Project Report - The stories behind Danmaku and anime data

Runzhe Zhan

Group G, CISC7201 Introduction to Data Science Programming
University of Macau
mb95543@um.edu.mo

1 Introduction

1.1 Background

Danmaku comment, which originated in Niconico, has been widely applied in Chinese video websites like AcFun, Bilibili and iQiyi in recent years. As the mechanism of Danmaku allows viewers to anonymously post marquee comments associating with the specific frames, Danmaku itself contains rich frame-related information such as feedback and emotion of video viewers. With the feature of Danmaku, it could expand the dimension for analyzing the users' interaction with video content which is advantageous for video producers and content providers.

1.2 Project Description

This project mainly focuses on the Danmaku data and other accessible data of the anime produced by Kyoto Animation from Bilibili, and tries to provide an easy-to-access form for Danmaku data based on data processing and visualization techniques. Besides, extra data and works that would facilitate analyzing the anime and Kyoto Animation also involved in this project. Subsequently, the analysis of the data would explore questions from three perspectives: 1) From a macro point of view, how does the Danmaku data and other information reflect the user group's views on the target anime? 2) Whether the composition of Danmaku can be decomposed and what factors will cause the difference in the variation of Danmaku. 3) Try to use Danmaku data and extra information to conduct a case study for the anime produced by Kyoto Animation.

On the other hand, as shown in Figure 1, the programming work of this project could be divided into three parts: gathering data, data visualization, and auxiliary computation. All operations related to data linked by one pivotal - MongoDB.

2 Data Collection

The data collected from Bilibili contains Danmaku data and detailed information of the anime which is authorized to play in P.R China and Taiwan produced by Kyoto Animation. A little challenge is that Bilibili doesn't provide the documentation of APIs for ordinary developers. So it took time to figure out the address and requesting parameters of the API with the help of network packet analyzing tool which is a built-in function for the most of browsers. Once getting the pattern of APIs, the rest of work would be done using the request library by constructing the fake headers and necessary cookies. A large proportion of data use semi-public APIs to collect, the traditional crawler also has been implemented to get additional data of each anime from Bilibili, Bangumi.tv and MyAnimeList for better conducting the analysis from other aspects. The Table 1 shows the API used in the project.

There is one thing needs to be clarified that the pool of Danmaku has the threshold number from $1000 \sim 6000$, and it depends on the popularity of different videos. That is to say, the historical Danmaku would be

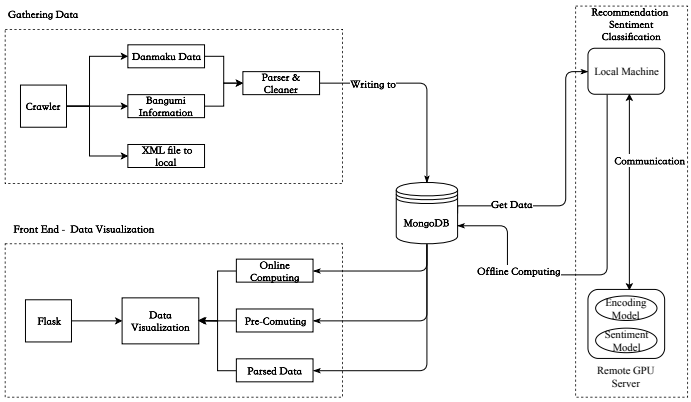

Figure 1: Overview of the project

hidden for the public whereas only the recent Danmaku would be shown in the current Danmaku pool. Calling the Danmaku API would get the current Danmaku pool data instead of all Danmaku data for some hot or long-standing videos. However, the historical API is still open for the verified users but has upper bound for the number of requesting each day. In this project, the historical data would be used only for the anime mentioned in case study part due to the limitation of historical Danmaku API.

