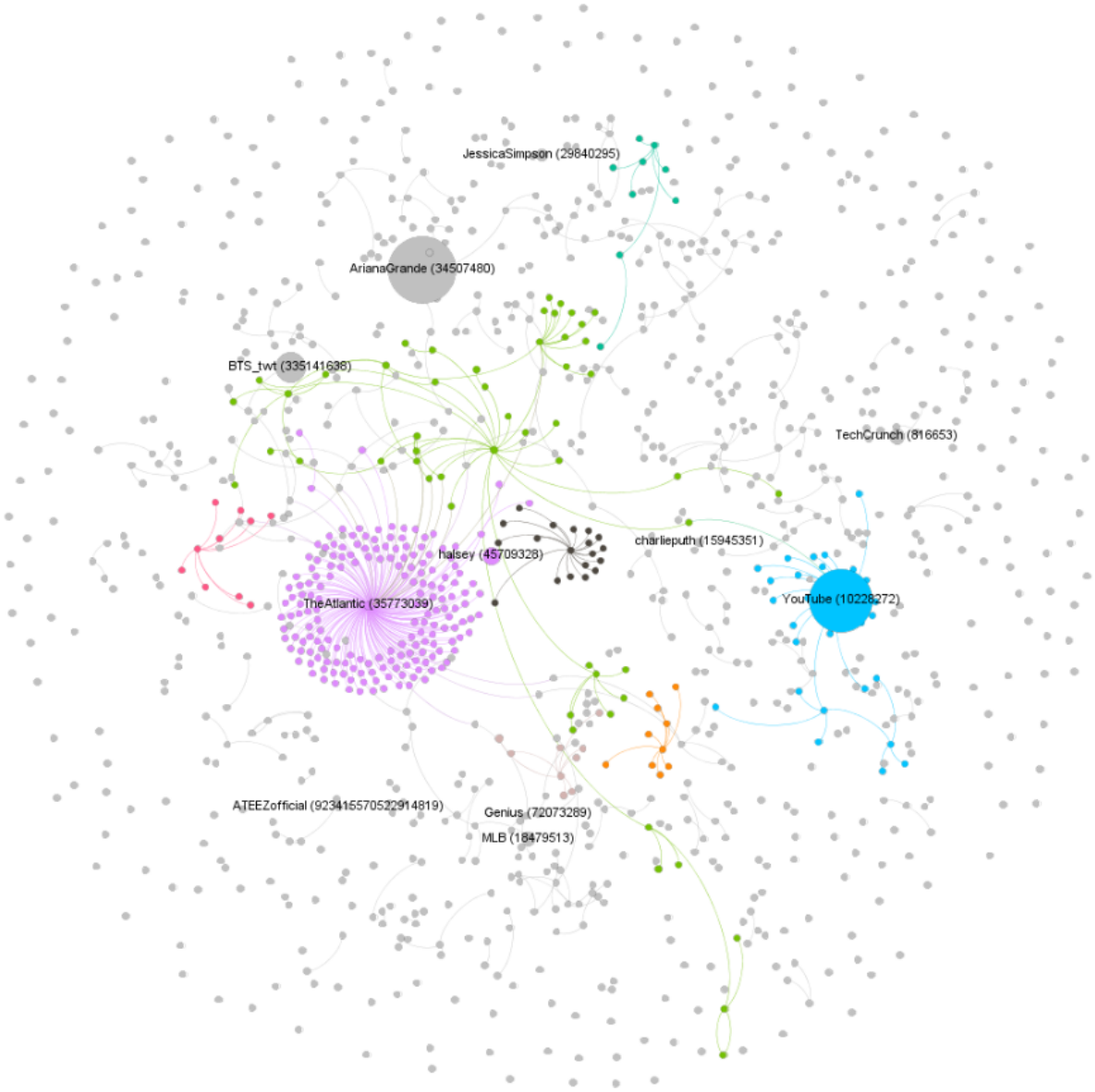


Figure 4: John Mayer Actor Network Node Size Follower Count + Modularity(Colour)



The Atlantic was the most influential user from earlier analysis using page rank, and whilst this user has a significant number of followers, from the graph we can see there are many users with a higher follower count. All users with more than 1 500 000 have a label in the visualization.

The two users with the highest follower count by far are ArianaGrande and YouTube. Ariana Grande is a very popular singer\artist within the pop genre with a very large twitter following, with YouTube understandably having many followers.

It is interesting to note that ArianaGrande is only mentioned within our John Mayer Twitter data(once in a tweet) and was not actually active. We could argue this skews this visualization

in this way. This same can be said about Jessica Simpson, also has a large number of followers but is only mentioned in our twitter data and does not directly tweet. In contrast, TheAtlantic also did not directly tweet, and is important in our dataset as many other users\actors retweeted or responded to TheAtlantic. Should we remove users that are only mentioned and leave users that have been retweeted? We could argue this would be a sensible approach.

In Conclusion, analysing the follower count for our John Mayer Twitter data has been insightful, yet slightly misleading. Further steps for this line of enquiry would be to remove users that are only referenced by mention and also to use a larger dataset.

3.3 Plot the location of tweets from your Twitter dataset on a map using Tableau. Add other attributes as additional marks to your map (e.g., as colour, size). Explain your choice of attributes and visual marks.

For question 3.2, we obtained extra attributes by using the AddUser Data and lookupUsers = TRUE parameter of the collect function. This allows for more columns in our data. The geo latitude and longitude columns were added to the data from Q3.2 to make 56 observations with 92 columns(variables) for use within our Tableau map.

John Mayer twitter data was then saved to JSON formatting using the jsonlite R library.

