Analysis of album comments based on NetEase Cloud Music

Dongfang Wang *, Cheng Huang

Department of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: wdf0517@163.com

Abstract. This paper aims to explore the user's comment behavior and the user's emotional tendency towards each song in the album. In this paper, the comments in the Netease Cloud Music are taken as the analysis objects, and Python is used as a tool to crawl data, clean data, segment words, generate themes, and analyze emotions. Finally, visual analysis is carried out through R language.

Keywords: Comment Mining, Sentiment Analysis, Visualization.

1. Introduction

In recent years, with the continuous development of information technology, the network has been closely connected with people's lives, which is difficult to give up. More and more people tend to express their ideas and insights on the Internet, some choose to express their views on a certain work (such as a film, a book, or a song), and some like to communicate with others and share their life dynamics on the Internet.

In daily life, we can often observe people wearing wired headphones or Bluetooth headphones listening to music everywhere. Music is all encompassing, carrying people's feelings of joy, sadness or pain. People often resonate in music and leave their own little feelings and footprints in the comment area.

At the same time, text mining [1] began in the last century, and now it has a mature theory and system. Text mining has also gone deep into various fields, such as film comment information mining and commodity comment information mining, but there is still less research on song comment information mining. Based on the above background, this paper will capture the comment data information of all songs contained in an album on the Netease Cloud Music through web crawler technology, and further clean and process the crawled information. Finally, we will conduct analysis and research. By analyzing the comment information of each song in the album, we can explore the attitude of the audience, the enthusiasm of the song, the theme of the song, and even the emotional trend of the entire album. Meanwhile, by analysis, it is also useful for audience to learn about this album and provide constructive suggestions for the platform to change the push mechanism.

2. Data Acquisition and Processing

This paper relies on the Netease Cloud Music to crawl the comment data of each song in the album Treasure Hunting as the basis for further research and analysis which is released by Chinese music male singer Vae Xu in 2018.

2.1. Data Acquisition

Since there are many anti-crawling mechanisms for directly crawling from the Netease Cloud Music website, it is difficult to get all comments and it is easy to block the IP address. Therefore, this article uses a project in Github -- Netease Cloud Music API as the interface to crawl, which is fast and has no anti-crawling. At the same time, it combines the Requests and JSON in Python to crawl data [2] [3]. The data obtained include: the user’s name, comment time, comment content, number of likes, and lyrics text of each song of the nine songs in the album--Treasure Hunting.

The basic process of a crawler mainly includes the following four steps: Step 1: Specify a URL to initiate a request and wait for a response from the server; Step 2: Get the response content. When the response is normal, you will get a response, whose content is the page content obtained, it may be HTML, JSON and other data types; Step 3: Parse the content. Parse the obtained content, HTML format is parsed using regular expressions; JSON strings are directly converted into JSON objects for parsing; Step 4: Store data. According to the required form of the data object, TXT, CSV, XLSX and other formats can be generated.

2.2. Data Preprocessing

After obtaining the corresponding data, we need to preprocess the data before further analysis. This paper mainly cleans the data by the following steps [4]:

(1) Data deduplication: Filter and delete duplicate data;

(2) Mechanical compression deduplication: Deduplication of continuous repeated words in the text;

(3) Short sentence filtering: Delete some short comment data;

(4) Emoji expressions filtering: Because the comments contain emoji expressions, this part of expressions is converted and filtered;

(5) Word segmentation: This paper uses Pkuseg [5] and user-defined dictionary to segment the text;

(6) Remove stop words: The text after word segmentation contains many meaningless words. Use the stop words dictionary to delete this part.

3. Analysis

In this paper, we will analyse four aspects: the overall distribution of comments, song buzzwords, the sentiments expressed by songs and the topic model.

