Final Report 7230ICT

Big Data Analytics

Rey Allen Permale
10-4-2024

Table of Contents

Setting **1**

I. Artist Information 2

Data Selection and Exploration **2**

II. Collection of Data 2

III. Top 5 Actors 3

IV. Unique Actors 4

V. Spotify Data 5

Text Pre-processing **9**

VI. Term-document matrices 9

VII. Semantic Network 10

Social Network Analysis **12**

VIII. Centrality Analysis 12

IX. Community Analysis 17

Machine Learning Models **23**

X. Sentiment Analysis 23

XI. Decision Tree 24

XII. Topic Modelling 26

Visualisation **28**

XIII. Dashboard 28

Analysis Review **29**

XIV. Alternative methods 29

Setting

I.Artist Information – Taylor Swift

Taylor Swift start as a country music singer who is in the music industry for 18 years now. She began her career in 2006 with her first album called “Taylor Swift”. Since then she evolved into one of the most prominent figures in music, transitioning from country to pop. Currently, she has been experimenting with indie and alternative genres.

Throughout her career, Taylor Swift has released 11 studio albums ang 4 re-recorded albums (Forbes, 2024). The total number of tracks that she published a total of 274 tracks after the addition of her latest album (Newsweek, 2024).

References

Toni FItzgerald. (2024). *Taylor Swift albums*. Forbes. https://www.forbes.com/sites/entertainment/article/taylor-swift-albums/

Sophie Lloyd. (2024). *How Many Albums Does Taylor Swift Have? Timeline of Song Releases*. *Newsweek*. https://www.newsweek.com/taylor-swift-how-many-albums-songs-tortured-poets-anthology-1892216 -:~:text=On%20Friday%20morning%2C%20Swift%20announced,number%20of%20tracks%20to%20274.

Data Selection and Exploration

II. Collection of Data

Reddit was chosen as the platform for data collection as the sentiments of the redditors are more related to the thread that was chosen compared to the comment sections in Youtube in the search strategy that was applied. The search strategy only chose the top threads related to the most recent 4 albums released by Taylor Swift including the re-released versions. It should also be noted that showbiz related feud and political thread was ignored to only focus on the improvement of the musical aspect of the artist. They key search words that gave the desired results and the chosen links are:

Search Words:

- Taylor Swift Latest Album
- Taylor Swift Midnights
- Taylor Swift Version
- Taylor Swift Album Megathread

Links:

- https://www.reddit.com/r/TaylorSwift/comments/y9ivjo/midnights_megathread/
- https://www.reddit.com/r/TaylorSwift/comments/17he0ir/1989_taylors_version_megathread/
- https://www.reddit.com/r/TaylorSwift/comments/1aj5rqq/the_tortured_poets_department_album_announcement/

- https://www.reddit.com/r/TaylorSwift/comments/139b5nl/speak_now_taylors_version_announcement_megathread/

The number of data points collected are in total of 3366.

rd_data	3366 obs. of 22 variables

Figure 1. Data Points

III. Top 5 Influential Actors

The dataset was removed of any rows that contains a null attribute and was then created into a network graph resulting into these 5 influential actors. The result also shows that people are more interactive to the comments of aran130711, PassionateAsSin and Lyd_Euh. Furthermore, there is a tendency that [deleted] is a false result as the activities of multiple deleted accounts in reddit can fall into the category of the [deleted] account.

```
> pagerank_scores <- sort(page_rank(rd_actor_graph)$vector, decreasing=TRUE)
> pagerank_scores[1:5]
     aran130711 PassionateAsSin         Lyd_Euh       [deleted]   wewerelegends
    0.405470854     0.215944992     0.204012327     0.001432508     0.001408293
```

Figure 2. Top 5 influential actors

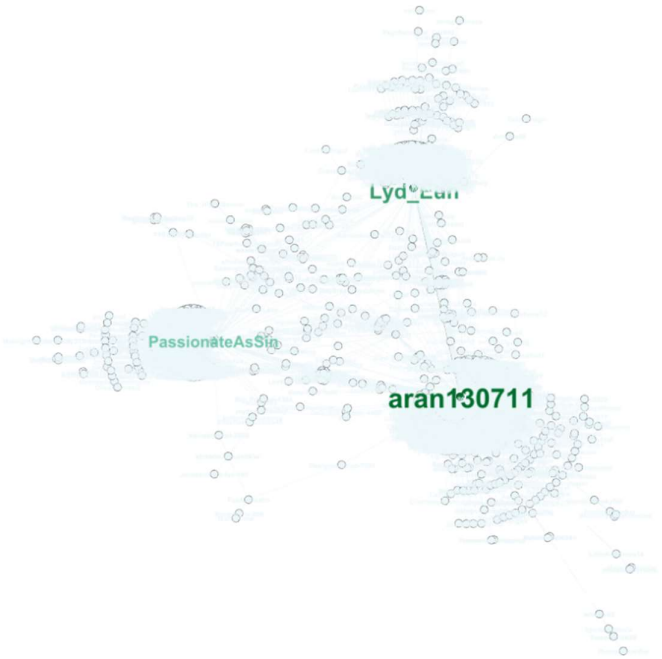


Figure 3. Actor Network Graph visualization

The reason there are only 3 prominent figures in the visualization is because the gap between the pagerank of the top 3 results compared to the top 4 and top 5.

IV. Unique Actors in the dataset.

There are around 1481 unique actors in the actor graph and was obtained by using the length function of R. This will return the number of unique actors which was converted into row in the actor network graph.

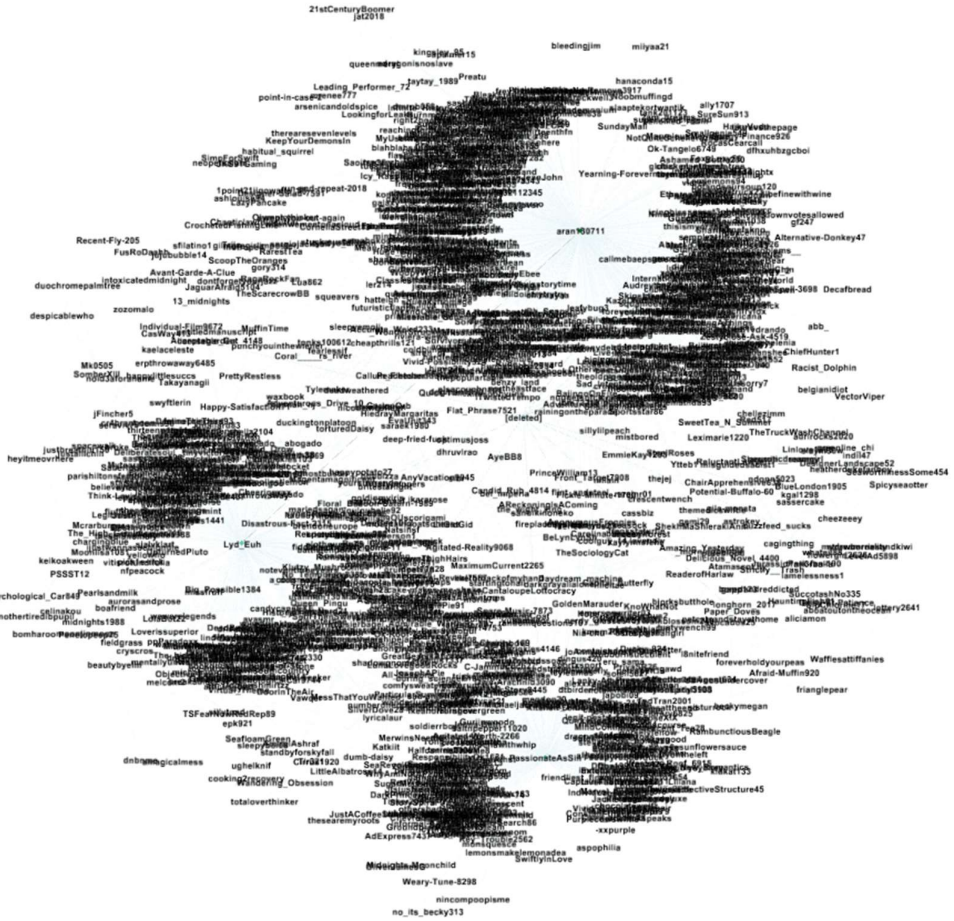


Figure 4. Unique Actors Network Graph visualization

V. Spotify Data

Active Years

The data collection in spotify for Taylor Swift showed that spotify has the oldest album released in 2014 where as in the searched in part 1 there are albums dating older than 2014. This is probably due to an issue back in 2014 which is not related to the case study.

ıages	name	release_date
variable	1989 (Deluxe)	2014-01-01
variable	1989	2014-01-01
variable	reputation Stadium Tour Surprise Song Playlist	2017-11-09
variable	reputation	2017-11-10
variable	Lover	2019-08-23
variable	folklore	2020-07-24
variable	folklore (deluxe version)	2020-08-18
variable	folklore: the long pond studio sessions (from the Disney+ s...	2020-11-25
variable	evermore	2020-12-11
variable	evermore (deluxe version)	2021-01-07
variable	Fearless (Taylor's Version)	2021-04-09
variable	Red (Taylor's Version)	2021-11-12
variable	Midnights	2022-10-21
variable	Midnights (3am Edition)	2022-10-22

Figure 5. Taylor Swift albums in spotify

Albums and Songs Published

The data collection showed a result of 20 albums released in Spotify and 359 songs published by Taylor Swift. These figures were obtained by searching for albums related to Taylor Swift and looping through each album to create a data set of all_tracks.

```
# Retrieve album data of artist

albums <- get_artist_albums("06HL4z0CvFAxyc27GXpf02",
                            include_groups = c("album", "single", "appears_on", "compilation"))
View(albums)

# Initialize variable to collect songs

all_tracks <- data.frame()

# Looping through each albums to get track

for (album_id in albums$id) {
  album_tracks <- get_album_tracks(album_id)
  all_tracks <- bind_rows(all_tracks, album_tracks)
}

View(all_tracks)
```

Figure 6a. album and tracks code

albums	20 obs. of 14 variables
all_tracks	359 obs. of 14 variables

Figure 6n. album and tracks count

The underlying reason for this result is that there are albums which have different versions such as the folklore having a deluxe edition.

folklore
folklore (deluxe version)

Figure 7. folklore album with different versions.