API	Address	Notes
Current Danmaku pool	comment.bilibili.com/[].xml	Open
Historical Danmaku	api.bilibili.com/x/v2/dm/history?type=1&oid=[]&date=[]	Verification
Video information	api.bilibili.com/archive_stat/stat?aid=[]	Open
	api.bilibili.com/x/web-interface/view?aid=[]&cid=[]	Open
Bangumi Information	api.bilibili.com/pgc/web/season/section?season_id=[]	Open
	www.biliplus.com/api/bangumi?season=[]	Open

Table 1: The APIs for collecting part of the data

3 Data Processing

3.1 Data Cleaning

Before writing data to MongoDB, the raw data of Danmaku is organized in XML format needs to be parsed using BeautifulSoup library. The information provided by Bilibili including time position, Danmaku type, font size, color, sending timestamp, Danmaku pool, hashed user ID, Danmaku id and Danmaku text. One of the features of Danmaku data is highly noisy caused by the casualness of users, therefore some different Danmaku may refer to the same meaning. The problems may come from the following points while doing statistical tasks:

1) The Chinese punctuation marks in Danmaku may separate the words into different combinations. For example, “吹响吧，上低音号！” and “吹响吧！上低音号” (Sound! Euphonium) should be viewed as one Danmaku. More specifically, the regular expression would be used to remove these punctuation marks.

2) The form of characters. The Danmaku constructed by repeated characters and different case form could be considered as the same Danmaku such as “233” vs. “2333333” and “AWSL” vs. “awsl”.

3) Cross-lingual phenomenon. As the biggest ACG community in Mainland China, the language of Danmaku in anime domain also includes the parallel language pair of Japanese and Chinese which mainly is the transcription of a specific role in the anime. For example, the Danmaku language pair “わたし、気になりますが” vs. “我很好奇” refer to the same meaning (I’m curious) collected from one anime Hyouka. Furthermore,

some homophonic language pairs like “滑滑蛋” vs. “ふわふわタイム” (cozy time) are more thorny to process compared to the previous situation. At the same time, it’s the special feature you could observe from Danmaku data collected from Bilibili which is different from other video websites in Mainland China (except AcFun). This is an issue that is currently **not resolved** by this project. A possible way to solve this multilingual problem (not limited to Japanese, the Danmaku may contain other languages) is using the machine translation to set a suitable pivotal language such as English then translating the raw Danmaku to this pivotal language may get the closer expression in one language form. But the homophonic language pairs case still won’t be solved in this way.

3.2 Sentiment Classification

For training the sentiment classifier, the main problem is lacking well-labeled in-domain data (Chinese Danmaku). The alternative choice is computing by calling the existing API (like Alibaba Cloud NLP), it’s efficient but expensive since one million records in the database. The solution of this project is conducting transfer learning using pre-trained model NEZHA based on BERT [1] and related-domain data weibo_senti_100k[2]. There are many Chinese sentiment resources but the reason of choosing this datasets is considering the relationship of language features between Bilibili and Weibo. Both of them are the SNS community of younger Chinese, and the usage of language is similar to each other, like short text and network-specific language. For other sentiment corpora may come from different domains (hotel comments, e-commerce comments and movie reviews) or other SNS community (ptt.cc, LIHKG) haven’t been taken into consideration due to the difference of language feature and cultural factors. On the other hand, the emotion of some irregular language like “awsl” and “233” would be calculated using a set of rules. But there’s another limitation of this project.

The fine-tuning process as shown in Figure 2, the observation from the result is fast convergence and high accuracy. There may be no doubt that the fast convergence phenomenon because the BERT-series model has been learned high dimensional language features leading to fast adaptation of the binary classification task. The **weird** thing is high accuracy during evaluating which is the point Dr. Ryan asked after presentation. After carefully checked the data for training this model, the possible reason is the quality of weibo_senti_100k dataset. Most training data contains emoji icons in the sentence like “太古汇的迪士尼展... 吃过午饭就跑去拍了 [抱抱] 各种米奇，各种欢乐 [哈哈]”[3] labeled as “Positive” and “... 新闻前几天刚被骂倾向反政府集会呢，今天就又被对立方骂，中立说事实的媒体记者成了里外不是人了。[衰][衰][衰][泪][泪][泪]”[4] labeled as “Negative”. Whether the model learned the language features of emotion or just the symbols of emotion is still a question worth exploring.