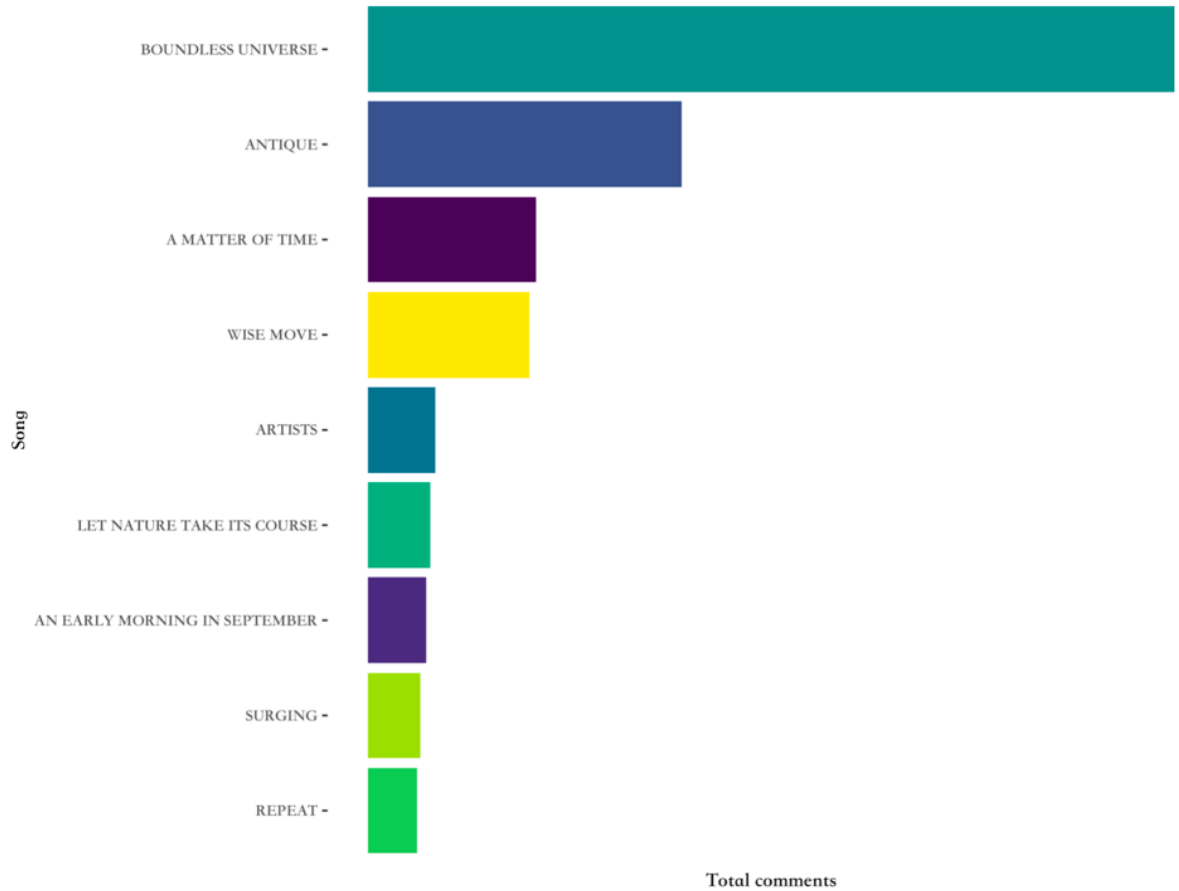


Figure 1. The distribution of comments for the album—Treasure Hunting

Table 1. Three Scheme comparing

Song Type	Song	Release Time	Platform
Title Song	Antique	2018.5.14	General Release
	Boundless Universe	2018.6.1	NetEase Cloud Music
	A Matter of Time	2018.6.12	QQ Music
	Wise Move	2018.6.21	TT Music
Non Title Song	Other	2018.7.12	General Release

As shown in Figure1, from the perspective of comment volume, the most popular track in Treatment Hunting is Boundless Universe, which has a much higher comment volume than other songs, followed by Antique. A Matter of Time and Wise Move have similar comment volume, while Repeat is relatively unpopular.

However, the popularity of a song depends not only on the number of comments, but also on the distribution platform, launch time, market publicity, market position and other factors. Based on the release information of the album (see Table 2 for details), it is not difficult to see that when the album is released, a song is usually selected to be released on the whole network with the highest exposure. Later, other special songs were introduced to the fixed platform for drainage, and finally the whole album is released. Obviously, Boundless Universe, which is first released on NetEase Cloud Music in a high exposure situation, has a far greater volume of comments in NetEase Cloud Music than the other songs. Although Antique is the first released title track, the audience and their comments are scattered on various platforms. Among non-title tracks, there is not a huge gap between comments. Artists has attracted more audiences and aroused more resonance among non-title tracks.