Collaborators

Collaborators in the Spotify data were obtained by unnesting the artist column of each album. The resulting artist was then filtered to remove the "Taylor Swift"artist_name to only display the list of artists that Taylor Swift did a collaboration with. Figure 8a shows the code to unnest and filter the collaborators and figure 8b shows the list of collaborators.

```
# Unnest the `artists` column

collaborators <- all_tracks %>%
  select(id, artists) %>%
  unnest(artists, names_sep = "_")

view(collaborators)

# Remove Taylor Swift in the list of artist

collaborators_cleaned <- collaborators %>%
  filter(artists_name != "Taylor Swift") %>%
  group_by(id) %>%
  summarize(collaborators_list = list(artists_name))

view(collaborators_cleaned)
```

Figure 8a. extracting collaborators

	id	collaborators_list
1	0y6kdSRCVQhSsHSpWvTUm7	Gary Lightbody
2	1wtOxkiel43cVs0Yux5Q4h	Lana Del Rey
3	2OzhQiSqBEmt7hmkYxfT6m	Post Malone
4	2Rk4JlNc2TPmZe2af99d45	c("Brendon Urie", "Panic! At The Disco")
5	2awNGIJHodfLZSCIB3PYhz	The National
6	2x0WlnmfG39ZuDmstI9xfX	c("Ed Sheeran", "Future")
7	3O5osWf1rSoKMwe6E9ZaXP	Bon Iver
8	3RaT22zZsxVYxxKR7TAaYF	HAIM
9	3ZVFcD8Wlw9T9kiGqmJf9F	Florence + The Machine
10	3k7ne7VmH43ZPWxPdvPUgR	The National
11	4ABYxlb92WBIjHu7TIKmmI	Hayley Williams
12	4AYtqFyFbX0Xkc2wtcygTr	The Chicks
13	4e3ZNTAV6PCrdYMUrUIMpQ	Fall Out Boy
14	4pvb0WLRcMtbPGmtejJJ6y	Bon Iver
15	55n9yjI6qqXh5F2mYvUc2y	ZAYN
16	5ExOm0dh4NyRyAdSAO9hyM	Florence + The Machine
17	5S4aYQAJOwJMAamANWIlCO	Bon Iver
18	6ADDIJxxqzM9LMpm78yzQG	Lana Del Rey
19	6Wlq9rqkxrqj5Kls4Kw14H	Bon Iver
20	6dODwocEuGzHAavXqTbwHv	Post Malone
21	6kz2hkq1xQvDqPcUyBkw6f	Bon Iver
22	6uwfVkaOM1mcMkFmSn35ix	HAIM
23	7GA86Uo2jYbj8vlXe2nyWd	Lana Del Rey
24	7HC7R2D8WjXVcUHJyEGjRs	Colbie Caillat
25	7qEUFOVcxRl19tbT68JcYK	Ed Sheeran

Figure 8b. List of Collaborators

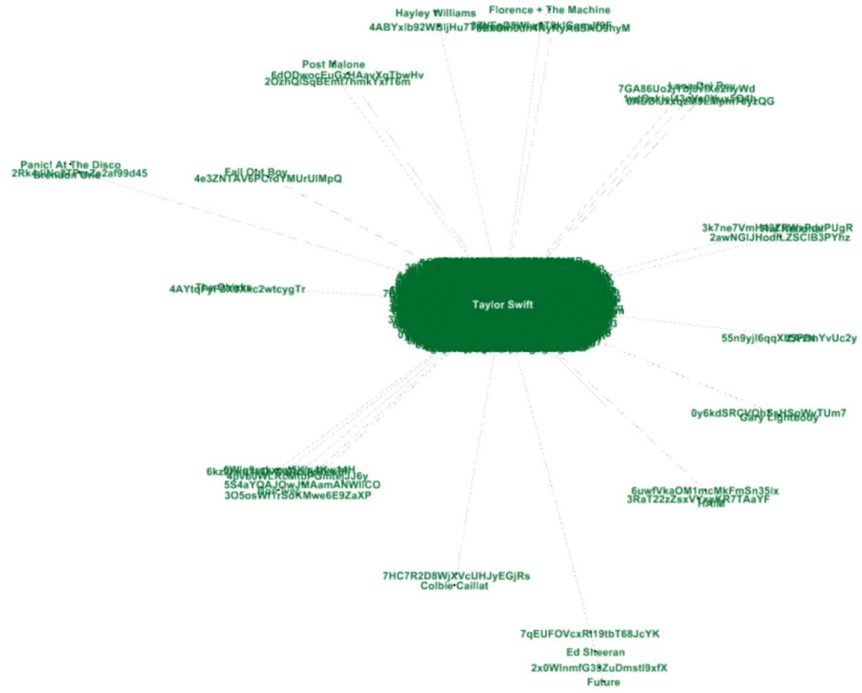


Figure 8c. Collaborators network with track_id

Prevalent Features

The noticeable features in Taylor Swift's tracks are the valence and danceability. These attributes can be compared as they behave in a proportional manner where if the valence is high (which means the song is positive), the more it is danceable.

From the results in these attributes, it can also be inferred that the lesser valued tracks in terms of valence and danceability are those songs that relates to heartbreak or nostalgia. Taylor Swift's evolution to pop music can also be observed as some of her songs has a high danceability and valence.

```
taylor_features <- get_artist_audio_features("Taylor Swift")
View(taylor_features)

data.frame(colnames(taylor_features))

taylor_features_subset <- taylor_features[ , 9:20]
View(taylor_features_subset)
```

Figure 9a. audio features code

danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	track_id
0.648	0.785	7	-5.414	1	0.1650	0.056100	0.00e+00	0.1480	0.943	160.020	1kKlhuQE0HXp1IwBRpaH2P
0.647	0.800	7	-5.384	1	0.1650	0.064700	0.00e+00	0.3340	0.942	160.078	0cqRj7pUJDkTCEsJkx8snD
0.729	0.748	0	-6.670	1	0.0245	0.307000	1.29e-06	0.0962	0.928	99.981	0F2S1QlnMaOfmNHIgmE53k
0.729	0.748	0	-6.670	1	0.0245	0.307000	1.29e-06	0.0900	0.928	99.981	1qLeEu4lXEclkLwoBlMlou
0.689	0.704	9	-10.813	1	0.2450	0.835000	4.83e-06	0.1340	0.920	151.884	6a8aUhYbaQBUI8PcJ5ZmQ6
0.689	0.704	9	-10.813	1	0.2450	0.835000	4.83e-06	0.1340	0.920	151.884	35rdVq36LMHQX0Suw9a6tK
0.632	0.805	7	-5.707	1	0.0690	0.011200	2.51e-05	0.1560	0.903	160.052	50yNTF0Od55qnHLxYsA5Pw
0.632	0.805	7	-5.707	1	0.0690	0.011200	2.51e-05	0.1560	0.903	160.052	3pv7Q5v2dpdefwdWlvE7yH
0.811	0.719	9	-6.553	1	0.0497	0.012900	1.36e-05	0.0742	0.865	103.979	4y5bvROuBDPr5fuwXblBZR
0.575	0.855	9	-4.827	1	0.0467	0.000315	1.61e-03	0.0419	0.840	139.920	7vBlnGdfOLzVEqlOQQxeU8
0.843	0.541	6	-7.361	1	0.0318	0.191000	1.93e-06	0.1080	0.839	116.002	0TYJbIHBtmWy9ugtU9BcbE
0.617	0.699	2	-5.712	1	0.0271	0.113000	0.00e+00	0.0740	0.827	79.979	5zyt5TR2g0l9psX2Z12ex6
0.843	0.543	6	-7.364	1	0.0317	0.174000	3.87e-06	0.0968	0.826	116.000	4fWDgRZa0GKE5KZNhE9MQa
0.843	0.553	6	-7.348	1	0.0317	0.168000	2.66e-06	0.1070	0.825	115.997	3Cl1f4yLwdHzcUnH7o0Xyh
0.658	0.877	7	-2.098	1	0.0323	0.173000	0.00e+00	0.0962	0.821	105.586	4BYejlNgfZF0qKDMEH2cim
0.618	0.687	2	-5.737	1	0.0271	0.111000	0.00e+00	0.0693	0.820	79.991	6WWvcA62QlnRguluWAk6bv
0.695	0.768	6	-4.883	1	0.0344	0.057700	0.00e+00	0.0426	0.813	123.054	5kHMfzgLZP95O9NBy0ku4v
0.570	0.747	4	-3.978	1	0.0426	0.445000	0.00e+00	0.2190	0.808	164.004	5IfvAo8ycf9mLwWr1ZhJj0
0.570	0.747	4	-3.978	1	0.0426	0.445000	0.00e+00	0.2190	0.808	164.004	6yM6QsnTCTVOkKEvg3hGlo
0.788	0.571	6	-6.135	1	0.0296	0.106000	0.00e+00	0.0934	0.797	115.990	550erGcdD9n6PnwxrvYqZT
0.576	0.755	4	-3.982	1	0.0427	0.449000	0.00e+00	0.2230	0.791	163.884	5rcHBt2aNcGFWW0JaAUe1L
0.571	0.547	10	-10.141	1	0.1010	0.704000	0.00e+00	0.0877	0.768	175.933	1SztNGCwEHJEVFx90E5g7D
0.571	0.547	10	-10.141	1	0.1010	0.704000	0.00e+00	0.0877	0.768	175.933	5kiZGSxgqPdv6rbqL9THdd

Figure 9b. audio features result

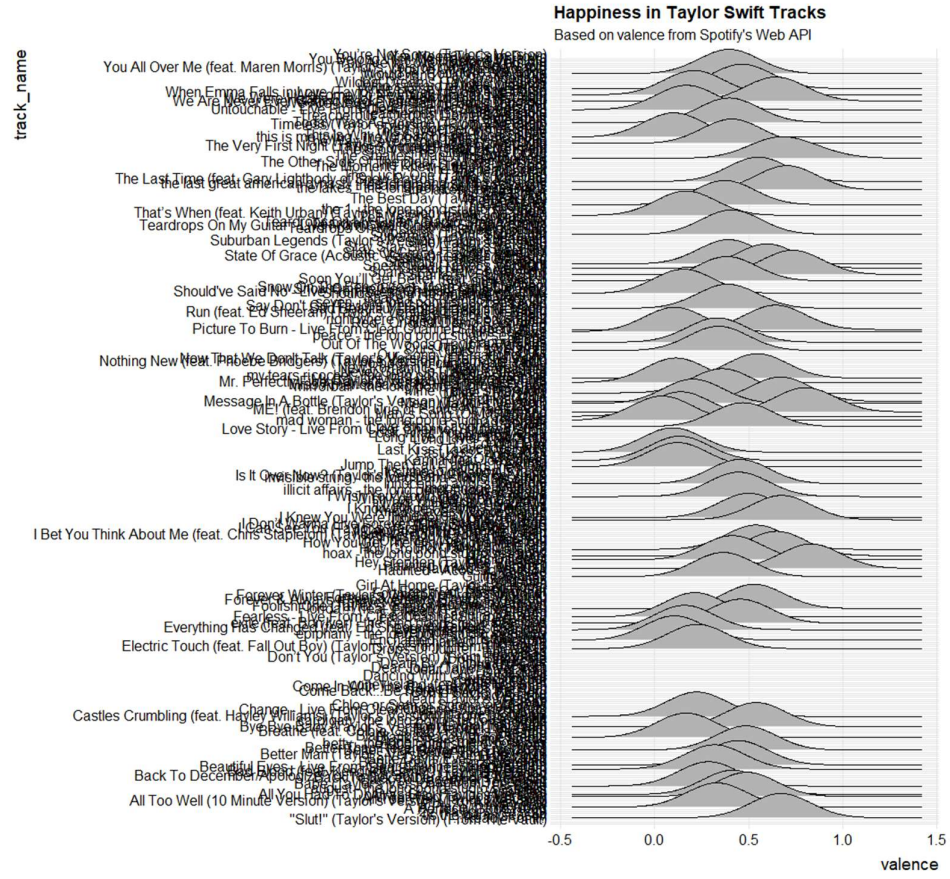


Figure 9c. valence per track graph

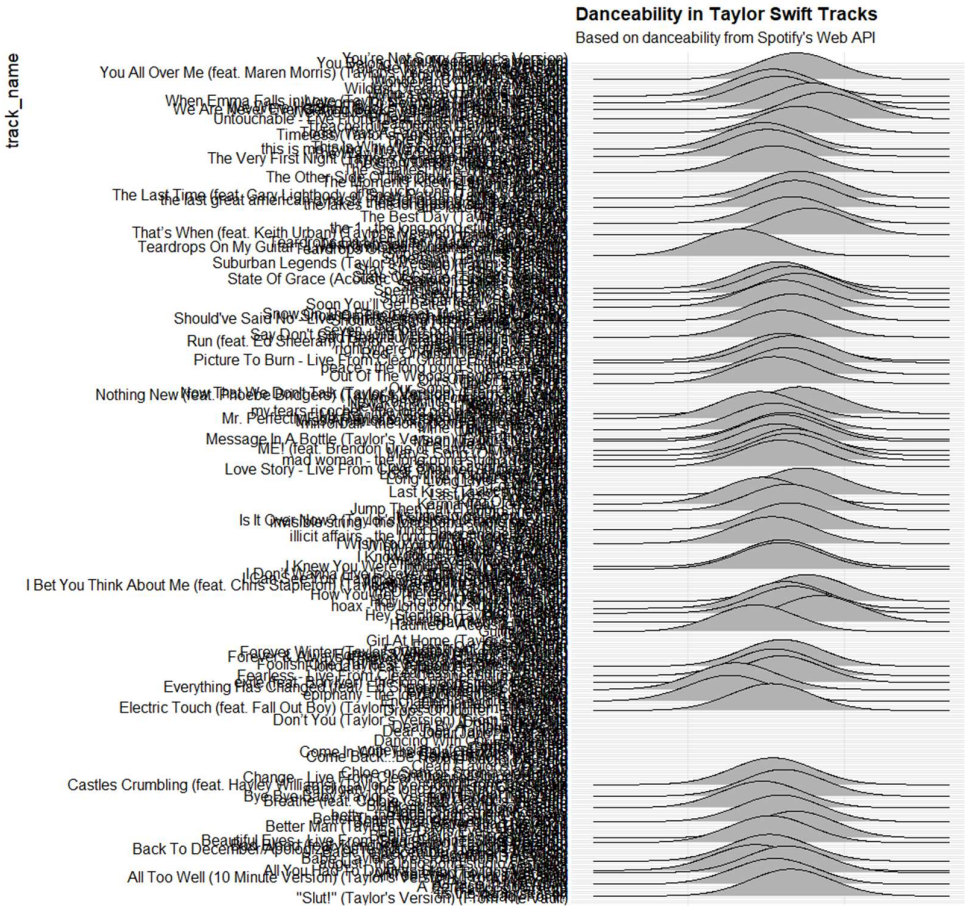


Figure 9d. danceability per track graph

Text Pre-processing

VI. Term-Document Matrix

The top 10 meaningful terms were achieved by removing unnecessary terms, converting each word to stem words, and removing stop words. The results are as follows:

