



3032ICT / 7230ICT / 1117ICT

Big Data Analytics and Social Media

Assignment Specifications

Report

Milestone 1

Prof Bela Stantic

Author: Thi Minh Oanh Luong

Student ID: s538355

School of Information and communication technology

Date of submission: 1st April 2023

Table of Contents

Abstract	3
Introductions	4
Overview	3
Case Study Setting	5
Band's overview	5
How many years have they been active?	5
How many albums & songs have they published?	5
Data Selection & Exploration	6
Twitter data retrieval	6
Top 5 most influential users	7
Top 10 most important terms	12
User accounts	13
References	16

Abstract

There has been a considerable increase in the use of social media analytics in order to learn about user behavior and presence. This assignment report presents a method of management and improvement in the popularity of a band by using social media analytics. The study case settings which are searched from several references are the preparation to access 1000 Twitter statuses in Twitter by using various tools. The study also applied data selection and exploration techniques to address the given problem. These findings highlight the top-5 influential users and top-10 most significant terms.

Introductions

By tools (Rstudio, Gephi) introduced in previous labs, data is collected through the Twitter API and several basic social network analyses using the vosonSML package in RStudio and graph visualization software Gephi. The study focusses primary information of a band named Twice to provide overview of their name, genre, members, etc. It also shows their years active and the number of songs and albums they have released. Furthermore, there consists of 4 key stages of Data selection and exploration. Firstly, from the case study, suitable keywords involving the band are chosen to seek 1000 tweets. Secondly, by Rstudio, the top five most influential users for the influential user are also listed, which helps to find out other interests/chrematistics beyond connected with the band. Next, they study shown top 10 most significant terms appearing together with found keywords related to the musicians. Finally, the number of unique user accounts in the dataset is calculated. The purpose of this report is to think about a case study, applying social media analysis to gain insight about how musicians can improve their popularity.

Case Study Setting

Band's overview

Twice is a girl band in South Korea formed by JYP Entertainment. There are 9 members: Nayeon, Jeongyeon, Momo, Sana, Jihyo, Mina, Dahyun, Chaeyoung, and Tzuyu. They follow K-pop, J-pop, bubblegum pop, dance-pop, and EDM genres.

How many years have they been active?

They have been active since 2015 (8 years). The band was formed under the television program *Sixteen* (2015) and debuted on October 20, 2015. At that time, their official fandom, ONCE, was also founded for all fans around the world.

How many albums & songs have they published?

They have released officially 202 songs, with 122 songs are recorded totally in Korean, 44 ones in Japan, and 8 ones in English lyrics. Besides, they have also published 21 albums and 12 Singles & EPs, 4 compilations, and 3 videos.

Data Selection & Exploration

Twitter data retrieval

twice kpop and **#twice** are chosen as keywords to filter out from the database with 1000 collected tweets.

```
twitter_data <- Authenticate("twitter",
                             appName = my_app_name,
                             apiKey = my_api_key,
                             apiSecret = my_api_secret,
                             accessToken = my_access_token,
                             accessTokenSecret = my_access_token_secret) %>%
  collect(searchTerm = "#twice OR twice kpop",
           searchType = "recent",
           numTweets = 1000,
           lang = "en OR ko",
           includeRetweets = TRUE,
           writeToFile = TRUE,
           verbose = TRUE) # use 'verbose' to show download progress
```

Below are explanations of code:

searchTerm = "#twice OR twice kpop"	By using hashtag "#twice", it is deemed more convenient to sift contents relating to the band in the database. Nevertheless, for more comprehensive search, keyword "twice kpop" can yield result involving both their name and genre, which ensures that no relevant recent data is missed during collection process.
searchType = "recent"	The "recent" in search Type benefit providing up-to-date data. Compared to "popular" or "mixed", the type is convenient as the data storage is a stack with the most recent information on the top. Therefore, conducting a "recent" can save more time in searching.
numTweets = 1000	There are 1000 tweets searched. When conducted fewer 1000 tweets might not provide a representative sample, while collecting more 1000 tweets are time-consuming, resource-intensive, and low-quality data for analysis.
lang = "en OR ko"	By selection of both English and Korean as the language option, all status in the result is displayed in either English or Korean. This makes data collection process is likely more global and diverse sources.

<code>includeRetweets = TRUE</code>	Retweets matching the research term are also included.
<code>writeToFile = TRUE</code>	The result should be written to a file.
<code>verbose = TRUE</code>	The collection process is shown in the console.


