02450 Social Graph and Networking Final Project: A Study of K-pop Lyrics

Chang-Liang Lu *(s220034)*[a] and Xueying Chen *(s216410)*[a]

[a]Denmark Technical University

This manuscript was compiled on December 14, 2022

Our biased choices will be influenced by culture, experience, education, information, etc. This article researched society's preference for Korean pop songs and the difference between international and national. Since we have heard the values and focused points from international and national (Korean) people on the songs are different. Also, a rumor says people prefer sentiment songs to the pure party or happy ones. To prove both rumors, we have scraped the data from the top 1 music entertainment stream platforms worldwide and in Korea, using them to select the top songs to analyze the similarity between the lyrics and the corresponding sentiment scores. Here we can see the main graphs about the similarity connections between each piece, the wordclouds to know the frequency occurrence of each word in the lyrics by each group, and the distributions of their sentiment scores. For the last two, we have separated into three sections— national, international, and combination.

Networks | Python Scrap | Data Mining | Data Visualization

Well- written lyrics can burnish a song, arouse great echo in people's hearts. Yet, not only happy cheerful songs are appreciated, sad songs are also popular, even when we do not subconsciously think about it. In this project, we want to find out if the top songs are more sentimental, regardless of its listener can understand its lyrics. Hence, we will compare the the K-pop songs from Korea and also the whole globe.

Data Mining. The general idea of constructing the national and international song charts can be divided into several steps:

1. According to popularity, the top 20 Korean bands would be selected.
2. Their top 10 songs will be extracted.
3. The lyrics of these 200 songs will be collected.

These three steps would be conducted separately on Korean platform and international platform. Since we needed to define what songs can be categorized as K-pop songs, we have downloaded the data set from Kaggle: K-pop Database 1, about the k-pop groups and idols from the year 1992 to the year 2020 in order to help us find and scrape our data from the website quickly.

Data for national chart. Melon 1 is the most famous music entertainment streaming platform in Korea. The tool for scraping the top 100 songs in Melon is the API made by a person who shared it on his GitHub page. We used it and received the data with a dictionary type. Because the data structure was not steady, lacked some components; it wasn't easy to use the function or code to clean it. To prevent the mistake of extracting the data we wanted, we took out the groups which showed on the Kaggle data frame and used a manual way to construct the top 10 songs data frame for each group that appeared in the data we scraped. However, some of the groups showed up in the dictionary not only once, so some of them we took 20, 30 pieces, or more than that. Depending on the proportion they have seen in the top 100 rankings. (Ex. BTS, this group has 5 songs in the dictionary, then we added 50 pieces in its data frame since 5 multiply 10 is 50) The reason to do this action is due to the k-pop songs we defined just taking a part of proportion in it. The ranking chart also included hip-hop songs, single singers, and the band does not belong to any entertainment company. Moreover, the data we wanted also included songs from the same group. Therefore, to avoid the size of data too small, we chose this method to continue our work.
One of the methods to receive the lyrics is by scraping the content on every group's k-pop fandom page. They have offered the English version. However, the challenges are that some of them had various forms for the lyrics, putting it in other positions or not giving. Also, the websites maintained an encryption system to prevent scraping. They pull out the letter "v" inside the lyrics. Because of these problems, we have a lot of manual operations taking us a lot of time.

Data for international chart. Spotify 1 , as one of the most influential and international music platform, which also has it own API, is the first choice when it came to build the international top chart. Its API has different classes and `Artist` and `Track` would be the mainly deployed ones. Both of them have an

Significance Statement

Korean-style pop music has become more and more popular in the world. Because of this situation, a few rumors gradually popped out. "The taste of k-pop songs, national and International people are different." , "The international fans focused more on the songs' vibe than the lyrics since they cannot understand the meaning at first listen", "People more like sentiment songs than others", etc. These rumors have been left on social media for a long time. Hence, it is pretty interesting to prove and correct for those rumors. It is also essential to figure out these questions. The stereotype is a classic example. Without disturbing and telling the truth, people will gradually believe unreal things. It eventually will turn into a significant impact on society. That is why "proof" is a crucial part of science. Another critical theme is that the way we analyzed is actually helpful for music therapy. They can use the method to explore what kind of music type will heal traumatized patients most effectively to improve this method.