Album – The highest reoccurring word; High probability because the threads search was about taylor swift's most recent album.
Song – One of the main items expected to find in a data collection for a singer. The main target is the sentiment of the users for the songs of Taylor Swift
Taylor – The first name of the artist.
Love – These term could either people use the term "love" to describe their sentiment to the tracks or is about how Taylor Swift's music is mostly about love songs.
Feel – This is expected in a love song. This could be part of the statement where people describe how they are feeling towards the album
Sound – By product of the song, or is used by the users to describe the song sounding like something familiar to them.
Listen – The verb users do in order to feel something from the track.
Track – This is another word for song or recording.
Version – Two of Taylor Swift's latest album are re-released version; hence versions of the album was discussed.

Time – This term could be related to how to users feel after Taylor's re-release of the album which happened years before.

```
> doc_term_matrix <- DocumentTermMatrix(text_corpus)
> dtm_df <- as.data.frame(as.matrix(doc_term_matrix))
> view(dtm_df)
> freq <- sort(colSums(dtm_df), decreasing = TRUE)
> head(freq, n = 10)
  album    song  taylor    love    feel   sound  listen   track version    time
    578     342     288     286     227     200     181     177     159     150
```

Figure 10a. Top terms used

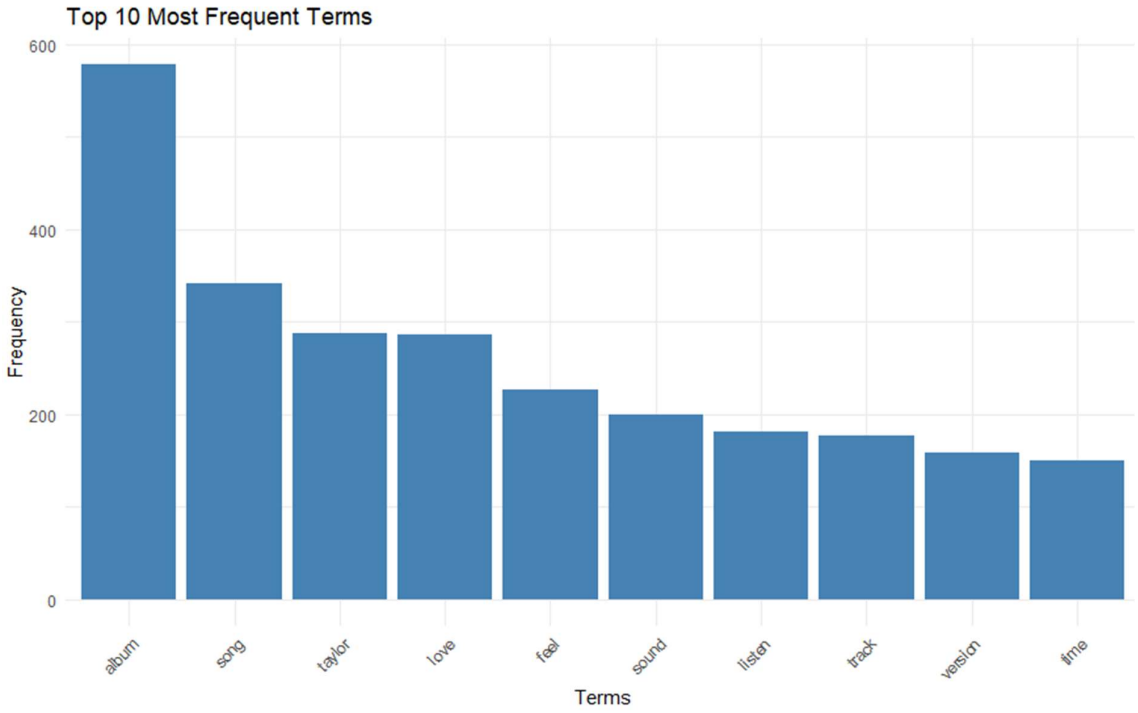


Figure 10b. Visualization for top 10 frequent words

VII. Semantic Networks

By obtaining the bigrams or the consecutive pair of words often appearing are obtained by using the functions in R. It was then sorted according to page rank which returned a result of the top 10 important words from the thread related to Taylor Swift. Those word are as follows:

```
> write_graph(rd_bigram_graph, file = "taylor_bigram.graphml", format = "graphml")
> rank_rd_bigram <- sort(page_rank(rd_bigram_graph)$vector, decreasing=TRUE)
> rank_rd_bigram[1:10]
      album          tv      taylor         pop     version     taylors       songs    favorite        love
0.027266626 0.014800547 0.011404335 0.011087341 0.009723830 0.009165392 0.008272280 0.008184287 0.008021740
    fucking
0.007121599
```

Figure 11a. Top 10 most important words according to page rank

It could be observed that the top term "album" has more then double the occurrence compared to the rest of the terms in the top 10 result. This could be due to the main topic of the subreddit to be the most recent album's of Taylor Swift. There are 2 results that hold the same meaning such as taylor and taylors where taylor could be when the users are talking about the artist "Taylor Swift" while taylors could be the stem word for the terms

taylor's version, as there are two albums related to the thread that are taylor's version release. The rest of the terms are music entertainment related and genre related such as pop, tv, songs and favorite. The last term is the profanity which is "fucking" which are also often used by the users as an expression.

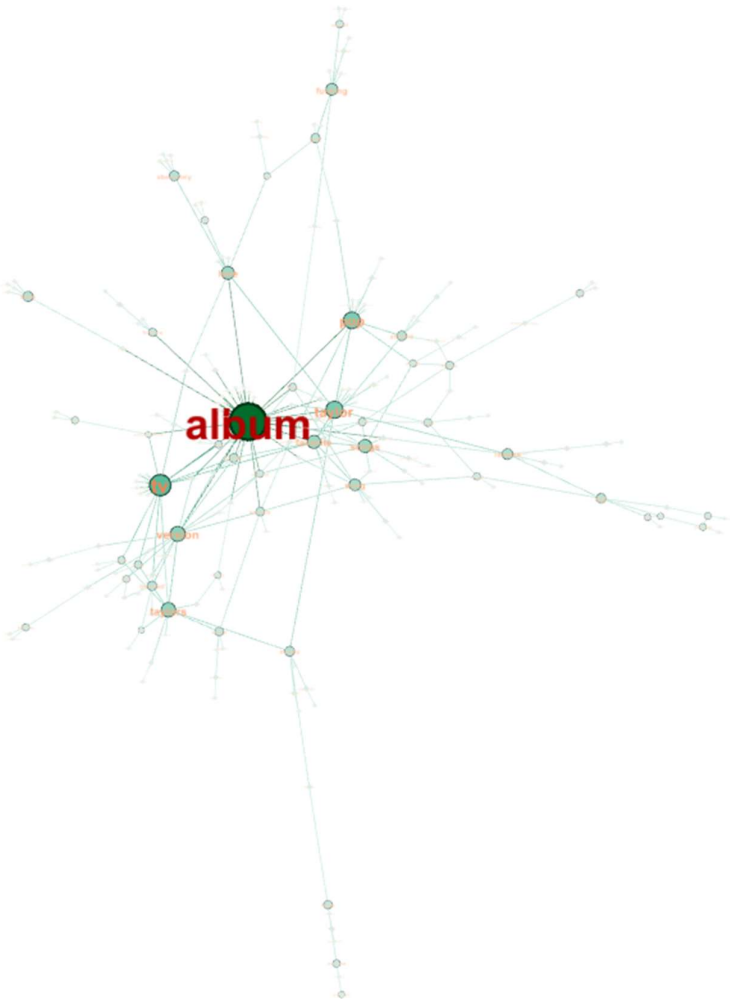


Figure 11b. Top 10 most important words visualization

Social Network Analysis

VIII. Centrality Analysis

The centrality analysis were performed to different datasets belonging to the main artists and two related artists.

A. Taylor Swift

Centrality analysis was performed to the actor network graph previously created to show the scores and determine if it is relevant to the artist. The related artist are obtained by the related artist function of the spotifyr.

Degree Centrality – The degree centrality measures the number of direct connections a user has. In this case, it would show the number of replies or comment the user has given out or received. The first sorting shows the aran130711's thread received the most replies whereas the [deleted] user comment the most. This may also be a false data, as the score [deleted] user received where an accumulation of the multiple deleted accounts that commented in the thread. The total sorted shows the total (in and out) activity the user has done or received.

```
> sort(degree(comp_subgraph, mode = "in"), decreasing = TRUE)[1:10]
      aran130711          Lyd_Euh  PassionateAsSin        [deleted]    wewerelegends Daydream_machine      cagingthing
             882              411              390               19               14                5                5
adragonisnoslave darkgrayallalone      astralrig96
               4                4                4
> sort(degree(comp_subgraph, mode = "out"), decreasing = TRUE)[1:10]
       [deleted]   Callum_Fletcher  Daydream_machine        SundayMail Accurate_Weird233  darkgrayallalone
             124               15               10                9                8                8
  fiddleleaffiggy    vainblossom249     tracyschmosby   coltsmetsfan614
               8                6                5                5
> sort(degree(comp_subgraph, mode = "total"), decreasing = TRUE)[1:10]
      aran130711          Lyd_Euh  PassionateAsSin        [deleted]  Callum_Fletcher    wewerelegends Daydream_machine
             885              414              391              143               17               16               15
darkgrayallalone       SundayMail  fiddleleaffiggy
              12               10                9
```

Figure 12a. Sorted degree centrality results

Closeness Centrality – This measure how each of the users are close to the rest of the discussion. This could mean that user may have reached out to more user via the reply they posted. The values shown in the results shows that most users have the same closeness score which means most users can be reached by other users on an equal level and vice versa.

```
> sort(closeness(comp_subgraph, mode = "in"), decreasing = TRUE)[1:10]
  tracyschmosby coltsmetsfan614     optimusjoss    maddiemaddie2     onebadnightx     JuanJeanJohn     BucketHeadJr
              1                1              1                1                1                1                1
 right2bootlick     bel_imperia      SundayMail
              1                1              1
> sort(closeness(comp_subgraph, mode = "out"), decreasing = TRUE)[1:10]
          Lyd_Euh       lustywench99          kidkat133        Paper_Doves        labrador007 Magentamagnificent
                1                  1                  1                  1                  1                  1
       astralrig96          sillytoad         jazzinitup      SwiftlyInLove
                1                  1                  1                  1
> sort(closeness(comp_subgraph, mode = "total"), decreasing = TRUE)[1:10]
      aran130711          Lyd_Euh        [deleted] Daydream_machine darkgrayallalone  Callum_Fletcher        mistbored
    0.0003684598     0.0003279764     0.0003232062     0.0003194888     0.0003174603     0.0003171583     0.0002883506
 fiddleleaffiggy    _morningbehbs             sibr
    0.0002881014     0.0002875216     0.0002874389
```

Figure 12b. Sorted closeness centrality results

Betweenness Centrality – This would indicate that these users posts are the most often nodes the appear on the shortest paths between other nodes. The result is also a bit reasonable as the degree centrality of aran130711 was the highest, hence this user's reply has the most interaction in the thread. Most of the results in the betweenness could be seen in the previous centrality analysis done.

```
> sort(betweenness(comp_subgraph, directed = FALSE), decreasing = TRUE)[1:10]
      aran130711          Lyd_Euh  PassionateAsSin        [deleted]    wewerelegends Daydream_machine   Touristforlife
      774046.732       448882.914       431847.851       348269.375        17078.948        12970.380         8853.000
darkgrayallalone      cagingthing  Callum_Fletcher
        7622.201         6763.687         5806.443
```

Figure 12c. Sorted betweenness centrality results

Relevance of Results – The relevance of the centrality analysis for Taylor Swift is through the revelation of how the users contribute to the topic of her albums and activities. It can also be shown that some users which in this case are fans of Taylor Swift, are relevant to increase her popularity by engaging with fellow users especially in the digital and information age we are in. The users with high centrality tend to drive the narrative that revolves around Taylor Swift and build different discussions which may in turn affect the sentiments other fans have for Taylo Swift.