Figure 5: Load and Save Twitter Data to JSON – Inspect Columns

```
> # load our tweet data that has geo data and extended columns(lookupUsers = TRUE)
> load("geo_tweets_loc_only_with_lat.RData")
>
> library(jsonlite)
>
> write_json(geo_tweets_loc, "twitter_data_John-Mayer_Geo_LookupUsers.JSON", pretty=TRUE)
>
> colnames(geo_tweets_loc)
 [1] "user_id"                 "status_id"               "created_at"              "screen_name"
 [5] "text"                    "source"                  "display_text_width"      "reply_to_status_id"
 [9] "reply_to_user_id"        "reply_to_screen_name"    "is_quote"                "is_retweet"
[13] "favorite_count"          "retweet_count"           "quote_count"             "reply_count"
[17] "hashtags"                "symbols"                 "urls_url"                "urls_t.co"
[21] "urls_expanded_url"       "media_url"               "media_t.co"              "media_expanded_url"
[25] "media_type"              "ext_media_url"           "ext_media_t.co"          "ext_media_expanded_url"
[29] "ext_media_type"          "mentions_user_id"        "mentions_screen_name"    "lang"
[33] "quoted_status_id"        "quoted_text"             "quoted_created_at"       "quoted_source"
[37] "quoted_favorite_count"   "quoted_retweet_count"   "quoted_user_id"          "quoted_screen_name"
[41] "quoted_name"             "quoted_followers_count"  "quoted_friends_count"    "quoted_statuses_count"
[45] "quoted_location"         "quoted_description"      "quoted_verified"         "retweet_status_id"
[49] "retweet_text"            "retweet_created_at"      "retweet_source"          "retweet_favorite_count"
[53] "retweet_retweet_count"   "retweet_user_id"         "retweet_screen_name"     "retweet_name"
[57] "retweet_followers_count" "retweet_friends_count"   "retweet_statuses_count"  "retweet_location"
[61] "retweet_description"     "retweet_verified"        "place_url"               "place_name"
[65] "place_full_name"         "place_type"              "country"                 "country_code"
[69] "geo_coords"              "coords_coords"           "bbox_coords"             "status_url"
[73] "name"                    "location"                "description"             "url"
[77] "protected"               "followers_count"         "friends_count"           "listed_count"
[81] "statuses_count"          "favourites_count"        "account_created_at"      "verified"
[85] "profile_url"             "profile_expanded_url"    "account_lang"            "profile_banner_url"
[89] "profile_background_url"  "profile_image_url"       "lat"                     "lng"
```

The JSON file was loaded as a data source into Tableau Public. First latitude and longitude were added to rows and columns, respectively. Noting the source field contains details on the device or app used to tweet, a calculated field was created to format a new field called Source.

Figure 6: Calculated Field for Source of Tweet

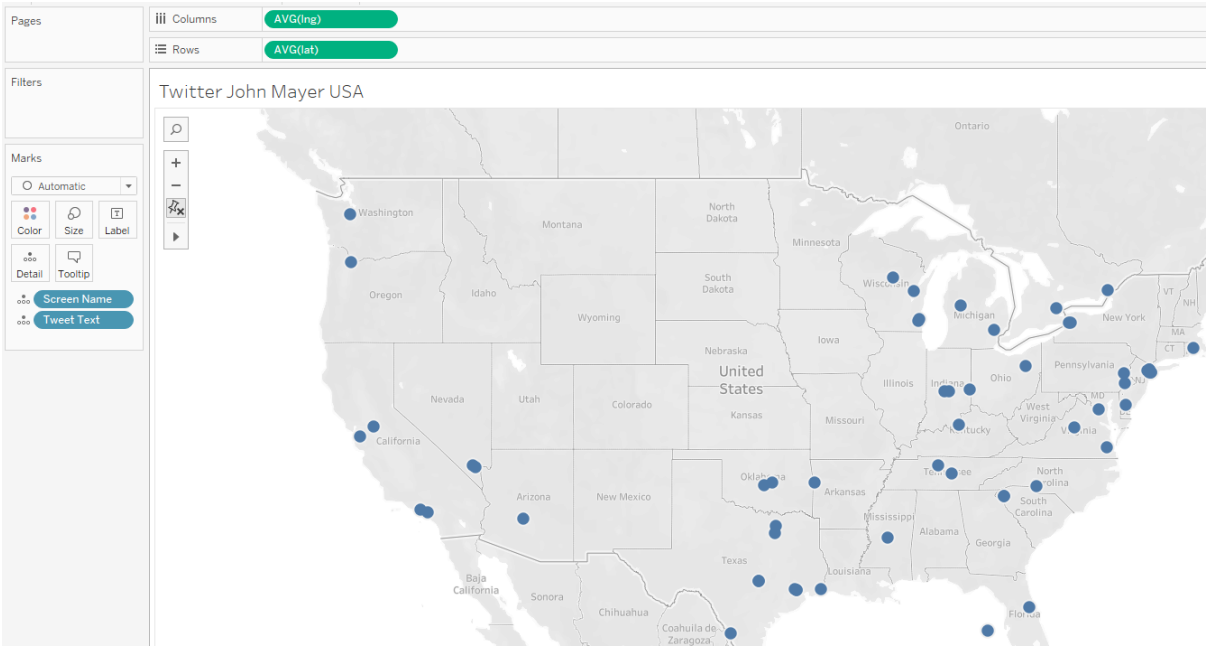
Abc twitter_data_John-Maye... source	=Abc Calculation Source
Twitter for Android	Android
Twitter for iPhone	iPhone
Twitter for Android	Android
Twitter for Android	Android
Twitter for iPhone	iPhone
Twitter for Android	Android

Source

```
CASE [source]
WHEN "Twitter for Android" THEN "Android"
WHEN "Twitter for iPhone" THEN "iPhone"
WHEN "Instagram" THEN "Instagram"
END
```

The text field was renamed to “Tweet Text” and the “screen_name” field was renamed to “Screen Name”, with both added as a mark detail. This provided for a plot of dot marks on our map of the USA for each tweet.

Figure 7: Initial Map Plot



Next, we added two additional marks, “Source” shown by colour and “Friends Count” shown by size. This allows the plot to convey spatial, source and friend count information using the same map plot format. Hovering over our dot marks, we are able to view information regarding that tweet.

