Rahul Rajeev
4/4/23
DSC 540

Factors of inorganic views in the KPOP industry

Purpose:

To study attributes of kpop idols and idol groups and how they could potentially affect the use of inorganic views. Do specifically formed groups from higher established companies require more botting and inorganic views?

Data Sources:

- CSV(s):
  - Link to Kaggle Zip and Site
  - This group of files will be used for the project as one large dataset, the four files that are available are a music video csv, an idol csv, and male/female idol group csvs
  - Only has information up to 2020, so perhaps I could use newer idol groups as a testing set
    - The music video csv is the main file
    - The other csvs provide insight about each idol and who they work under
  - Includes information about soloists, groups such as birth date, debut date, group, gender, company, etc.
  - Most importantly, the music video csv provides a youtube link, which can be used in an api call per video to gather further information
- Youtube API:
  - Link to Documentation
  - Get information about statistics: likes, dislikes, view count
  - Get information about content details: duration, content.rating - mpaa rating
  - Use youtube link from the csv to make api call and gather the information
  - There are more sections to the api call, but the two I wish to work with are the statistics and content details
- Website:
  - Link to Website
  - The website provides tabular data about the amount of inorganic views, real views, and a percent of real views. Depending on the complexity of scraping, I may have to create my own columns using only the inorganic view range and the real views from youtube, as it provides the most accurate and recent amount
  - Checks status based off youtube or algorithm (how much views are real)
    - Represented by symbols in the table, not sure how it's going to read
  - There is a switch that helps show more data, but I don't know how to call for it to be switched on as it shares the same exact url.

Relationships:

- Music video csv artist names → csv of artists from solo, boy or girl group
    - (if artist name doesn't exist, match by korean name)
    - this gives information about soloist or group origin
- Music video youtube links from csv → individual api calls
    - to get likes, dislikes, view count if necessary (recorded by youtube), and ratings
- Music video names from csv → html table video names to get the inorganic view count, status of view count, percentage of real views

Ideal columns of dataset:

- **(8) Music video csv:** Date of music, artist, song name, korean name, director, video link, type, release
- **(6) group**: debut, company, number of members (original and final), active?, fanclub name
- **(9) Solos/idols csv:** full name, korean name, k stage name, date-of-birth, group, country, birthplace, other group, gender
- **(4) Youtube api:** views, like count, dislike count (if available), rating
- **(2) Website:** nonorganic views, status (validated)
- **(4) Created columns:** percent of real views to inorganic, inorganic views lower bound, inorganic views higher bound, range of inorganic views
- There is still room for improvement

Summary:

First I will download the csv zip file and perform the necessary joins to form the 4 files into one dataset. Following the imports, I will then clean up the dataframe, renaming columns as needed, and creating ones that I can for the first set of data. Then I will create a helper function that performs api calls on each video to get the information I need. Finally, I will have to scrape the html with the information about the inorganic views.

The data I collect will help illustrate the relationship between different attributes of idols and idol groups and the need for our existence of inorganic views. The only columns I'll actually end up not using for calculations are the string columns such as korean name and k stage name, although the previous will be used to match idols if their stage name doesn't match their idol name. Fanclub name and activeness only really affects some aspects, but it could be an interesting project to see whether an established and active fanbase could affect viewbotting.

The ethical implications are obviously going to be visualizing and presenting some of the flaws of the Korean pop industry. While viewbotting isn't necessarily an uncommon occurrence, as it does also affect video views in the US, it could be interesting to see whether viewbotting is used to spread influence or gain a larger audience outside of the nation.