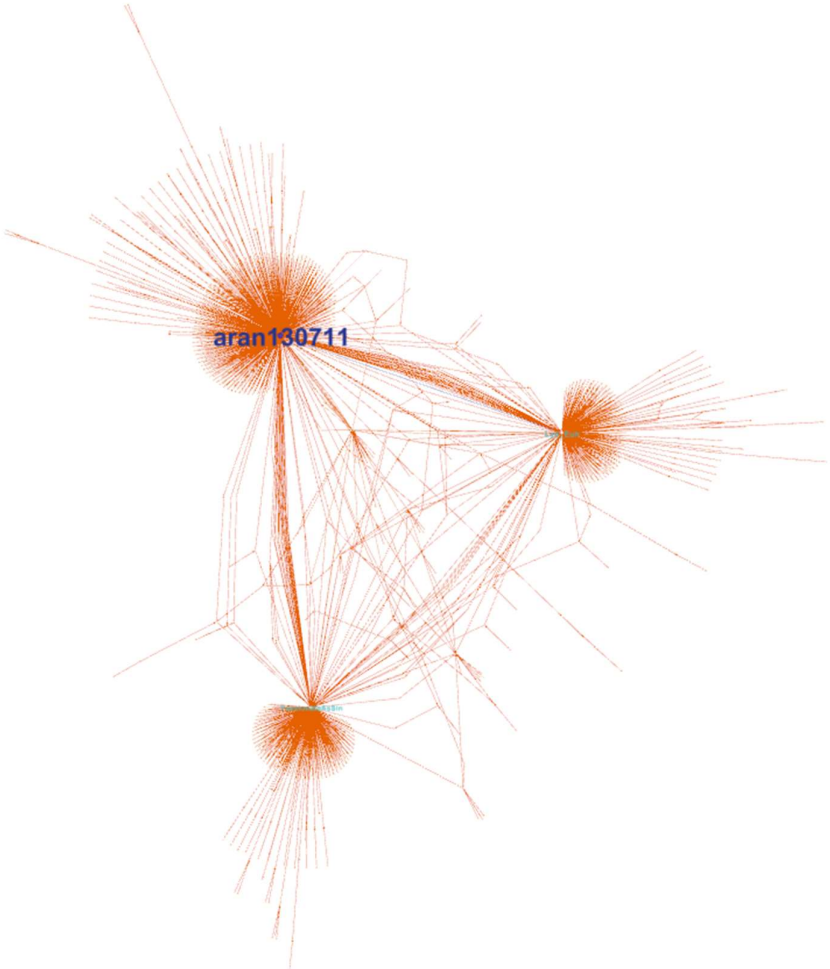


Figure 12d. Actor Graph Network Taylor Swift

B. Olivia Rodrigo Centrality Scores

The component size of Olivia Rodrigo's search threads are up to 659. This would mean that there would be a gap in the centrality scores when comparing Olivia Rodrigo's to Taylor Swift's.

```
> V(network_graph)$name[1:20]
 [1] "LevelAd5898"          "biplane923"           "Competitive-Till5255" "JimmyJizzim"          "Suspicious-Froyo2181"
 [6] "Confident-System-562" "msskmssk"             "simonsmart16"         "spiritsandstories"    "Quick-Sherbert-975"
[11] "Pigsfly13"            "dodieadeux"           "wizzy_da_wizard"      "TopConsideration6834" "charjx"
[16] "CatchingTerror"       "InsomniaticAlien"     "[deleted]"            "New-Importance-3664"  "Western_Quote_3954"
> comps <- components(network_graph, mode = c("weak"))
> comps$no
[1] 1
> comps$csize
[1] 659
```

Figure 12e. Component Size Olivia Rodrigo

Degree Centrality – The fairly noticeable as only one person in this subreddit received a reply of more than a 100, the rest only falls up to a range of 50. This would mean that the subreddit was mostly affected by a single significant person which is "NominalPerson".

```
> # Display top 10 nodes from the sub-graph ordered by degree centrality
>
> sort(degree(comp_subgraph, mode = "in"), decreasing = TRUE)[1:10]
      NominalPerson           [deleted]   No-Restaurant3922          livielouis chalamalabingbong19
                463                  51                  44                  36                  11
Economy_Housing7257         JimmyJizzim         Ygomaster07 Elegant-Thought5170             Amalekii
                  8                   7                   6                   6                   5
> sort(degree(comp_subgraph, mode = "out"), decreasing = TRUE)[1:10]
          [deleted]             Amalekii          Messigoat3          livielouis Wholesome_Turtle078
                 77                  11                  10                   9                   8
     throwawaedawae        charmedroses         Ygomaster07         hillpritch1   No-Restaurant3922
                  8                   7                   7                   7                   7
> sort(degree(comp_subgraph, mode = "total"), decreasing = TRUE)[1:10]
      NominalPerson           [deleted]   No-Restaurant3922          livielouis             Amalekii
                466                 128                  51                  45                  16
chalamalabingbong19          Messigoat3         Ygomaster07         hillpritch1         JimmyJizzim
                 14                  14                  13                  12                  11
```

Figure 12f. Sorted degree centrality results Olivia Rodrigo

Closeness Centrality – The result is still somewhat similar compared to Taylor Swift's result. This is because how the thread system is structured where it is easy for people to connect to another hence the closeness centrality of all users are fairly even.

```
> # Display top 10 nodes from the sub-graph ordered by closeness centrality
>
> sort(closeness(comp_subgraph, mode = "in"), decreasing = TRUE)[1:10]
    wizzy_da_wizard     InsomniaticAlien             zeixble            nzgirl25  Lego_332nd_trooper
                  1                    1                   1                   1                   1
theloyaldogoffenrir          acephoenixx           j_king25           MJustin80           DirtStill
                  1                    1                   1                   1                   1
> sort(closeness(comp_subgraph, mode = "out"), decreasing = TRUE)[1:10]
        LevelAd5898         JimmyJizzim             msskmssk              charjx         frankiestree
                  1                   1                    1                   1                    1
  Lego_332nd_trooper theloyaldogoffenrir          raysworld94 Agile-Huckleberry337           zomgz0mbie
                  1                   1                    1                   1                    1
> sort(closeness(comp_subgraph, mode = "total"), decreasing = TRUE)[1:10]
      NominalPerson           [deleted] AbbreviationsSea5962  chalamalabingbong19         hellboy_2900
       0.0009225092        0.0007002801        0.0006426735        0.0006301197        0.0006281407
    Goodforyouhoney Confident-System-562         Ygomaster07       youve_lost_me  Wholesome_Turtle078
       0.0006281407        0.0006269592        0.0006269592        0.0006265664        0.0006257822
```

Figure 12g. Sorted closeness centrality results Olivia Rodrigo

Betweenness Centrality – With a huge difference in community size, it can be observed in the betweenness centrality results. This is because since there are fewer edges, there is a direct proportionality to the difference in number of routes.

```
> # Display top 10 nodes from the sub-graph ordered by betweenness centrality
>
> sort(betweenness(comp_subgraph, directed = FALSE), decreasing = TRUE)[1:10]
      NominalPerson           [deleted]   No-Restaurant3922          livielouis       Jahidulislame
          199092.299           74848.395           29417.811           23613.131            4740.346
        JimmyJizzim chalamalabingbong19     Goodforyouhoney         Ygomaster07            msskmssk
          3663.481            3608.096            3281.981            2636.069            2622.000
```

Figure 12h. Sorted betweenness centrality results Olivia Rodrigo

the difference in number of routes.