Figure 8: Hover Over for Tweet Information

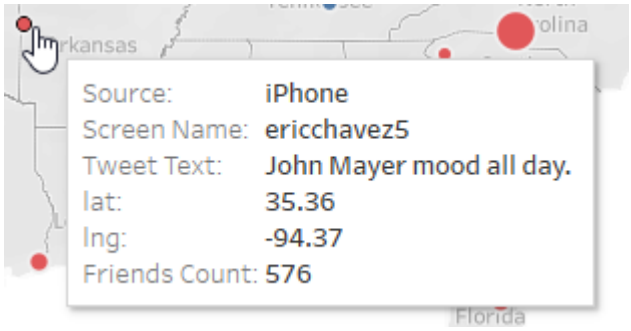
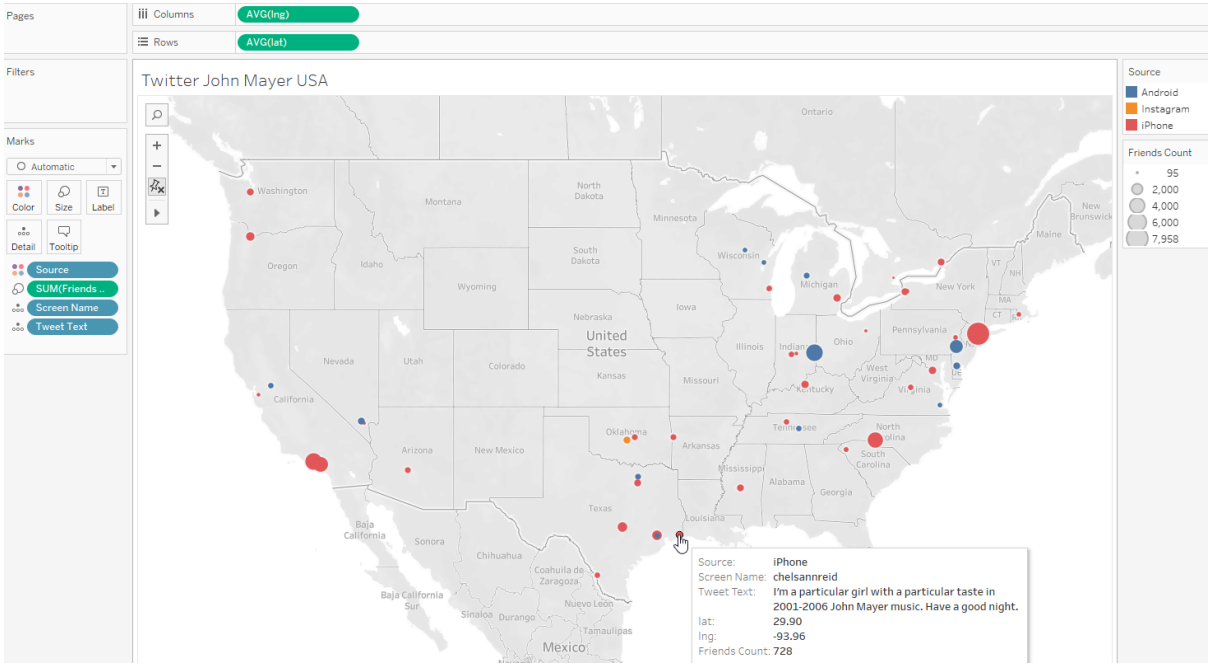


Figure 9: Completed Configuration



The Source field, as per the legend, shows the source of the tweet. We can see this only has three possible values. Source was chosen to use colour mark due to the having a few values and the ease at which this can be communicated visually. Almost all of the 53 tweets are from a mobile device, being either Android or an iPhone. There is one tweet that has a source of Instagram (orange in the lower mid-state of Oklahoma. Most users with a large friend count use iPhone(approx. 6 users). We note there are three iPhone users clustered around the New York area, all with significant friend counts, with only three users across the country using Android that have similar friend count.

Friend count was allocated to a size mark as this value is quantitative and discrete, allowing for a size legend to easily delineate the follower count visually. Users with a high friend count

are mostly centred over densely populated areas, namely Los Angeles on the west coast and New York area on the east coast, another located in Mexico, with one possible outlier located in Indiana – likely close to Indianapolis.

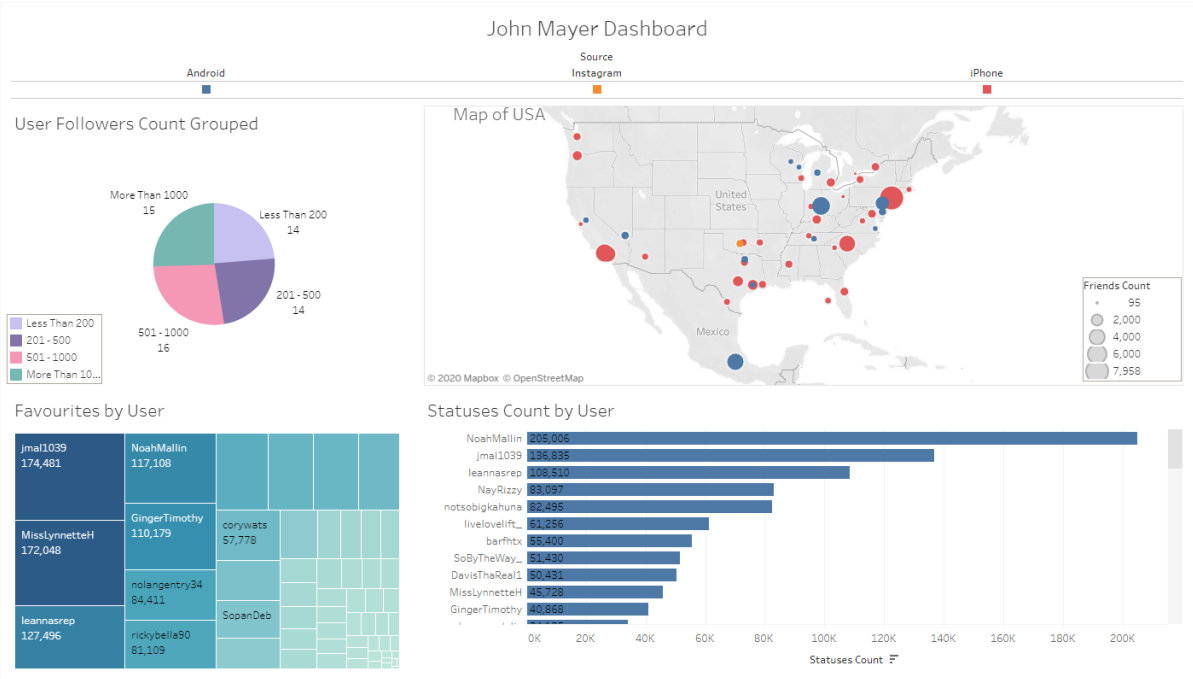
It would be exciting to gather a larger dataset to analyse further using the above plot configuration.

Please view the full interactive visual:
https://public.tableau.com/profile/daniel7447#!/vizhome/JohnMayerTwitter/Map

3.4 Create at least two other charts from your datasets using Tableau and combine them together with your plot from the previous question into a dashboard. Explain the functionality of your dashboard.

Building on our Tableau map of the USA, a bar chart, a tree map and a pie chart have been created and added to a dashboard.
https://public.tableau.com/profile/daniel7447#!/vizhome/JohnMayerTwitter/JohnMayerDashboard