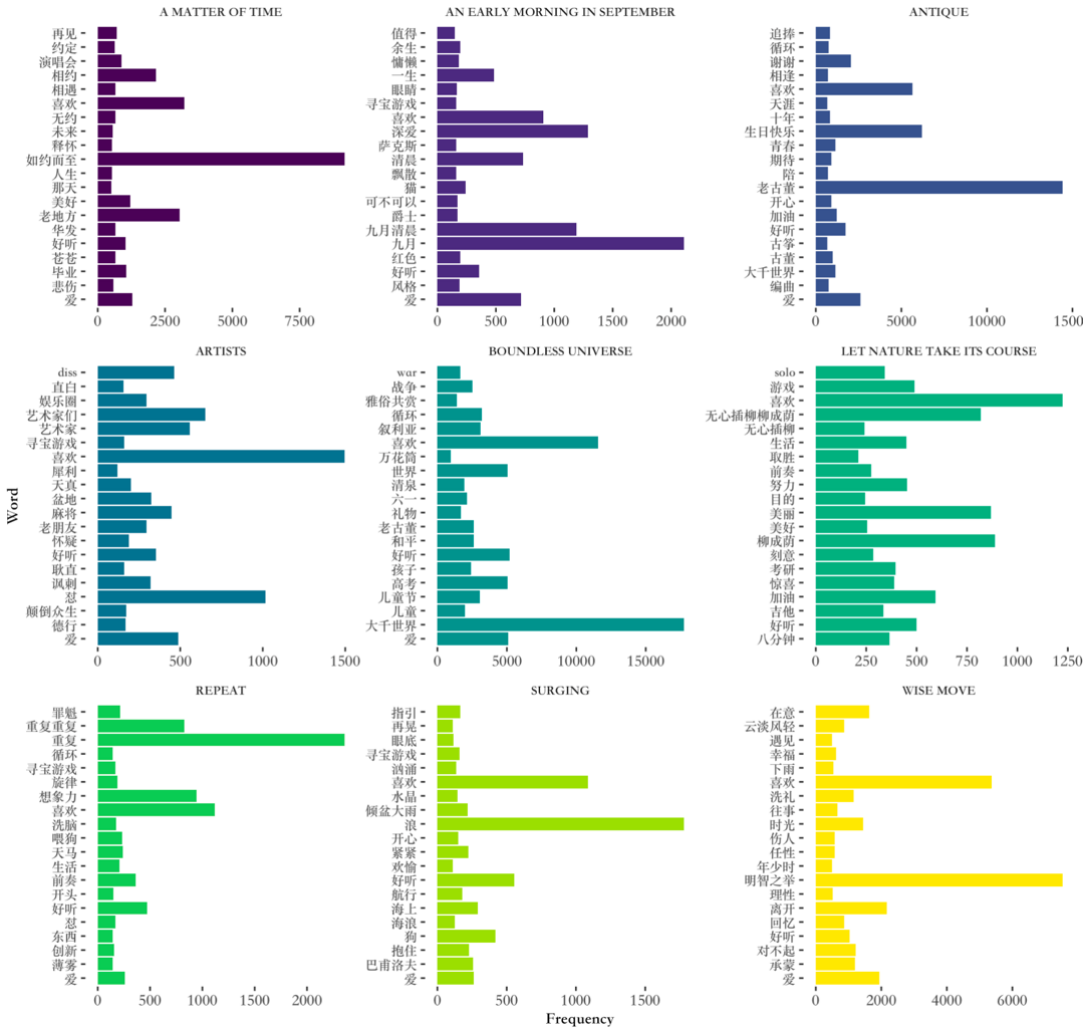


Figure 2. Top 20 words in each song

As shown in Figure2, the top 20 words of each song in the Treasure Hunting are obtained by deleting the words related to the creator and processing. According to the statistics of word frequency, there are some words expressing the audience's attitude such as "Nice", "Like", "Love". There will also be cross-appearance of other song names or album names. Also, when a part of the song or instruments impress the audience deeply, the audience will leave their own footprint, such as " Orchestration ", "Guitar", "Saxophone", "Guzheng", "Jazz". In addition, there are theme words in each song, such as "June 1", "Child", "Children", "Children's Day", "World", "Peace", "Syria", "War" in the Boundless Universe. It can be seen that the theme of the song is related to looking forward to peace and paying attention to children. Or the "Promise", "appointment", "Old place" and "Encounter" in the A Matter of Time, which means that the theme of the song is related to the appointment.