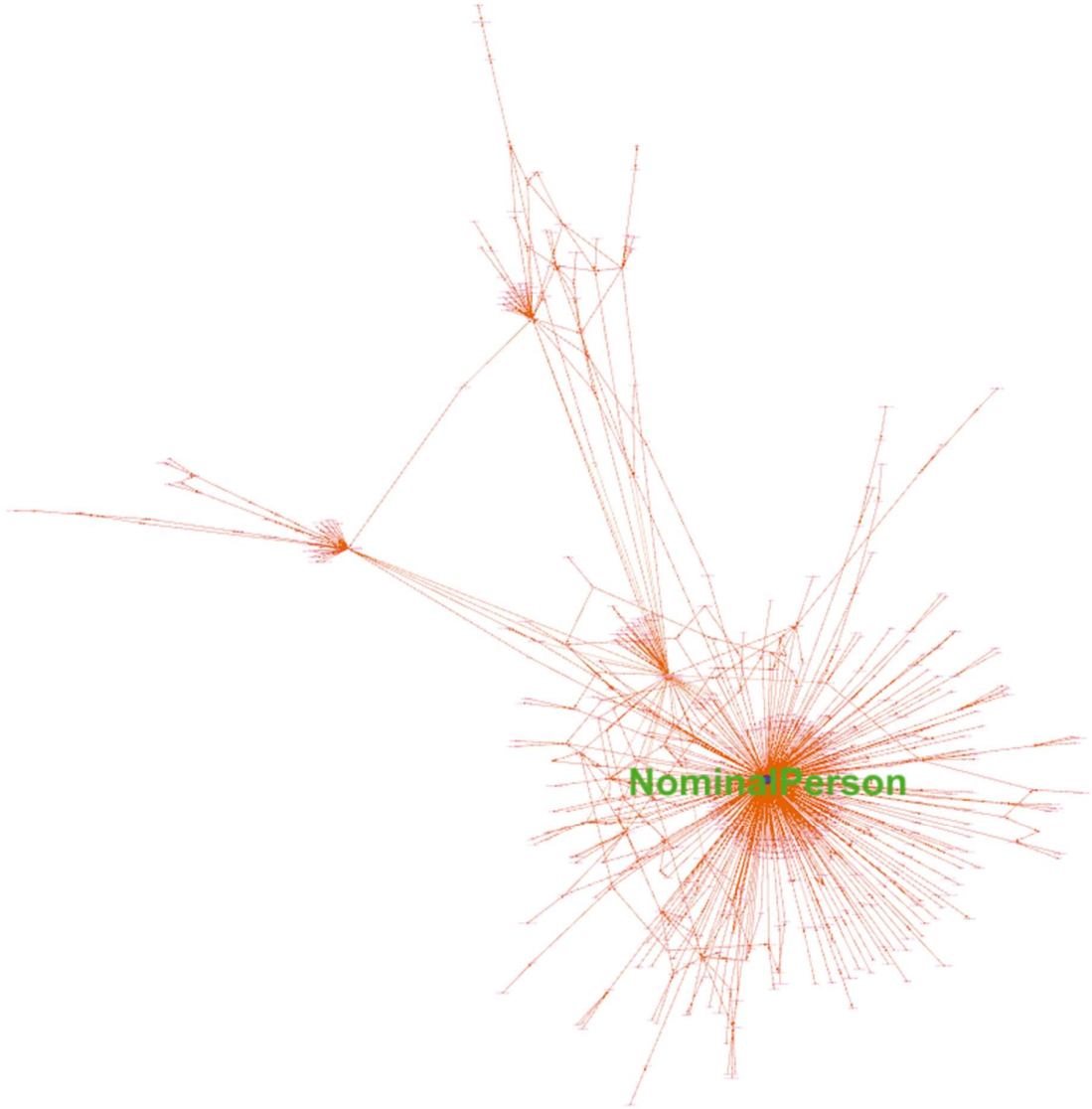


Figure 12i. actor network Olivia Rodrigo

C. Ariana Grande Centrality Scores

Having a component size of only 160, it can be seen the drastic difference between the centrality scores of the actor graph in Taylor Swift's centrality analysis compared to Ariana Grande's analysis.

```
> V(network_graph)$name[1:20]
 [1] "PhotoFreakk"          "lexirobertsx"         "Bright-Hat-6405"      "NotteStellata"        "Jamburger88"
 [6] "[deleted]"            "ConsistentDonkey3909" "levinofriends"        "Cheap_Trifle4524"     "CardiganTSwiftie2005"
[11] "Civil-Two-9572"       "Salt_Meringue_2031"   "Art3misDisney"        "Umbreon---"           "throowowowawaayyyy"
[16] "OkStudent3629"        "mo0nlightangel"       "OfficiAldark"         "theman247g"           "steel_magnolia_med"
> comps <- components(network_graph, mode = c("weak"))
>
> comps$no
[1] 1
> comps$csize
[1] 160
```

Figure 12j. Component Size Ariana Grande

Degree Centrality – The difference in degree centrality compared to Taylor Swift is rather noticeable. Although, it can also be observed that the distribution among the activities per user is more even compared to Taylor Swift's where only 4-5 users received more than 10 replies where as here, it can be seen that all of the top 10 users received replies of more than 15.

```
> sort(degree(comp_subgraph, mode = "in"), decreasing = TRUE)[1:10]
Brilliant_Mushroom_1       mo0nlightangel             filafits            [deleted]       _xmoonlightbae
                 185                   36                   35                   28                   26
    oceanmoonmermaid              lolyana            pyrominic     JamesAlexandra67      Yeet-Meister420
                  24                   24                   24                   20                   16
> sort(degree(comp_subgraph, mode = "out"), decreasing = TRUE)[1:10]
Brilliant_Mushroom_1              lolyana            [deleted]      Yeet-Meister420     oceanmoonmermaid
                  61                   48                   46                   24                   20
      flappy_zachary    Temporary_Act4731            pyrominic      SnooOranges2016            Torttiaaa
                  16                   12                   12                   12                    8
> sort(degree(comp_subgraph, mode = "total"), decreasing = TRUE)[1:10]
Brilliant_Mushroom_1            [deleted]              lolyana     oceanmoonmermaid      Yeet-Meister420
                 246                   74                   72                   44                   40
      mo0nlightangel             filafits            pyrominic       _xmoonlightbae     JamesAlexandra67
                  39                   37                   36                   27                   22
```

Figure 12k. Sorted degree centrality results Ariana Grande

Closeness Centrality – The result is similar with Taylor Swift's results. This could mean that the structure of the thread system of Reddit also plays a role on how the users are close to other users.

```
> sort(closeness(comp_subgraph, mode = "in"), decreasing = TRUE)[1:10]
        PhotoFreakk CardiganTSwiftie2005   Salt_Meringue_2031 gettingcarriedaway86       Kooky-Buy-8462
                  1                    1                    1                    1                    1
             svnboo          Roachman420         peach-prince  comeformecuzimright      -1-throwaway-1-
                  1                    1                    1                    1                    1
> sort(closeness(comp_subgraph, mode = "out"), decreasing = TRUE)[1:10]
        kittycattss     JamesAlexandra67         Aviana_Panns              7_rings-        Equal-Medical
                  1                    1                    1                    1                    1
          Heir_Kia Illustrious_Pick_733         spiceboy2109            Albiiii294         ezravinyl002
                  1                    1                    1                    1                    1
> sort(closeness(comp_subgraph, mode = "total"), decreasing = TRUE)[1:10]
          [deleted] Brilliant_Mushroom_1       mo0nlightangel             filafits     JamesAlexandra67
        0.003039514          0.002457002          0.002331002          0.002331002          0.002320186
      _xmoonlightbae   Rocket_Raccoon_XXX            tore_laps                tn_nt        Tou-lip_louse
        0.002262443          0.002183406          0.002183406          0.002155172          0.002061856
```

Figure 12l. Sorted closeness centrality results Ariana Grande

Betweenness Centrality – With a huge difference in community size, it can be observed in the betweenness centrality results. This is because since there are fewer edges, there is a direct proportionality to

```
sort(betweenness(comp_subgraph, directed = FALSE), decreasing = TRUE)[1:10]
      [deleted] Brilliant_Mushroom_1            filafits       mo0nlightangel     JamesAlexandra67
       10083.333            5520.816            4045.667            4031.600            3689.400
_xmoonlightbae       MajorOctofuss     rabnabombshell             Sahith17     oceanmoonmermaid
       3244.000            774.000             471.000             471.000             314.000
```

Figure 12m. Sorted degree centrality results Ariana Grande

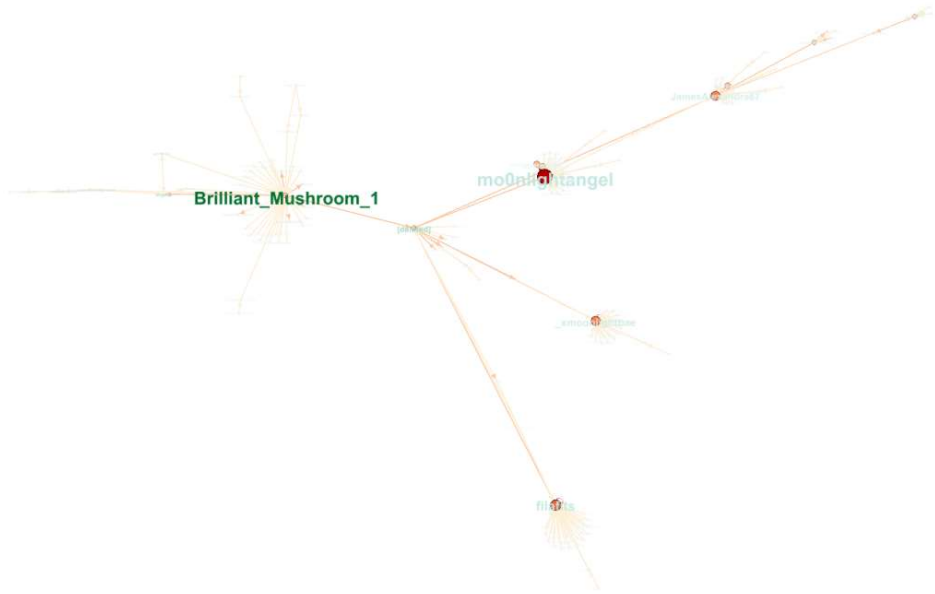


Figure 12n. actor network Ariana Grande

IX. Girvan-Newman and Louvain methods

A. Taylor Swift

Louvain Algorithm – The modularity-based algorithm that helps us identifies communities which have dense links within the clusters. The relevance would be knowing which communities in the data sets are more active in this case Is the community number 7. Furthermore, we could also identify the user groups, in the community and see how their conversation develop to further promote Taylor Swift's popularity.

```
> undir_network_graph <- as.undirected(network_graph, mode = "collapse")
> louvain_comm <- cluster_louvain(undir_network_graph)
> sizes(louvain_comm)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
345   3   7   4   2   4 531   4  56   2   2   7   2   5   5   4   2   2   2   5   2   5   5   2   2 339   4   2   2
 30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
  2   2   2   2   3   2   2   3   2   2   2   3  36   2   4   2   2   2   2   5   3   2   3   3   4   3   3   3  12
 59  60
  3   3
```

Figure 13a. Louvain algorithm result.

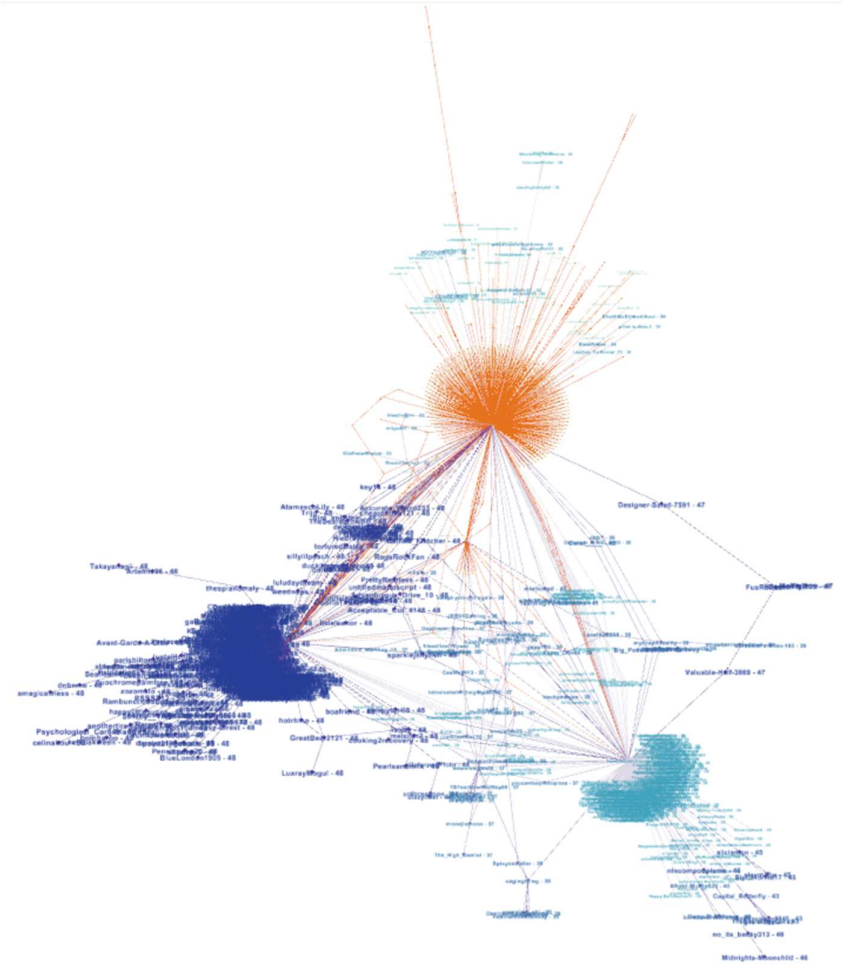


Figure 13b. Louvain modularity class visualization

Girvan-Newman – This algorithm has the same goal but has a different approach where it progressively removes edges with the highest betweness centrality, to split the network to smaller subgroups hence identifying the communities. This algorithm has a relatively similar result with the Louvain algorithm only having a difference on a minimal amount of users in a community.

```
> eb_comm <- cluster_edge_betweenness(undir_network_graph)
> sizes(eb_comm)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
350   3   3   2   2   4 533   4  26   2   2   7   2   5   5   4   2   2   2   5   2   5   5   2   2 349   3   2   2
 30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
  2   2   2   2   3   2   2   3   2   2   2   3  36   2   4   4   2   2   2   2   5   3   2   8   3   5   3   3   4
 59  60  61  62  63  64  65
  3   3   3   8   4   3   3
```

Figure 13c. girvan-newman algorithm result.