Figure 10: John Mayer Dashboard

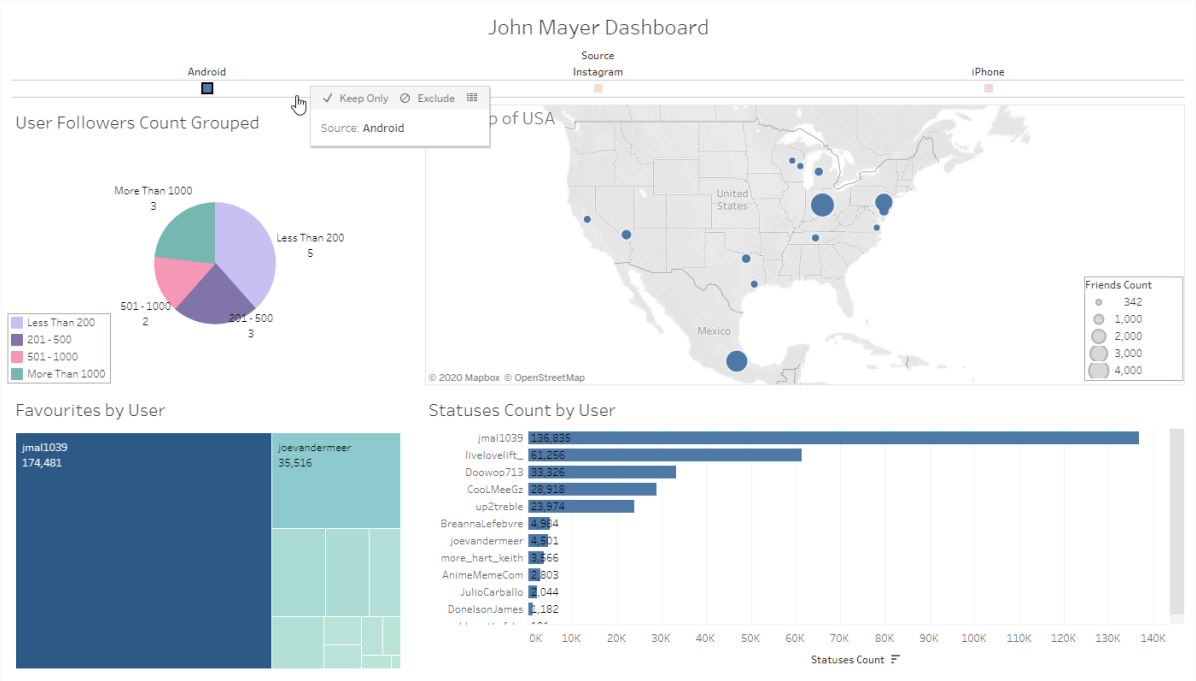


The map of the USA displays the source of the tweet by colour and the friends count for the user using size. The bar chart displays the statuses count by user; the tree map favourites by user; the pie chart is used to group follower counts into buckets.

The dashboard is interactive, allowing us to click elements of each chart to filter, also affecting the other charts.

Firstly, we can click the Source legend across the top to filter all charts by the source system of the tweets, namely Android, Instagram, or iPhone. This will change the Map to show only tweets made using that system, also changing the other charts to reflect the data of only those tweets. This allow us to quickly view the location of those tweets, the follower, friend, favourite, and statuses count for those users.

Figure 11: Android Tweets Only



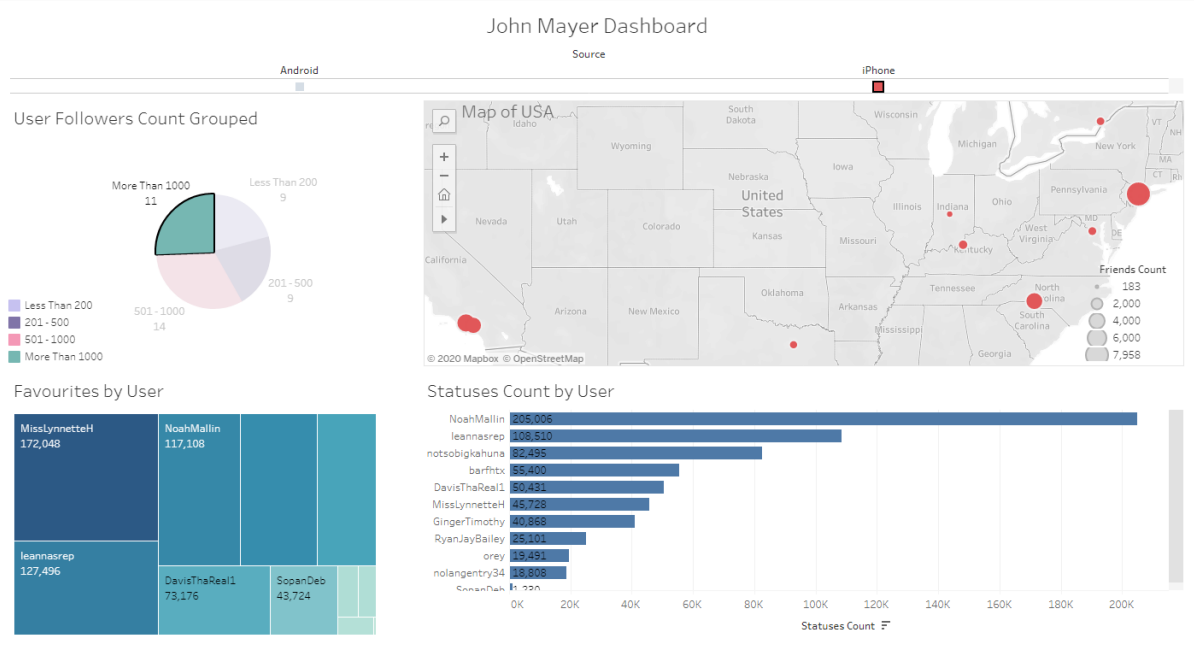
It should be noted that once we have a filter in place on a chart, we need to click the white space within that chart to clear the filter and revert back to all data.

Next, we note the function to click on the pie chart sections to display only the tweets within that followers count group segment. This allows us to quickly analyse users with a large or low followers count.

Furthermore, we can drill into specific users from either the map, by clicking on a specific dot(a tweet), or by selecting the user from either the favourites by user or statuses by user charts. Hovering over the map we can see the specific tweet text along with screen name and friend count for the user. This allows us to quickly identify users with high fried, follower and status count as they are likely more influential that users with lower values.

Lastly, we can combine interactive filter functions for tweet source and user follower count group to get a feel for say, all iPhone users with followers more than 1000. We can quickly see that iPhone source has 11 users more than 1000 followers, with Android only having 3 users with more than 1000.

Figure 12: Combining Source and Follower Count Group



In Conclusion, we have a nice interactive dashboard that can be used to visualize tweets on a map, compare and contrast source, friend, follower, favourite and statuses counts along with drilling into individual twitter users and seeing the individual tweets. This could be used to target specific users or twitter platforms for more strategy investment in gaining more popularity for John Mayer.

Evaluation

3.5 What are the findings of your social media analytics?

Our research was guided by the hypotheses created in the beginning of our study. These are:

- Advertisements help increasing John Mayer visibility.
- More interaction with social media users provides good publicity.
- Collaboration with artists increases John Mayer popularity.
- Covers are a good option to rise John Mayer's reputation.