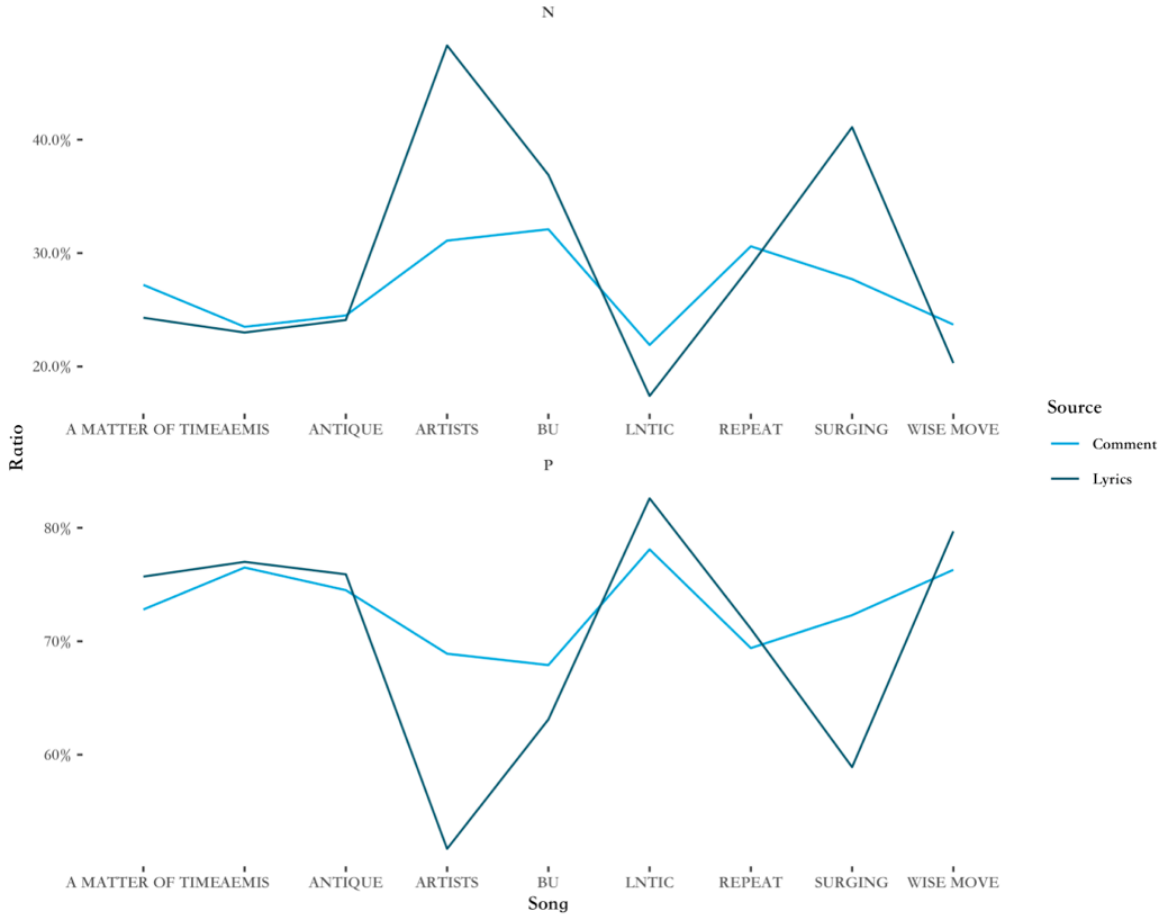


Figure 3. Comparison of emotional distribution between lyrics and comments

In this paper, the emotional analysis [6] [7] [8] in Snownlp is used to compare the emotional distribution of lyrics and comments of each song, where N represents negative and P represents positive.

First of all, the emotional expression of lyrics is relatively positive. The emotional expression of Let Nature Take Its Course is the most positive, while the emotional expression of Artists is relatively negative. It is not difficult to see the author's emotional attitude by referring to the hot words "Satire" and "Grudge" in the comments above. Perhaps the song is a revelation and satire of some social phenomena. Secondly, the emotional expression of the comments is also very positive, and the proportion of positive emotion in each song's comments is maintained at a relatively gentle state.

In general, comments are more positive than lyrics, which is not unrelated to the audience's personal preference in the comments. The emotional inclination of the lyrics of a song will affect the emotional inclination of the comments of the song to a certain extent.