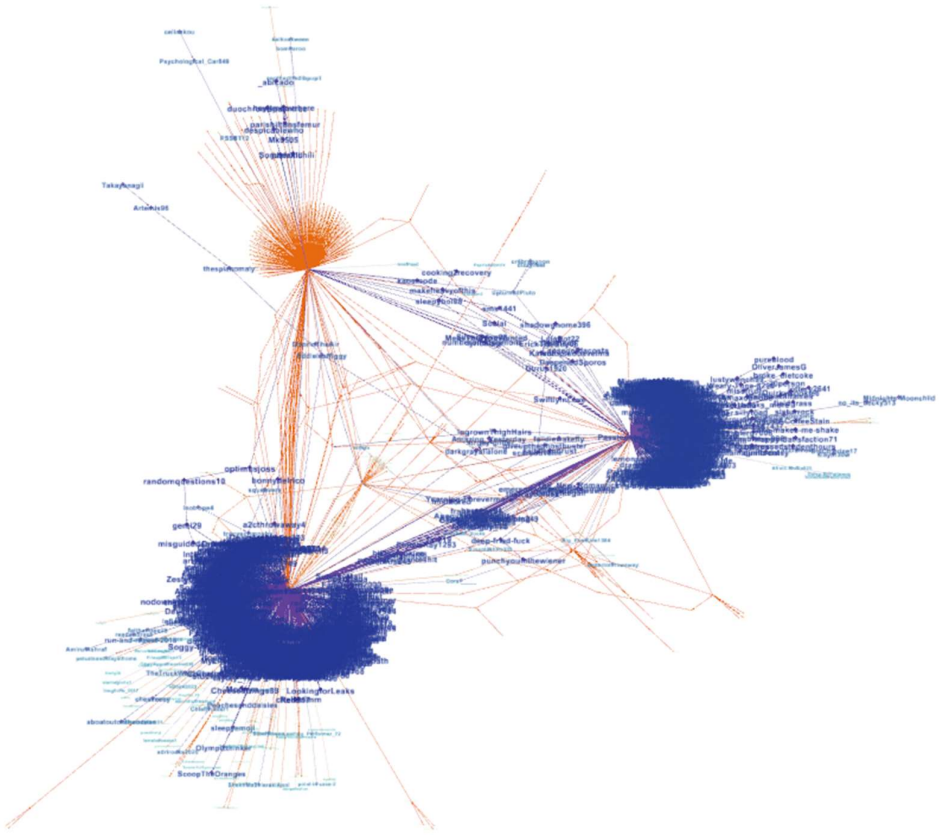


Figure 13d. girvan-newman algorithm visualization.

Relevance – Overall these algorithms are important in order to have a sub-topic exploration and also observe user engagement patterns. Both of these aspects will be a vital information to find a way to further promote Taylor Swift's positive semantic response by the public,

B. Olivia Rodrigo

In general, Olivia Rodrigo's threads demonstrated a different structure in community analysis, as the distribution of activity only falls mostly in one community comparing to Taylor Swift's 3 major communities.

Louvain Algorithm – The community size in Olivia Rodrigo compared to Taylor swift is far lesser and there is only one community that reached more than a hundred of interactions among users.

```
> undir_network_graph <- as.undirected(network_graph, mode = "collapse")
> louvain_comm <- cluster_louvain(undir_network_graph)
> sizes(louvain_comm)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
  3  16   5 210   5  40   7   5   2   2   2   2   2   3   6  18   2   2   2   3   2   5   2   2   2   3   2   4   2
 30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
  2   2  50  16   2   2   2   4   4   2   2   2   2   2   2  62  49  28   2   6   8   2   6   2   3   2   2   3   3
 59  60  61  62  63  64  65  66  67  68  69
  3   2   3   2   2   2   2   2   2   2   2
```

Figure 13e. louvain algorithm results olivia rodrigo.

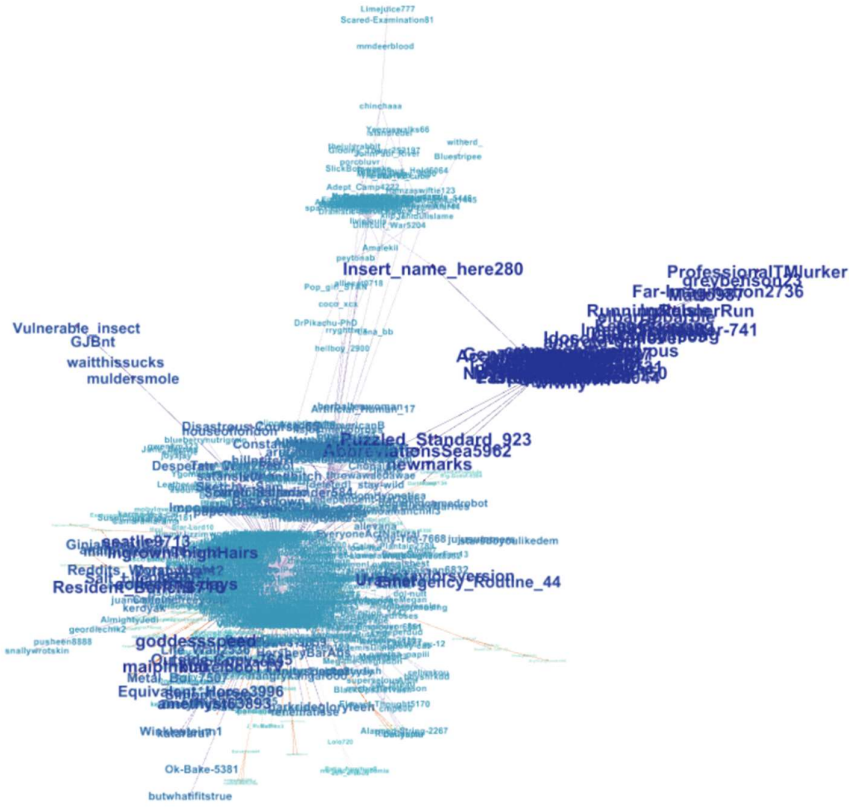


Figure 13f. louvain algorithm visualization olivia rodrigo.

Girvan-Newman – This algorithm shared a similar result with it's Louvain counter part. Comparing to Taylor Swift's results, given that there are fewer users in the threads, it is also expected that these algorithm will show a far lesser results.

```
> eb_comm <- cluster_edge_betweenness(undir_network_graph)
> sizes(eb_comm)
Community sizes
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
  3  16   5 210   5  44   7   5   2   2   2   2   2   3   6   5   6  13   2   2   2   3   2   5   2   2   2   3   2
 30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
  4   2   2   2  50   2   2   2   4   4   2   2   2   2   2   2  62  12  27   9  16   2   6   8  11   4   2   6   2
 59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75
  3   2   2   3   3   3   2   3   2   2   4   2   2   2   2   2   2
```

Figure 13g. girvan-newman algorithm results olivia rodrigo.

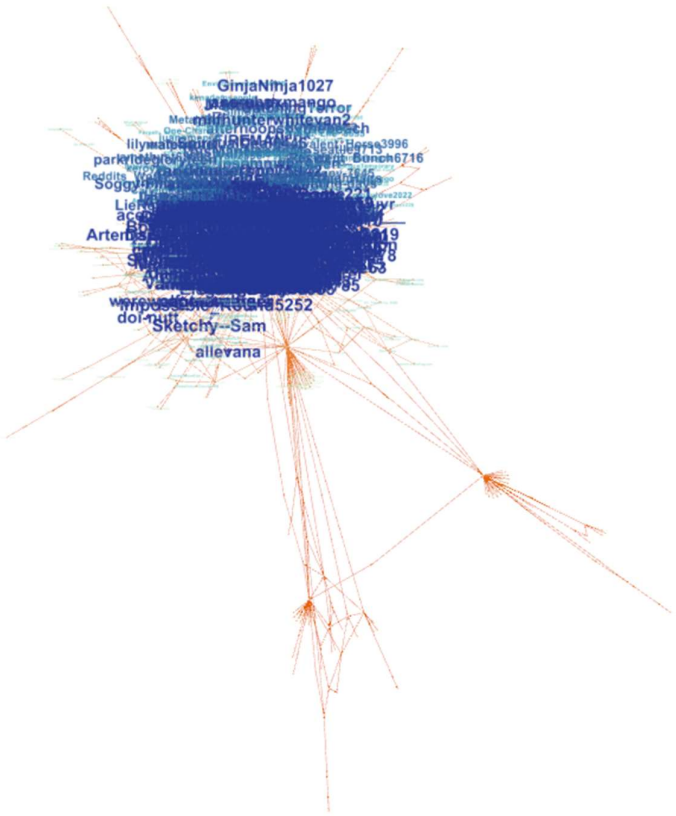


Figure 13h. girvan-newman algorithm visualization olivia rodrigo.

C. Ariana Grande

In general, the huge difference of users active in Ariana's threads compared to Taylor Swift resulted in a far lesser community generated. It can also be observed that in the communities are fairly distributed among each users.

Louvain Algorithm

```
> sizes(louvain_comm)
Community sizes
 1  2  3  4  5  6  7
27 11 23  6 41 29 23
```

Figure 13i. louvain algorithm results ariana grande.

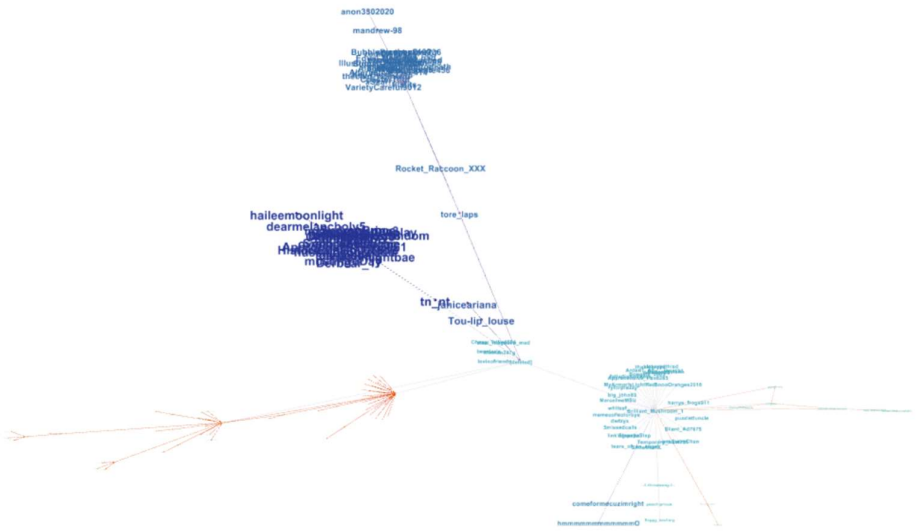


Figure 13k. louvain algorithm visualization ariana grande.

Girvan-Newman

```
> sizes(eb_comm)
Community sizes
 1  2  3  4  5  6  7
28 10 22  6 41 29 24
```

Figure 13k. girvan-newman algorithm results ariana grande.

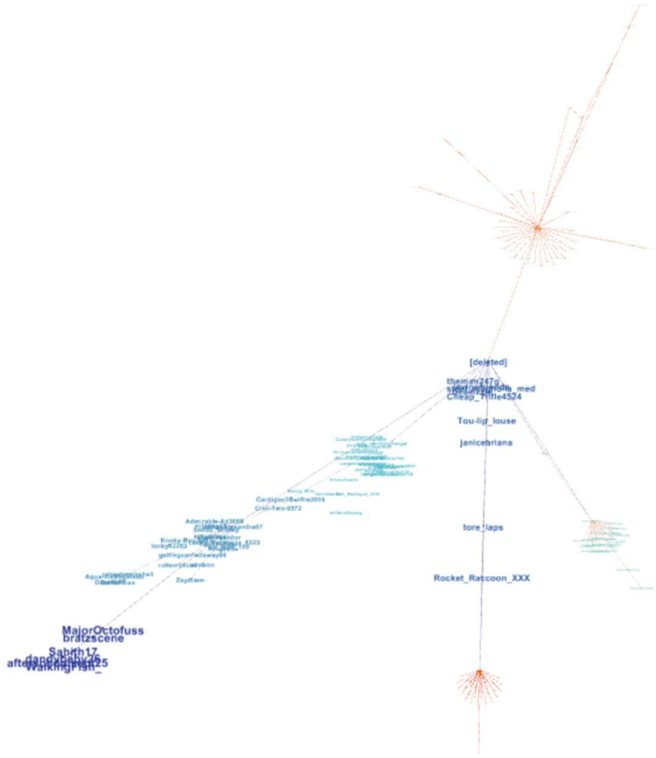


Figure 13l. girvan-newman algorithm visualization ariana grande.