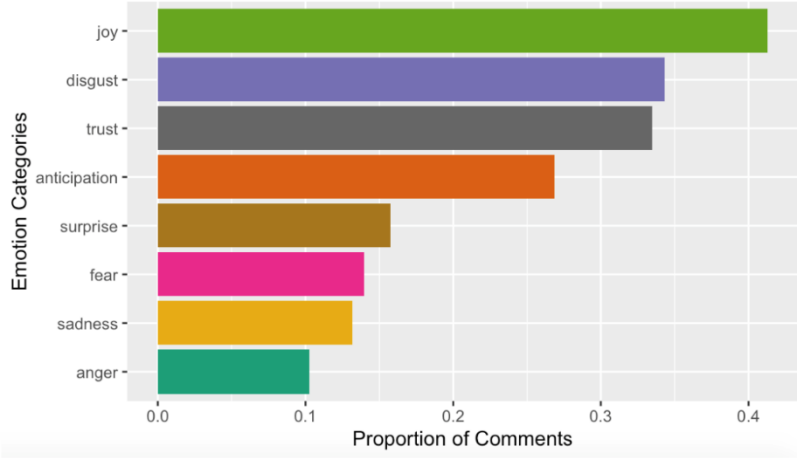
With social media analytics it was possible to test these hypotheses and the findings are detailed as follows:

a. *Advertisements help increasing John Mayer visibility.*

This hypothesis was confirmed. It became evident that the Land Rover add had a very positive impact on John Mayer's visibility. John is promoting Land Rover in conjunction with Atlantic's marketing team (which is a newspaper with approximately 2 million followers) and this user

was evaluated to be very influential within the singer's network, mostly due to posts related to the add. A TF-IDF analysis count confirmed that. Within the top 10 most frequent terms in the Twitter data, the majority were terms connected to the Land Rover add such as "time", "johnmay", "landroveruse", "remind", "moment", "redwood", "sponsor" and "California". Also, sentiment analysis conducted for this advertisement shows that people's reactions to it are generally very positive as per Figure 11.

Figure 13: Emotion Analysis of Land Rover Advertisement with John Mayer



The Land Rover advertisement shows scenes of road trip and nature reserve parks and suggests adventure and enjoyable time outdoors. Therefore, this leads us to conclude that John Mayer's image can have a positive association with these categories and more partnerships can be done within these topics.

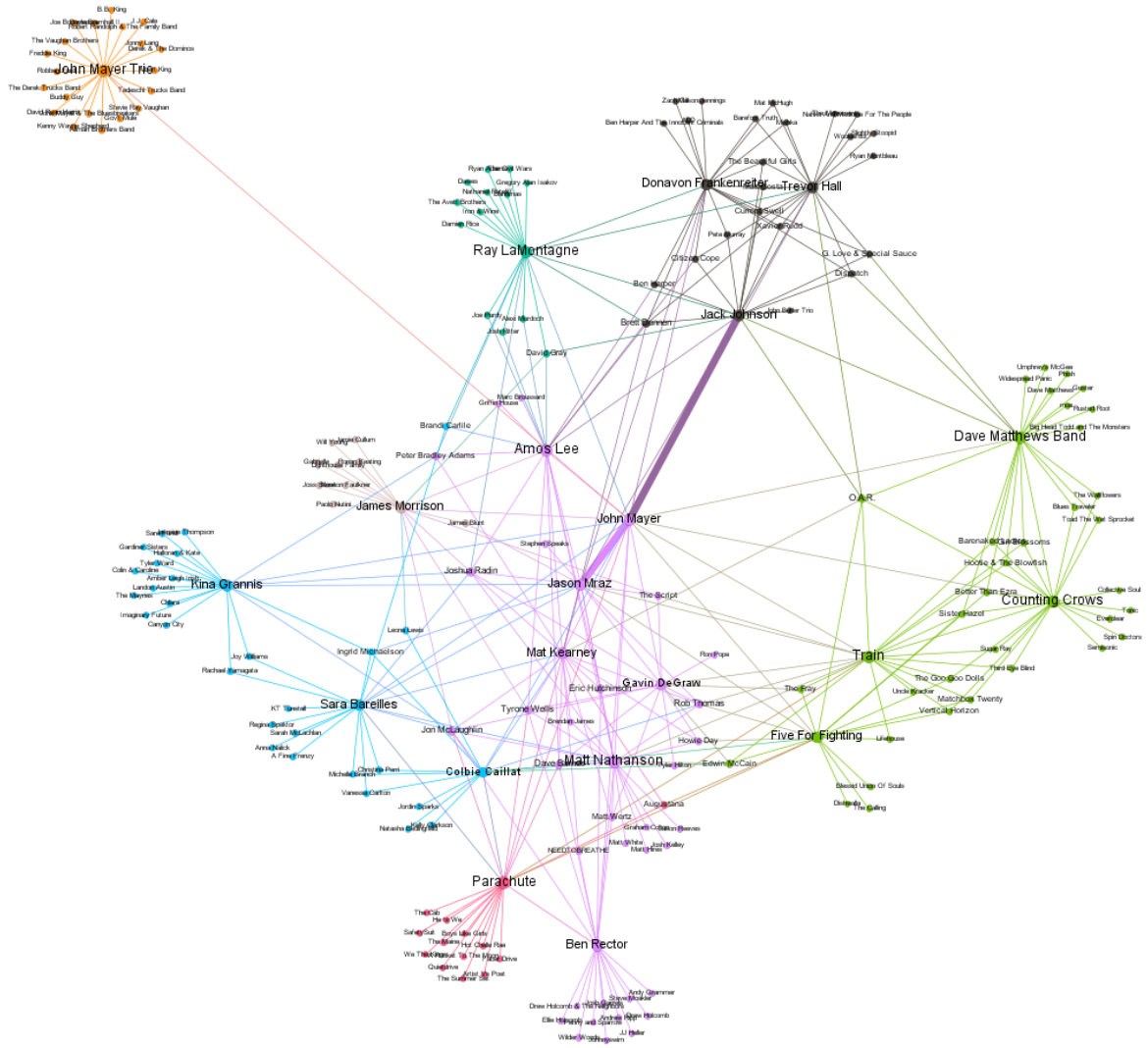
b. More interaction with social media users provides good publicity.

This hypothesis was confirmed. It was clear that John Mayer's fans use more replies and retweets than mentions. As it was observed from the tweet analysis, 35% of the tweet content in our dataset was based on already existing or repropagated tweets. Similarly, replies are also significant, comprising almost 20% of the dataset. This in important to understand the way how the users interact with the content associated with John Mayer and how communication of the content flows throughout the network.

c. Collaboration with artists increases John Mayer popularity.

This hypothesis was confirmed. Collaboration with other artists have contributed to rising John Mayer's profile. Some of the analysis pointed that names such as Taylor Swift and Kanye West were still related to John Mayer, even years after the collaboration. Taylor Swift also shares fans with John Mayer and this could be due to this partnership. Furthermore, it was noted that only two of John's related artists (Taylor Swift and Kanye West) have John as a related artist, which means that there is a high number of other artists that are recommended to a user who likes John Mayer's song, however, the opposite is not true.