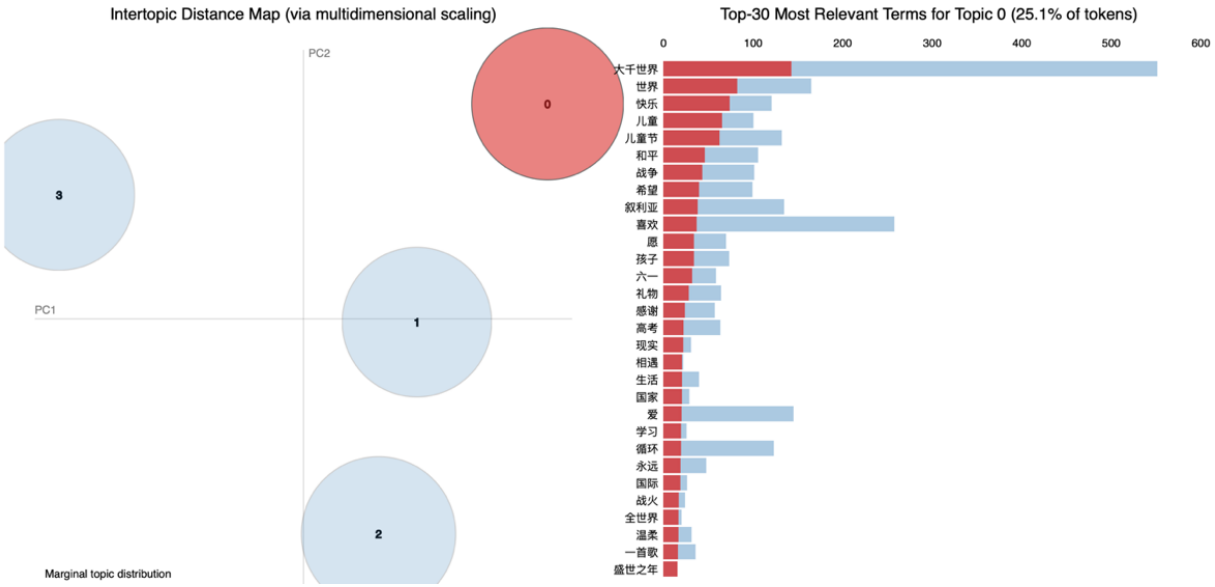


Figure 4. LDAvis

Latent Dirichlet Allocation (LDA) [9] [10] is one of the topic clustering models commonly used in text mining. Through LDA model, we can mine the potential important feature words in the text, which is convenient for text analysis. Therefore, this paper takes the comment data of Boundless Universe as an example to conduct theme analysis. By drawing the relationship between coherency and the number of topics, it is found that when the number of topics is set to 4, the coherence is the highest, and the topics obtained at this time are evenly distributed and not coincident. As shown in Figure4, it shows the visualization graph after LDA clustering.

4. Summary

This paper captures the song comments in an album on Netease Cloud Music, analyzes and visualizes the comments to help audience quickly understand song popularity, theme, and emotion of the album. The song named Boundless Universe is the most popular song in the Treasure Hunting, with the theme of peace and children. Artists has the most comments among non-title tracks, expresses the most negative emotion in the whole album no matter in the comments or in the lyrics, because its expression contains satire on social evils. Therefore, when audience tend to listening this album, it would be better if the platform first recommended these two songs.

References

[1] Aggarwal C C, Zhai C X. An introduction to text mining. Mining text data. Springer, Boston, MA, 2012: 1 - 10.

[2] Pan Xiaoying, Chen Liu, Yu Huimin, Zhao Yizhe, Xiao Kangning. Survey on research of topic crawling technique. Application Research of Computers, 2020, 37 (4): 961 - 965, 972.

[3] Diouf R, Sarr E N, Sall O, et al. Web scraping: state-of-the-art and areas of application. 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019: 6040 - 6042.

[4] Gao Xiang, Li Bing. Research on method to detect reduplicative Chinese short texts. Computer Engineering and Application. 2014, 50 (16): 192 - 197.

[5] Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, Xu Sun. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation. Arxiv. 2019.

[6] Wei W. Survey on Chinese text sentiment analysis. Journal of computer applications, 2011, 31 (12): 3321.

[7] Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 2014, 5 (4): 1093 - 1113.

[8] Karim M, Das S. Sentiment analysis on textual reviews. IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2018, 396 (1): 012020.

[9] Blei, D. M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3 (Jan), pp.993 - 1022.

[10] Jelodar H, Wang Y, Yuan C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 2019, 78 (11): 15169 - 15211.