[1] Cheng-Liang Lu collected the national data and Xueying Chen the international data to this work. The rest part are equally contributed.

attribute called `popularity`, which is exactly the measuring unit for the chart. Thus, the top 200 songs were gathered in no time. The first challenge that we encountered is, that the Spotify API does not support any kind of lyrics fetching, we have therefore turned to Beautiful Soup 1 and Selenium 1. As one of the most-known python scraping API, Beautiful Soup is good at capturing elements on websites. Combining with Selenium, which could simulate all sorts of website interactions, like clicking and entering information, they are able to bypass the obstacles set by Spotify: one must log-in to see the lyrics and click twice to see the song credits.
While we were scarping lyrics and song credits from Spotify, we could feel that Spotify is not happy about this behavior. There were two fairly common problems that occurred:

- The responses from Spotify were not stable, hence the time-out error
- Spotify has some sort of anti-crawler protection, hence the access were blocked from seeing some of the lyrics from time to time

As a consequence, we set a highly tolerant waiting time for the website, as well as changing accounts and using VPN a few times, finally the 200 lyrics were successfully gathered.
The most challenging part of this data collection is that the lyrics are primarily in Korean, partially in English, and marginally in Chinese, Japanese, French and Spanish. And the translator API that we used, can only handle the first appeared foreign language (as foreign consider to English) and ignore the rest at auto mode, even the first foreign language appeared again after the second one. Therefore we need to first define what is the primary foreign language of the song, then this language will be translated throughout the whole song, and the secondary language can be translated in next session, until there are only a few foreign characters left, which would require manual fix.

Data Analyzing. First, we make a distribution of similarities to see if the top songs have similar lyrics. To visualize clearly, we have created a beautiful connection graph that has included colors of nodes representing the songs, whether they belong international chart, national chart, or both. The links between them are similarity scores between each other. The color bar explains the meaning of the edges in the graph. We raised the standard for building links that the scores need to surpass 0.95 and the connect would be made. To prove our hypothesis, break the rumors, and discover an interesting phenomenon, we make sentiment scores to observe human preferences and society's tastes in Korea, the world, and both. Then, used word cloud to analyze what words have the higher frequency appearance in the same group's lyrics of all songs. In order to realize what kind of words are the most frequently applied in specific groups and what most fans favorite. Also, we make sentiment scores to observe human preferences and society's tastes to prove our hypothesis and the rumors on the Internet.

Results. For the first part, it was interesting to see that the top songs we collected from both music media stream platforms have a result shows the high peak of sentiment distribution would lean to 0.8 and 1.0. This means that most of the songs' lyrics content is similar. Hence, we could infer that they may contain some common crucial keywords or special meanings inside the content that hit everyone's point. However, there is still a cluster in the range of 0.0 to 0.4, centered at 0.2. That means it still has some unique style that hit a few individuals' minds, although the figures are much lower than the previous trend.

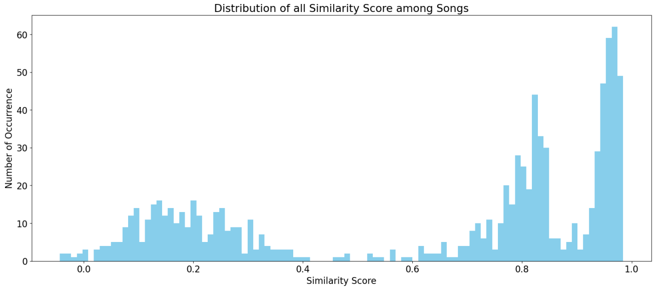

Fig. 1. The Similarity Distribution among all songs

Since we knew the highest occurrence cluster group from the previous section, we would like to understand more about their relationships and gain more details from this beautiful graph. We can pay attention to the nodes' color first. We separate them to be red (national group), blue (international group), and purple (from both). There are strong connections between them, and it is obvious that they are divided into two clusters, one national and one international, and the purple ones mix between them. The significant attraction of this graph is the biggest cluster. If we look carefully, we can find the connections between nodes are powerful, and the scores are almost close to 1.0.