Figure 14: Network Graph of John Mayer and Related Artists



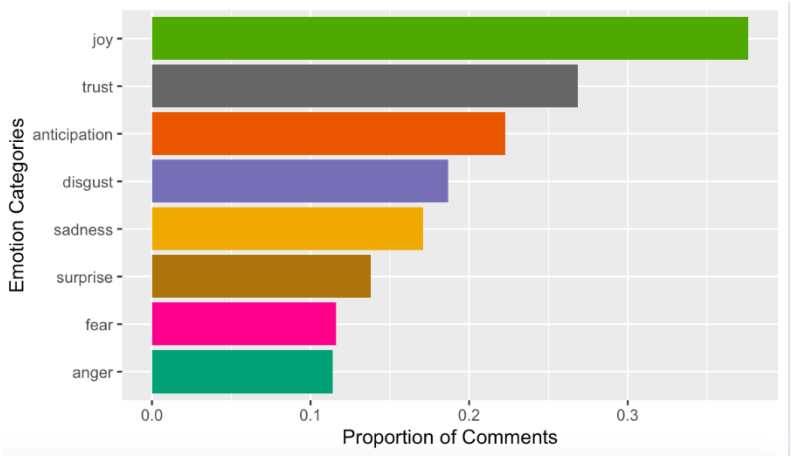
d. Covers are a good option to rise John Mayer's reputation.

This hypothesis was confirmed. Covers are a good strategy to increase John Mayer's reputation. In this study it was analysed "Free Fallin"[2], which is one of the John Mayer's cover.

[2] Free Fallin is a song originally written and recorded by Tom Petty and Jeff Lynne, in 1989, for their album Full Moon Fever.

One of the analysis showed that “Free Fallin” was John Mayer’s most watched video on youtube and it was the video with the second highest number of likes, only after “New Light” (which was considered an outlier). It is also apparent, but sentiment analysis, that it is a video with a good acceptance by the public as per figure below.

Figure 15: Emotion Analysis of “Free Fallin”



Apart from that, there are other interesting findings that were not related to our hypotheses, but which appeared to be relevant to consider when implementing the strategies to raise John Mayer’s profile. For instance, John Mayer’s video clip “New Light” was very much commented for resembling the pandemic period. At the same time, it is evident that (democrat) politics, women empowerment, and issues affecting black people are dear to John Mayer’s top five influential users. This can be due to the proximity to American presidential elections as well as the movements “Me too” and “Black Lives Matter”, respectively.

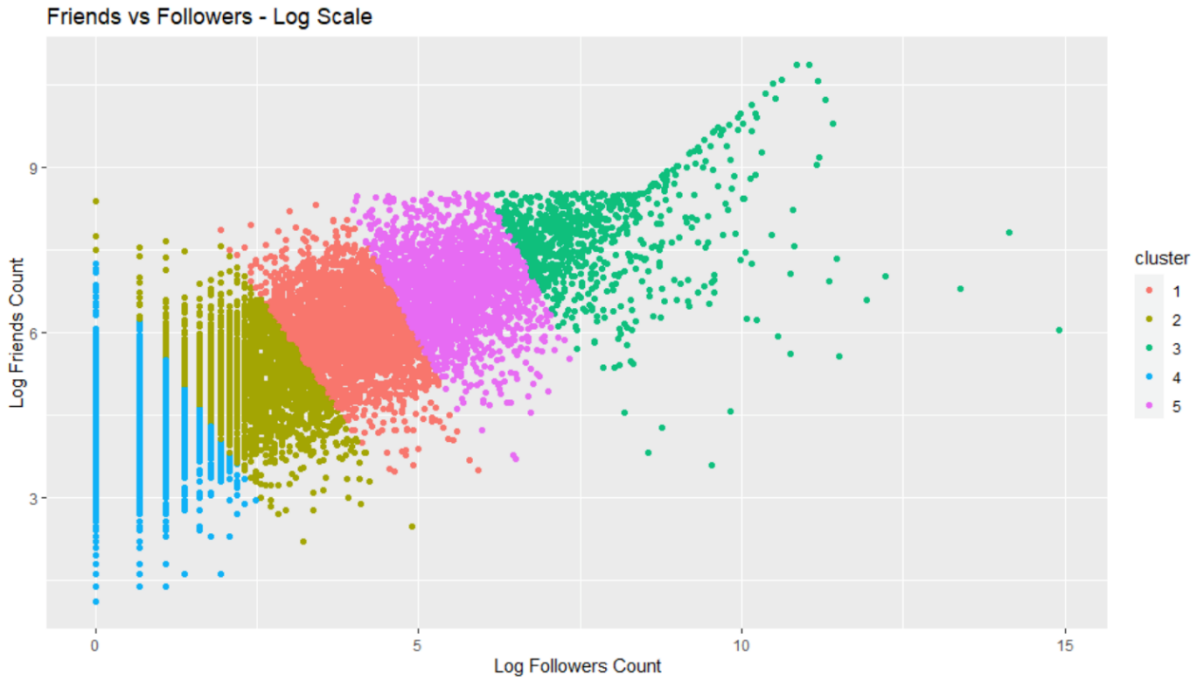
3.6. What actions for improving the popularity of your artist/band do you suggest based on your findings?

Based on our findings it is possible to propose several actions for improving John Mayer’s popularity. However, it is important to not lose sight of his profile as a musician. In one of his interviews, he declared: “I don’t make music for the club, I make music for the omelette on Sunday after the club” (Billboard, 2018). This statement makes it clear that his intention is to play soft and calm music for a relaxing atmosphere. Therefore, our recommendations will be aligned with his personal brand.

First of all, joining other company’s publicity campaign would be highly beneficial for his reputation. However, it is important to take into consideration the topics that were identified as a having a positive association with his profile. These are road trip, nature, adventure and relaxed moments. This also relates to the way how he declares his personal brand and therefore will highlight his personality at the same time that will raise his popularity.

Second, it is possible to say that more posts and replies from John Mayer would give users content on which to retweet and reply and it is likely to contribute to increasing John Mayer's followers count and improve this popularity. The content of his posts could be associated with the hot topics such as (democrat) politics, women empowerment, black movement and the pandemic. Additionally, engaging with influential users, especially those identified as part of the green cluster shown in Figure 14, could make content about John Mayer "go viral", which would also call attention to his music and increase revenues.