Machine Learning Models

X. Sentiment Analysis

On a high-level analysis of the data collected from reddit. It can be observed that mostly are just categorized on a neutral level. This was achieved by categorizing the phrases into sentiment labels which are Positive, Neutral, and Negative using the function below:

```
sentiment_df$sentiment <- factor(sentiment_df$sentiment, levels = c(1, 0, -1),
                                 labels = c("Positive", "Neutral", "Negative"))
view(sentiment_df)
```

Figure 14a. sentiment labels

	text	sentiment
1	Midnights Theories and Easter Eggs Megathread	Neutral
2	I wish the lyrics were preloaded on Spotify	Positive
3	God this takes me back to buying a CD and laying on my be...	Positive
4	I am looking the lyrics up on Genius	Neutral
5	*NA*	Neutral
6	Ironic because it was spotify who was doing the lyrics billbo...	Negative
7	*NA*	Neutral
8	*NA*	Neutral
9	It is me hi I am the problem it is me	Negative
10	She changed her bio to this and I love it	Positive
11	can not stop thinking about that one TikTok sound Am I the...	Negative
12	*NA*	Neutral
13	*NA*	Neutral
14	When my depression works the graveyard shift all of the pe...	Neutral
15	*NA*	Neutral
16	Lordes album Pure Heroine defined my teenage years I am ...	Negative

Figure 14b. sample data set

Using the ggplot function in R, a bar graph was produce for easier comparison among the sentiments of the redditors. According to the obtained data, the redditors were mostly anticipating or happy about the newly released albums. The rest of the emotions in the visualization portrays the genre of the album Taylor Swift released.

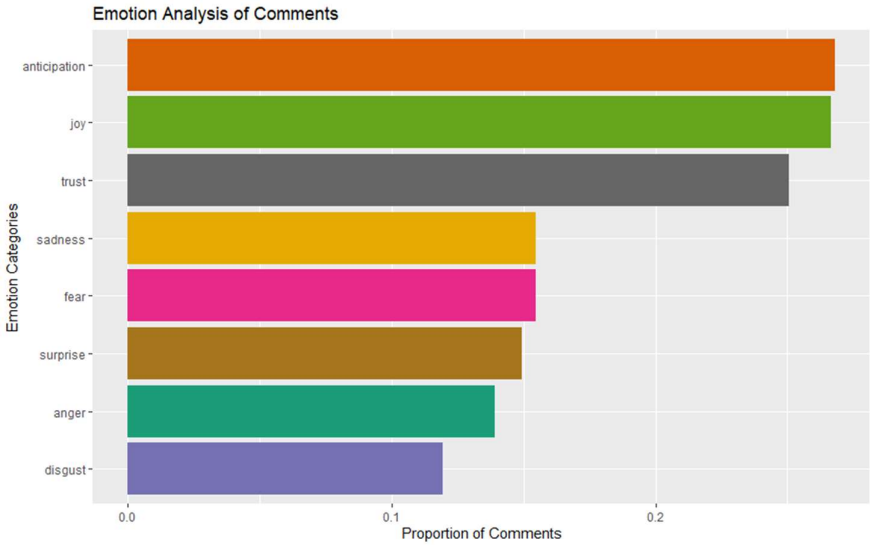


Figure 14c. visualization of sentiment analysis

XI. Decision Tree

Using a machine learning algorithm used for both classification and regression task, we are going to predict the likelihood of being a Taylor Swift song against 2 playlist from spotify "Daily Top 50" and "Daily Mix". This would be interesting as Daily Top 50 are the top 50 popular songs of the world, whereas the Daily Mix is a personalized one.

The process is done by creating a binary attribute in both dataset (isTaylor). Value 0 indicates that it is not a Taylor Swift song and value 1 indicates that it is. Both of these data will then be combined by music of Taylor to be utilized when we are training the data.

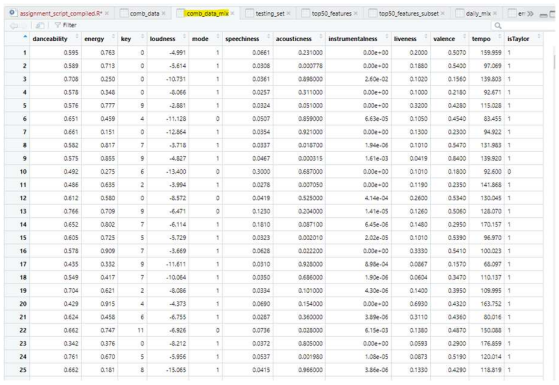


Figure 14d. Daily Mix mixed with Taylor Swift

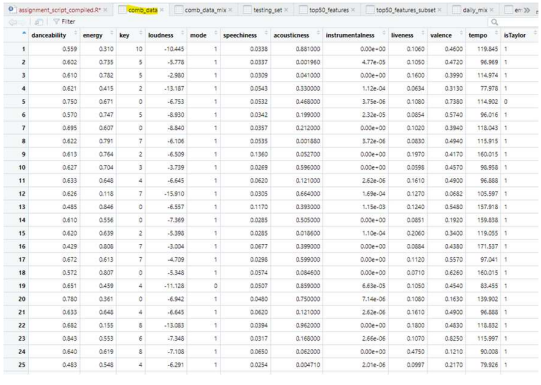


Figure 14e. Top 50 mixed with Taylor Swift

Both datasets are then separated into a 80:20 split, where 80 percent will be used for training the machine whereas the remaining 20 percent will be the testing. On the testing phase of the dataset, the results held a high accuracy of classification. 93.12% for the Top 50 Global mixed with Taylor Swift and 91.78% for Daily Mix with Taylor Swift Songs.

```
[1] "Prediction is: 1. Correct!"
>
>
> # Analyse the model accuracy with a confusion matrix
>
> confusionMatrix(dt_model, reference = testing_set$isTaylor)
Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0  1.8  0.9
         1  6.0 91.3

 Accuracy (average) : 0.9312
```

Figure 14f. Top 50 Testing Data Result

```
>
> if (predicted_label_mix == testing_set_mix[prediction_row_mix, 10]){
+   print(paste0("Prediction is: ", predicted_label_mix, ". Correct!"))
+ } else {
+   paste0("Prediction is: ", predicted_label_mix, ". Wrong.")
+ }
[1] "Prediction is: 1. Wrong."
>
>
> # Analyse the model accuracy with a confusion matrix
>
> confusionMatrix(dt_model_mix, reference = testing_set_mix$isTaylor)
Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0  0.6  0.8
         1  7.4 91.2

 Accuracy (average) : 0.9178
```

Figure 14g. Daily Mix Testing Data Result

XII. Topic Modelling

Topic modelling is a technique in natural language processing (NLP) which aims to discover the underlying themes or topics in a collection of documents. This will help identify patterns and group them according to similarity. This would be helpful in the music industry as underlying trends and topics can be discovered through this algorithm. Furthermore, feedbacks from fans can also be obtained allowing the artist's management team to adjust accordingly.

The algorithm was done by using the data collected from reddit and was cleaned and stripped off of any noisy data. The collected data was reduced down to its stop words and is then converted into a document term matrix. Using the Latent Dirichlet Allocation (LDA) with a set number of clusters in this case is 6. This will then return the top topics from data collected data which can then be used to collect information from the redditors.

This algorithm was ran through 3 datasets which are Taylor Swift, Ariana Grande and Olivia Rodrigo. Below are the results of these popular artists and what their fans talk in their threads.