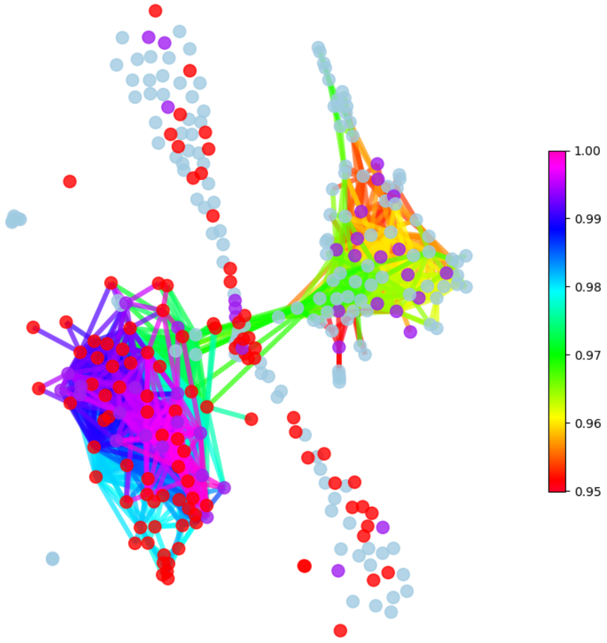

Fig. 2. Similarity Connections shown in FA2 Graph

Evident to see most of the nodes are from the national

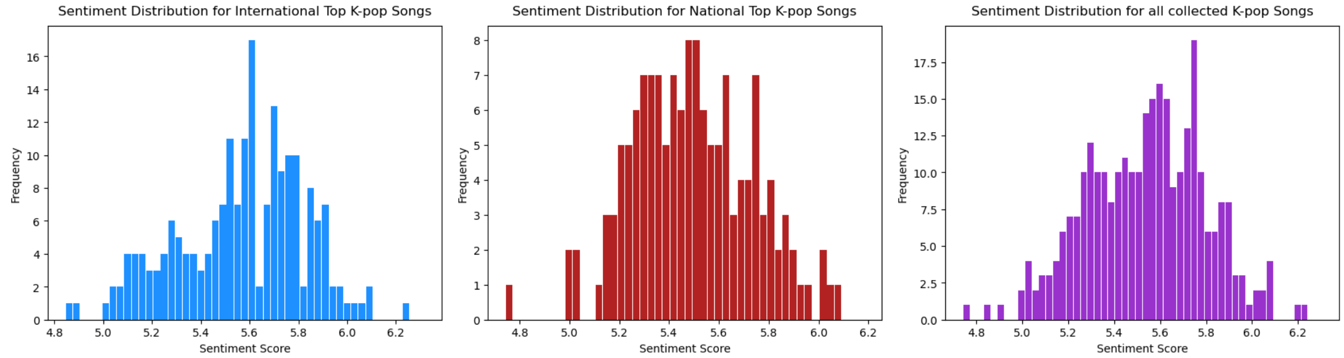

Fig. 3. Three sentiment scores from left to right are international(blue), national(red), and the combined(purple).

chart in the biggest cluster. This means that in most cases, compared to international people, they will have familiar tastes or favorite styles in Korea, which make sense for people from the same region or countries. They will have similar perspectives compared to other people from elsewhere. However, it is still surprising that many International music listeners have a close preference for Koreans, as the score of 0.94 is not a low number.

Turn into the next section. We focus on comparing the difference in sentiment scores distribution between national pieces and international one. We can see that both patterns seem similar; however, if you look at the horizontal axis, you will find the national one is slightly at the left, contrasting the international distribution. This means that the whole pattern of sentiment score is lower than the international one. This situation represents that Koreans may prefer sad songs a little more than the ordinary people in the world.

In the last section, the diagram shows the frequency tokens in the word cloud. The words appear more common in the set of lyrics, and the text will be greater. We divided into groups to show that each of the tokens performed in their lyrics.

Readers may notice that some of the groups' canvases look empty, and three possible reasons might cause these:

- The content inside the lyrics is highly duplicated.
- The lyrics lost the essentials after translation.
- Lack of lyrics resources on the Internet. This means either they may be new groups, so people will not pay much effort to build lyrics or they were not famous international groups, which leads to we need to translate them and lose some keywords.

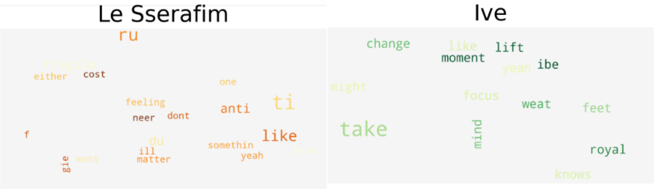

Fig. 4. Wordclouds of groups, Le Sserafim and Ive, with less tokens

On the other hand, several sub-graphs filled many keywords in their frames. This phenomenon is attributed to three factors.

- The groups emerge on both national and international charts, so the groups have more tokens than others to pick by wordcloud.
- Since these groups are famous worldwide, many passionate fans must translate lyrics and build the wiki fandom pages to let us have resources to use.
- These kinds of songs may have lyrics on wiki fandom to skip the translation section(since those came directly from the lyrics of wiki fandom.) without translating to other words. Or they repeated catchy sentences, again and again, for the music style.