Figure 16: TheAtlantic Follower Users: Friends and Followers with 5 Clusters



Third, collaboration with other artists is vital to John Mayer's career and should be intensified. As it was demonstrated previously, past collaborations with Taylor Swift and Kanye West show positive impact on John Mayer's profile until today. Also, in our analysis of related artists, these were the only artists who have John as related artist, with the rest not linking back to John. This means that if a person likes John Mayer's song they will be recommended to other similar singers. However, these same singers will not link their fans back to John Mayer. Therefore, it would be useful to partner with the artists not linking back to John, to create the bidirectional link and grow John's profile as a recommended artist.

Finally, the last recommendation is for him to release cover songs from time to time. This would call attention to his profile by using songs that are well-known and appreciated by a larger audience.

3.7 How could you refine your social media analytics? For example: - Could you use different data sources? - Could you choose different parameters? - Can you think of ways to obtain more relevant data? [1-2 paragraphs, 0.7 mark]

Although Twitter is the most popular social media analysis data collection source, and YouTube and Spotify have much relevant data for the analysis in this course, these sources are, to some extent, limited by languages as well as are focus in specific areas of data sources. Sources such as Instagram and Facebook also provide much rich data in a broader scale. For instance, Facebook (started in 2004) is not as young as Twitter (started in 2006) and has a monthly active user count over two billion with over four million likes per minute, tweeter has over 360 million posts per minute.

Instagram is also a great option for source of social analysis data, the company is much young with only ten years of age, however, it already have reached over one billion monthly active users and over 95 million posts per day. Instagram allows not only social interaction between people, but it goes further with communities' clusters related to eCommerce as well as interactions with Facebook's communities (one is now own by the other and are linked by API integrations). For starters, Instagram is one of the quickest growing platforms in the planet and Twitter, YouTube and Spotify do not even get close in data amount by a long shot. What makes Instagram most interesting is that Instagram posts are often motivated. People post just about everything related to what they love and what people most want in this world.

It is also known that Facebook users shows much more about themselves than typically found in platforms like Twitter or Instagram. There, users more often than not use their real names, as well as profiles that usually includes everything from brands favouritism to cuisine most liked. Business that tries to find data related to new audience demographics and learn more about your consumers. Therefore, Facebook it is also an unbeatable opportunity for targeted social media communities' analysis. Facebook Groups and conversations are very useful are parameters finding niche groups people and their communities. Although, most social media users are likely on Facebook these days the data collected there must be analysed with care as much as any other social network data source. Parameters are very similar to all the mentioned social media platforms. This is a great thing as much of the methodologies used for their analysis can also be used in combination with Facebook and Instagram parameters. Follower count, reactions, impressions, mentions, and hashtags, posts, video views, engagements, page likes as well as post reaching counts.

The Asian social media sources such as Weibo & WeChat have also much relevant data from an eastern perspective. It is also worth mentioning The Little Red Book (specific for fashion and beauty communities) and Tencent networks who own QQ which is a very popular instant messaging platform. These networks provide a great volume of mentions, sentiment, engagements, and reactions. They also have a promising follower growth in the eastern of the globe.

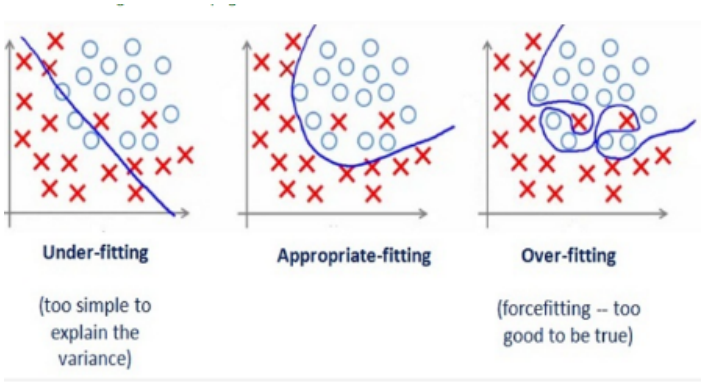
3.8 Research and evaluate other methods/algorithms for network analysis, machine learning models, or visualisation. Compare them to the methods you used in these milestones. Did you find a method that could give you better insights or more promising results for your social media analytics? Explain why you think so.

Support Vector Machine vs Naive Bayes

Although, Naïve Bayes (here forth NB) algorithms was one of the choices in this course development, Support Vector Machine (here forth SVM) could have been used more strongly instead. SVM is also a supervised machine learning model which uses classification algorithms for two group classification challenges. NB's variants and SVMs are quite often used as standard methods for classification of text, however, the performance of these two models much vary based on their model variants, as well as their features chosen, task, and dataset.

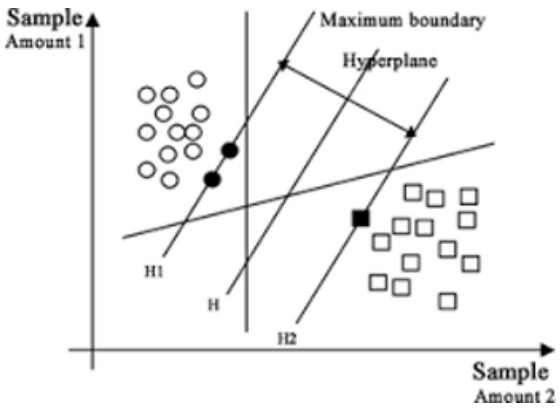
A SVM model establishes of labelled training data for individual categories, after is given, they are clever enough to classify new text as result of its original input. NB extends generative models' classes in order to given classifications. In the NB' models. posterior probability from the class conditional get strengthened. Therefore, the output is a probability of belonging to a particular class. SVM on the other hand is based on a discriminant function ($y = w.x+b$). In this case weights w and predisposition parameter b and training data is used for estimation. The action here is an attempt to discover a hyperplane that boosts the margin, as well it throws optimization functions in this regard.

Figure 17: Figure Classification



When talking about performance, SVMs will definitely have a netter one as using the radial basis function kernel are even more prone to perform better due to the fact, they can cope with non-linearities in the given data. NB performs best as well, however, once the features are independent of one another, this case is unrealistic as just a few happens in real case scenarios. Although, that is the case, NB still performs nicely even when the features are not completely independent. SVM is definitely the choice to go when using a great number of case data (is better at full-length content), but NB is not a bad choice for small case sets (i.e.: is better at snippets). In a nutshell, the combination of both methodologies, seems to be a suitable choice as well as very strong standard for clever classification text data.