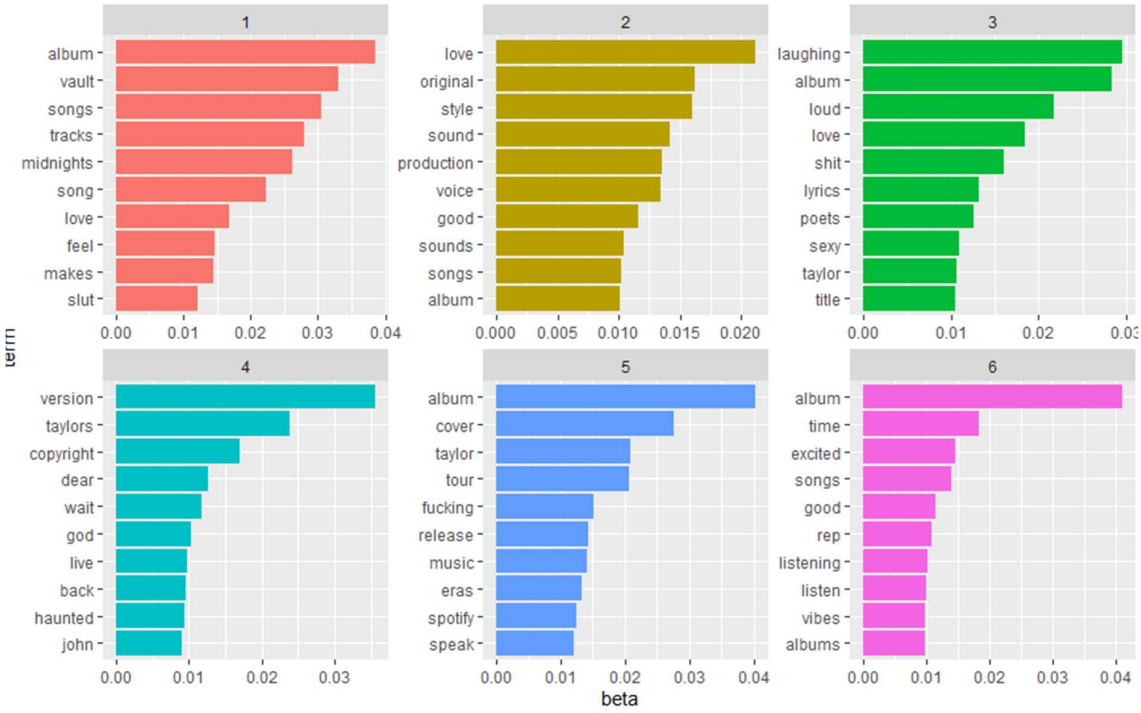


Figure 15a. Taylor Swift Topic Modelling

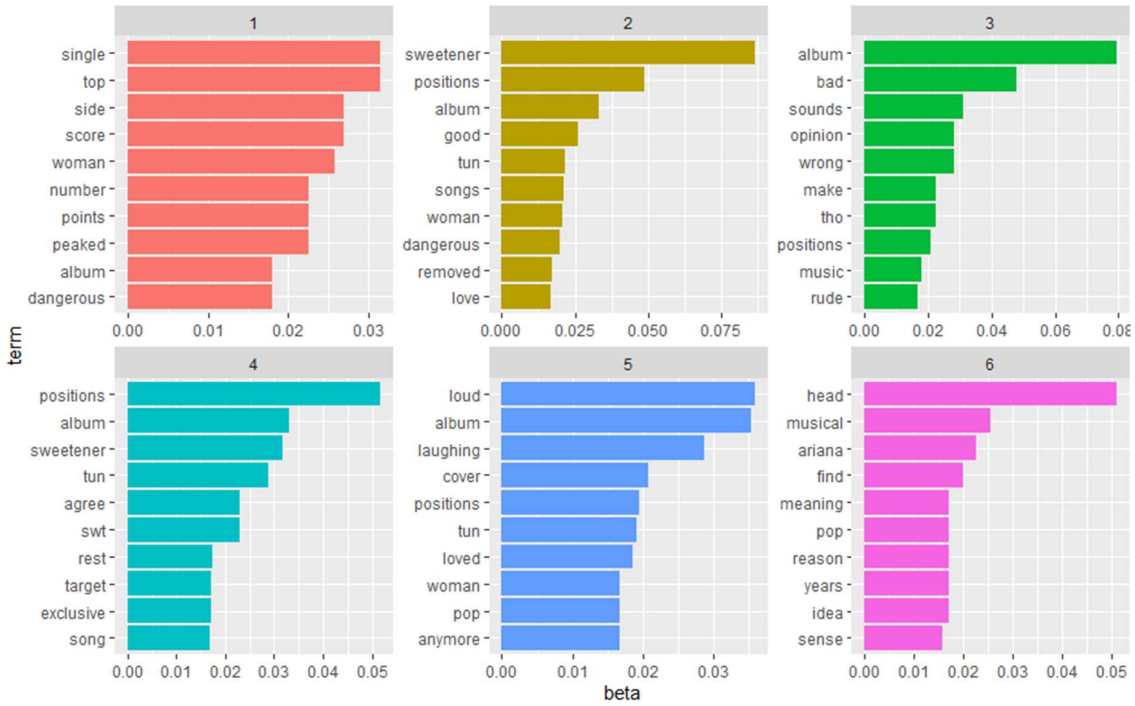


Figure 15b. Ariana Grande Topic Modelling

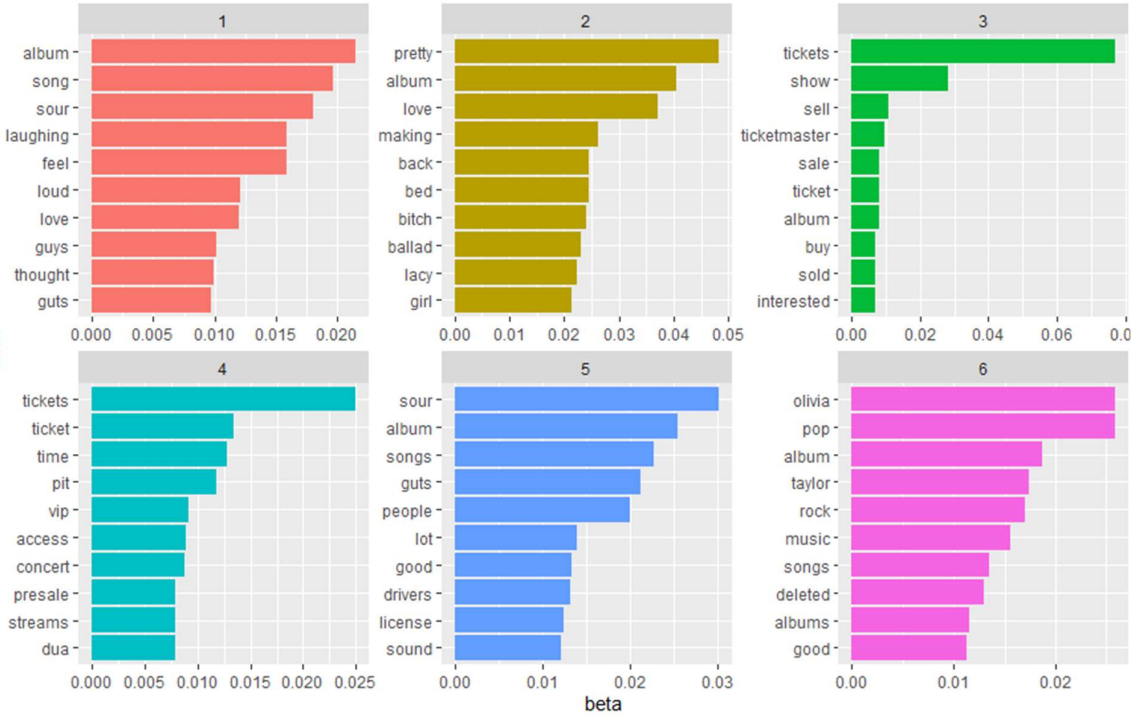


Figure 15c. Olivia Rodrigo Topic Modelling

DATA VISUALIZATION

XIII. Data Visualization for Reddit and Spotify

A. Reddit Data

The visualization using PowerBI will be showcasing mainly of the redditors statistics across the for threads that was data collected. These subreddit thread represents the album released by Taylor Swift in the past 3 years. This visualization we can see how one fan or in business case one social manager can effectively increase an artists popularity

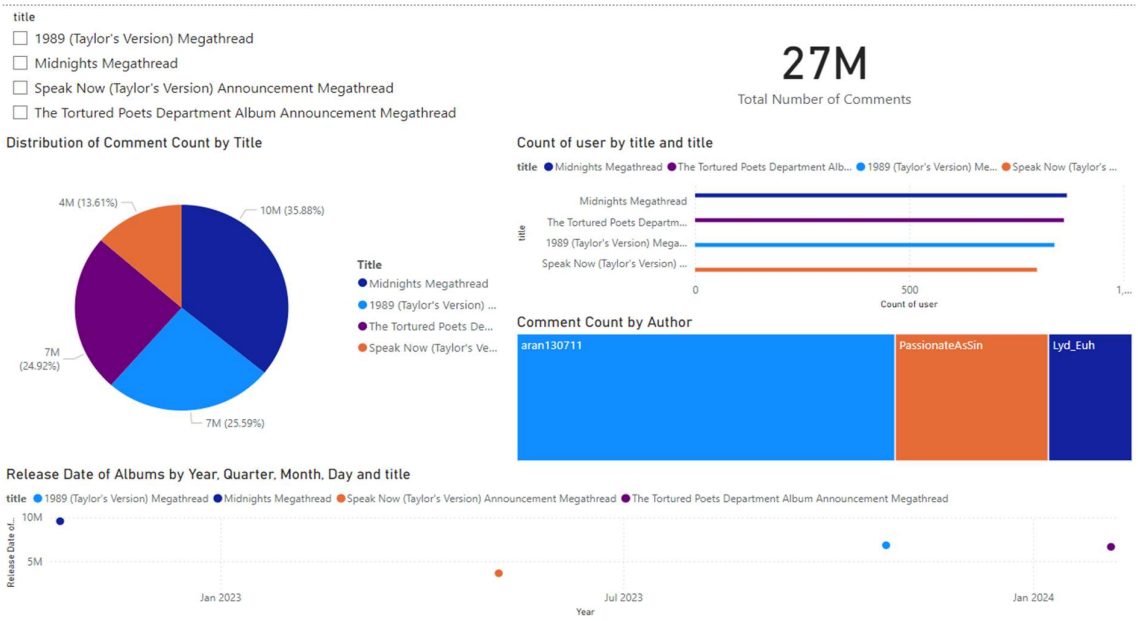


Figure 16a. Reddit overview for Taylor Swift Data

The most interesting findings in this chart is how one user garnered more than half of the interactions in reddit under the thread of Taylor Swift. It can be safely assumed that this user is working for Taylor Swift as a social media manager where the users post gained 16M number of comments which is massive for a reddit's space.

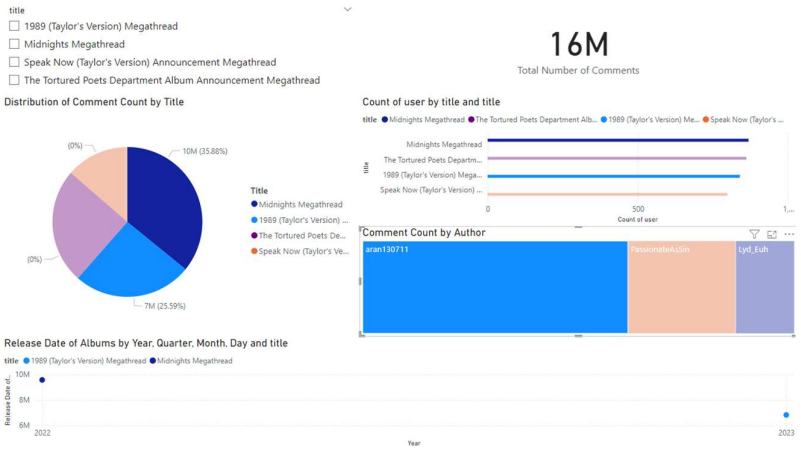
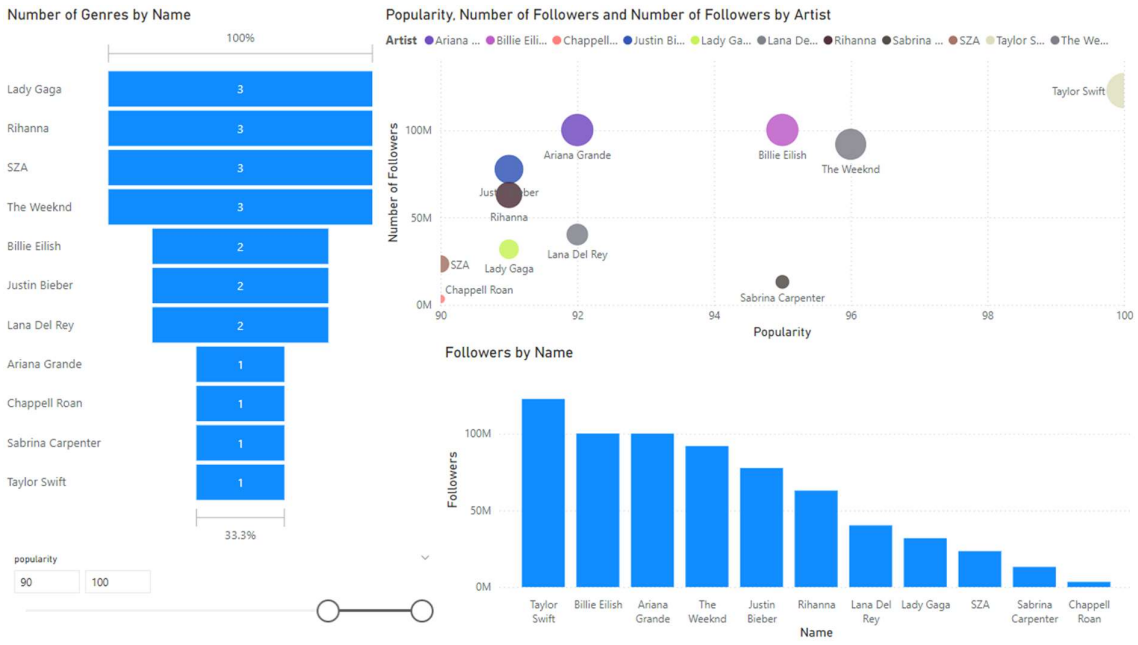


Figure 16b. Social media manager impact

B. Spotify Data

The key visual for the Spotify compares Taylor Swift towards other artist. It can be seen how Taylor Swift is dominating in terms of popularity and follower count. One of the things that can be noticed here is about Ariana Grande. Despite having low presence in Reddit, she has around 100M of followers, but her popularity is only at 90th percentile. With this visual we can safely conclude the fan to fan interaction drives popularity of the artist.



ANALYSIS REVIEW

XIV. Different Models and Visualizations

Conducting social media analysis using various methods and algorithms provides unique insights, and there is no single approach that fits all purposes. For tracking current trends in social media or the music industry, NLP algorithms are highly effective. If the goal is to classify artists, decision trees and their counterpart, Bayesian algorithms, can be useful. For analyzing non-linear relationships within data, advanced algorithms like Bayesian Networks and Neural Networks offer robust options.

Additionally, there are diverse visualization techniques that can effectively present data. For instance, heatmaps are valuable for displaying an artist's popularity across different geographical locations, while network graphs can illustrate relationships between users and topics. When it comes to topic analysis, word clouds are useful for quickly conveying the essence of a topic and its associated terms.