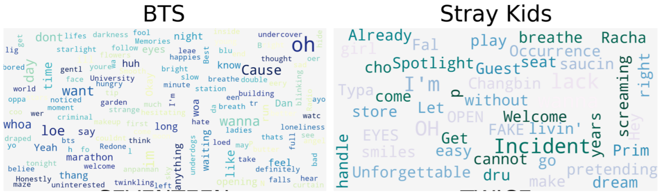

Fig. 5. Wordclouds of groups, BTS and Stray Kids, full of tokens

Although we have these factors exist in our project, we still surprisingly found that many worldwide popular groups have two same patterns. One is putting key onomatopoeia inside multiple positions in the lyrics, and another keyword is “love”. (readers may see “loe”, which also represents “love”) Love may be pretty common sense, for love songs play a huge part in the music world. Nevertheless, It is fascinating to prove that the K-pop industry loves to use onomatopoeia as keywords inside songs. This style is a unique part of the global music industry, and we are also excited that we have proved this doubt to be true.

Discussion. Through this experiment we have proved and disprove some rumors on the internet and fulfill our curiosity. However, our system is still not stable. Most of the challenges came from the the data mining part:

- Faced with translation punctuality,nature speaking, and multiple languages making translator cannot work.

- Webpages encryption that cause words which contained "v" without that letter.
- The form of lyrics in k-pop fandom pages, sometime, are not the same.
- The songs famous in Korea but not international brought about not able to find lyrics on the Internet.

These problems took us a big trouble. We had to fix them with either manual ways or bunch of coding. Although we had checked it throughout our dataframe, it was still a large numbers of content for human eyes. Hence, we hope we can find a suitable website offering a stable lyrics resource including some powerful APIs or build a strong API by ourselves that we can use in the future.
We also want to implement our theme not only for "korean pop songs" but also the pieces from the world. K-pop songs are just one part. We anticipate we could extend our range to the pop songs around the world. Develop it to become a more complete tools that can be implemented in different areas. For example, music therapy for choosing the right song with right lyrics, or for the music makers as a guiding composer, etc.

Materials and Methods

Data. A data set includes K-pop idols from 1992-2020, which is the data set we build our data set upon. It has 10 attributes includes: `Stage Name`, `Full Name`, `Korean Name`, `K. Stage Name`, `Date of Birth`, `Group`, `Country`, `Birthplace`, `Other Group`, `Gender`. Here is the link to it.

Websites and their APIs. A list of the website source, that we collected our data from by deploying their APIs.

Melon. A digital music shop and music media streaming platform. The name originates from the English "Melody on". It is the biggest music subscription platform in Korea. The scores or ranking on it also reflects the music trend. "Melon prize" held every year is seen as one of the main ceremonies in this country. Click here to see the main page of Melon and the page of Melon API.

Wiki Fandom. A website people can create and build a page with knowledge, similar to Wikipedia. But they focus more on different media, and a lot of websites are built by fans. People who visit these web pages are not only readers but also writers.
To see an example, please click here.

Spotify. It is a Swedish own audio streaming and media services provider. As one of the largest music streaming service providers, with over 456 million monthly active users, including 195 million paying subscribers, as of September 2022 (1). Click here for its Web API reference.

Python APIs. This project is built on several python package, which we will briefly introduce here, and please refer to the code 1 for more detailed usage.

Beautiful Soup. It is a python based HTML parser, which is known for it web scarping ability. Here is its documentation page.

Selenium. It is originally designed for website testing, and now is a browser automation, which could simulate a range of human-browser interactions. Link to its website.

spaCy. It is an open-source software library for advanced natural language processing (2). In this project is used to calculate the similarity score among songs. Click here to view its website.

1. Wikipedia, Spotify (2022) Last accessed 14 December 2022.
2. Wikipedia, spacy (2022) Last accessed 14 December 2022.

Appendix.

Datasets. We collected two datasets: The national K-pop song dataset, includes `index`, `NO`, `song`, `group`, `lyrics` in total 5 attributes.
The international K-pop song dataset, includes `track`, `artists`, `uri`, `album_name`, `popularity`, `valence`, `composer`, `producer`, `lyrics`, `translated`, `similarity` in total 11 attributes.

Project Code. Here is the internal link of the project on DTU oneDrive. And if you prefer github.

Figures. Please refer to the explainer notebook 1 for more graphs and figures in detail.