Figure 18: Figure Classification

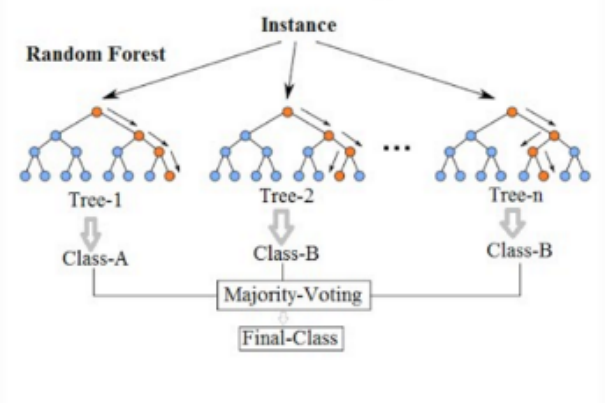


Random Forest vs Decision Trees

One could argue that Random Forest (here forth RF) is better than Decision Tree (here forth DT) and leave at that. Having said that, which is a known fact, the reality is that RF is based on DT and cannot exists without it.

A RF is in its core just a collection of DT. A DT is built on an complete set of data, and uses all the features/variables of interest, on the other hand the RF randomly pick out observations/rows as well as specific features/variables in order to build multiple DT from and averages the output as results. After many DTs are created, each DT point and pick the class. The class that has received points in the majority pick for predicted class. DTs are often comparatively inaccurate, and many other predictors algorithms perform better with same data given. This is when RF come in and help with the issue, as if replacing each decision tree with a RF of DTs. Having said that, RF is not as easy to interpret as an individual DT. RF does shows more promising results for social media analytics against DF. The benefits of social media analytics are great to businesses exposure through social commerce and using RF can scientifically help business to make decision with lass margin or error as in DT due to the large amount of data.

Figure 19: Random Forest Simplified



Gradient Boosting and Decision Trees

Gradient boosting (here forth GB) is a machine learning technique for used for regression and classification case scenarios and typically in the form of a decision tree. It works with the usage of several algorithms such as GBM, XGBoost, LightGBM, and CatBoost. This technique can produce a prediction model in an ensemble of weak prediction models form. Gradient boosting is typically used with decision trees (especially CART trees) of a fixed size as base learners. For this special case, Friedman proposes a modification to gradient boosting method which improves the quality of fit of each base learner.

GB is especially CART trees. Friedman proposes a conversion to GB method which enhances considerably the quality of fit of each base learner in a DF. In short, the combination of GB and DF can considerably increase the quality of results of data processed.

In conclusion, the methodologies and algorithms used in the course have provided much insight and easy capabilities for social network analysis. The combination with mentioned methods' alternatives could improve considerably the output results. For instance, larger dataset with the usage of Random Forest for its analysis as better predictions results in larger datasets when using RF. As the work here shown was mostly related to "John Mayer" which is not as popular as some "bigger" singers and data is not as available. These alternative methodologies could help in linking the missing pieces in a larger scale for better output results.

References

After Dark. (2020). J*ohn Mayer*. Retrieved from: https://afterdark.co/events/london/the-o2/john-mayer--london

Ali, J., Khan, R., Ahmad, N., Maqsood, I. (2012). Random Forests and Decision Trees. *IJCSI International Journal of Computer Science Issues 9*(5), 1694-0814.

Billboard. (2018). *John Mayer's 20 Best Songs: Critic's Picks*. Retrieved from: https://www.billboard.com/articles/columns/rock/8484184/john-mayer-20-best-songs

Deng, H. (2018, October 28). Why random forests outperform decision trees. [Blog post]. Retrieved from: https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5

Friedman, J. H. (2002). *Stochastic Gradient Boosting*. Computational Statistics & Data Analytics, 38(4), 367-378. doi: https://doi.org/10.1016/S0167-9473(01)00065-2

Julianhi. (2014, May 13). Cluster your Twitter Data with R and k-means [Blog post]. Retrieved from https://www.r-bloggers.com/2014/05/cluster-your-twitter-data-with-r-and-k-means/

Kim, J. (2018, May 9). This social media whiz explains WeChat, Weibo, and China's online consumers. [Blog post]. Retrieved from: https://www.techinasia.com/talk/ashley-dudarenok-wechat-weibo-chinese-consumers

Liu, M. (2018, August 7). Analysis of the emerging Chinese social media - The Little Red Book. [Blog post]. Retrieved from: https://towardsdatascience.com/analysis-of-the-emerging-chinese-social-media-the-little-red-book-ad7edd05459e

Mason, L., Baxter, J., Bartlett, P. L., Frean, Marcus (1999). *Boosting Algorithms as Gradient Descent*. In S.A. Solla and T.K. Leen and K. Müller (ed.). Advances in Neural Information Processing Systems 12. MIT Press. pp. 512–518.

Ognyanova, K. (2016). *Network visualization with R*. Paper presented at PolNet 2016 Workshop, St. Louis.

Pew Research Center. (2014). *How we analyzed Twitter social media networks with NodeXL*. Retrieved from: https://www.pewresearch.org/wp-content/uploads/sites/9/2014/02/How-we-analyzed-Twitter-social-media-networks.pdf

R Project. (2020). *Introduction to VosonSML*. Retrieved from: https://cran.r-project.org/web/packages/vosonSML/vignettes/Intro-to-vosonSML.html

Search Engine Journal. (2020). *Instagram Has 1 Billion Monthly Users, Now the Fastest Growing Social Network*. Retrieved from: https://www.searchenginejournal.com/instagram-1-billion-monthly-users-now-fastest-growing-social-network/258127/#close

Song List. (2020). *John Mayer's Songs*. Retrieved from: https://www.song-list.net/johnmayer/songs

Soundigest. (2019). *5 of Our Favorite John Mayer Collaborations to Ever Exist*. Retrieved from https://soundigest.com/2019/02/07/best-john-mayer-collaborations/

The Atlantic. (2020). *John Mayer goes outside.* Retrieved from: https://www.theatlantic.com/sponsored/land-rover-2020/john-mayer-goes-outside/3424/

Tidy Text Mining. (nd). *Topic Modeling*. Retrieved from https://www.tidytextmining.com/topicmodeling.html

Twitter. (2020). *Tweet Location FAQs*. Retrieved from: https://help.twitter.com/en/safety-and-security/tweet-location-settings

Want, Sida, Manning, Christopher. (2012). *Baselines and Bigrams: Simple, Good Sentiment and Topic Classification.* Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 90–94.

Wikipedia. (2020). *Decision Tree*. Retrieved from: https://en.wikipedia.org/wiki/Decision_tree#Advantages_and_disadvantages

Wikipedia. (2020). *Gradient Boosting*. Retrieved from: https://en.wikipedia.org/wiki/Gradient_boosting

Wikipedia. (2020). *John Mayer Discography*. Retrieved from: https://en.wikipedia.org/wiki/John_Mayer_discography#Studio_albums

Wikipedia. (2020). *Random forest*. Retrieved from: https://en.wikipedia.org/wiki/Random_forest

Zubrinic, K., Milicevic, M. and Zakarija, I. (2013). International Jounal of Computers, 7(3). 109-116.