Top 5 most influential users

Find the top 5 important users in your actor graph:

```
# Run Page Rank algorithm to find important users
rank_twitter_actor <- sort(page_rank(twitter_actor_graph)$vector, decreasing=TRUE)
head(rank_twitter_actor, n=5)

# Overwrite the 'name' attribute in your graph with the 'screen name' attribute
# to replace twitter IDs with more meaningful names,
# then run the Page Rank algorithm again
V(twitter_actor_graph)$name <- V(twitter_actor_graph)$screen_name
rank_twitter_actor <- sort(page_rank(twitter_actor_graph)$vector, decreasing = TRUE)
head(rank_twitter_actor, n = 5)
```

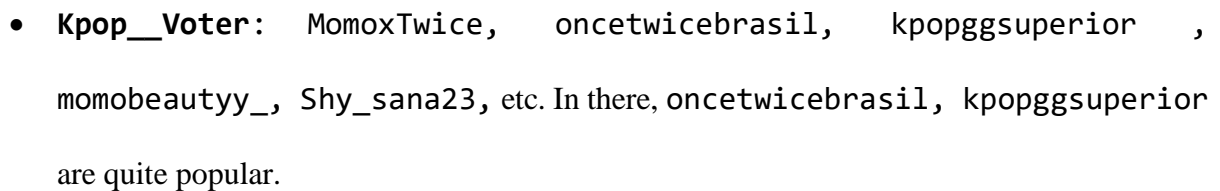
The below table is top-5 most influential users for “Twice” band in Gephi:

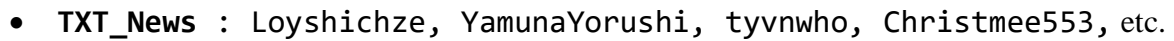
Id	Label	Interval	name	screen_name	PageRank 
n910	n910		1566393722199474176	CeiiLevis	0.037874
n194	n194		1956910224	SerieTV46	0.023835
n570	n570		1617751720922316801	Kpop_Voter	0.019869
n143	n143		939906095426007041	blackpizzah	0.016934
n915	n915		1085062354017050624	TXT_News	0.015744

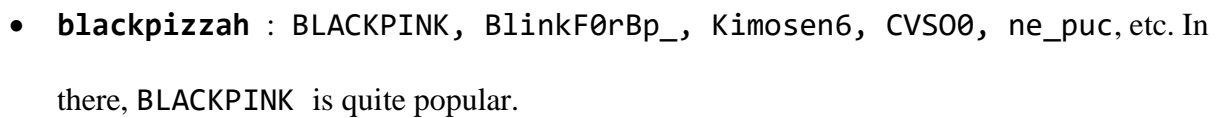
In the top 5, there are some other interests/characteristics they have besides those related to your artist/band:

-

- ## Assignment Specifications Report – Milestone 1









```
# Run Page Rank algorithm to find important terms/hashtags
rank_twitter_semantic <- sort(page_rank(twitter_semantic_graph)$vector, decreasing = TRUE)
head(rank_twitter_semantic, n = 10)
```

Assignment Specifications Report – Milestone 1

Id	Label	Interval	name	type	n	PageRank
n3	n3		top	term	8.0	0.046072
n0	n0		kpop	term	110.0	0.044518
n4	n4		songs	term	8.0	0.026319
n47	n47		#kpop	hashtag	34.0	0.024829
n10	n10		mv	term	9.0	0.022479
n33	n33		#twice	hashtag	58.0	0.022344
n17	n17		blue	term	7.0	0.022342
n69	n69		#pasabuy	hashtag	3.0	0.016603
n46	n46		group	term	11.0	0.01545
n11	n11		views	term	10.0	0.014338

Explain:

Term	Description
top	trending topic
kpop	primary genre of band
songs	their products
#kpop	hashtag of main band's genre
mv	their products
#twice	band name
blue	a word of "baby blue love" - one of their songs
#pasabuy	a hashtag of a Filipino slang where a person will ask another one to buy him/her stuff like some foods, groceries, essentials etc.
group	group
views	view

These terms provide a glimpse of discussion, top-titles, and other contents relating to "Twice" band, including their musical career and personal lives, are shared in Twitter.

User accounts

My code in Rstudio:

```
#calculate unique user in the database
library(dplyr)
n_distinct((twitter_data$tweets)$user_id)
```

In this section, I use `dplyr` library to handle it.

I access `twitter_data$tweets` to use `n_distinct()` function. It counts the number of distinct combinations in a `user_id` set. This method is faster and more concise to others.

The result is:

```
> n_distinct((twitter_data$tweets)$user_id)
[1] 900
```

The result shows that there are 900 unique accounts in my dataset who are interested in the band. This means, in the list of 1000 top-recent Twitter statuses, 900 different accounts have been identified. Moreover, it is inferred that these accounts are active, so they can be used for other purposes.

Conclusion

The report has indeed proven that social media analytics through case study and given software are applied to help Twice band improve their popularity. It is evident that all keywords involving the band are selected to retrieve 1000 recent status in Twitter database. The study also lists the top 5 most influential users in social media and their interests except those related to the band through tools. As well as the top 10 most important term conjunctions with keywords are shown. Additionally, these findings show there are 900 unique accounts in the dataset. Thus, I can leverage the results to deal with Milestone 2 in assignment 2.

References

- 1) Discogs.(n.d.).*Twice (25)*. <https://www.discogs.com/artist/4786543-Twice-25>
- 2) Spotify.(n.d).*Twice*.<https://open.spotify.com/artist/7n2Ycct7Beij7Dj7meI4X0>
- 3) Kat Moon.(2019).*Everything to Know About K-Pop Group Twice*.Time.
<https://time.com/5671342/twice-k-pop-everything-to-know/>