3.3 Tag-based Recommendation

Since the exact user ID has been hidden due to hash encryption, a tag-based recommendation algorithm has been proposed to recommend the similar anime in this section. The key idea is using pre-trained model as a language model to encode the tag of each anime to get the vector representation of each anime. Given the n encoded tags $Embed(T_i)$ of one anime which are M-dimensional vector (default $M = 768$), taking the averaging of these vectors then getting the embedding of the anime A,

$$A = \frac{\sum_{i=1}^{n} Embed(T_i)}{n}$$

The subsequent task is calculating the cosine similarity for each anime pair (A, B) to construct the similarity matrix. The most time-consuming operation is encoding the different tags to a vector representation. For

[1] https://github.com/huawei-noah/Pretrained-Language-Model/. “NEZHA-base-WWM” model released in 9, December is used in this section.

[2] https://github.com/SophonPlus/ChineseNlpCorpus

[3] **Meaning:** Disney’s Show at Taikoo Hui ... I went to take the pictures after lunch, [Hug] all kinds of Mickey, all kinds of joy [haha]

[4] **Meaning:** Just a few days ago, the news was scolded as inclining to the anti-government rally. Today, it was scolded by the opposite party. The media reporters who said the truth in a neutral way have become a two-faced dissident. [bad] [bad] [bad] [bad] [wept] [wept] [wept]

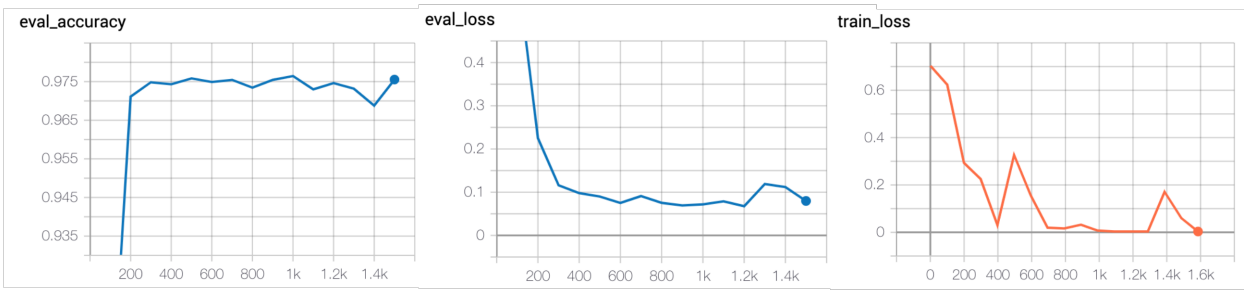

Figure 2: Fine-tuning process in weibo_senti_100k datasets exported by TensorBoard. The accuracy and loss of evaluation set is 0.9783 and 0.0687.

example, constructing 40×3593 similarity matrix would take nearly 20 minutes in GPU server with NVIDIA P100 graphics card and Intel Xeon E5-2682v4 (4 vCPU) [5]. While deploying phase, the pre-computed similarity pair would be written into MongoDB to achieve instant recommendation. In this project, the top-nine anime similar to the current anime will be recommended to users (as shown in Table 2).

$$Sim(A, B) = \frac{\sum_{j=1}^{M} A_j B_j}{\sqrt{\sum_{j=1}^{M} A_j^2}\sqrt{\sum_{j=1}^{M} B_j^2}}$$

Anime	Recommendation results
CLANNAD ~After Story~	CLANNAD -クラナド-, CLANNAD もうひとつの世界智代編
AIR	夏目友人帳肆, 夏目友人帳
涼宮ハルヒの憂鬱	デュラララ!!×2 承, 涼宮ハルヒの消失
日常	小林さんちのメイドラゴン, 干物妹！うまるちゃん R

Table 2: Examples of recommendation results

4 Data Visualization

As shown in Figure 3, the visualization part was mainly implemented by Flask front-end framework and eChart.js library. By creating the query API, all data required for visualization will be extracted from MongoDB then transferring to the front end using JSON as the data carrier. The front-end page will request these APIs for the purpose of rendering the visualization graph asynchronously. Another point is for saving the computing power and speeding up the response time, the task needed high computational resources would be pre-computed rather than online computing which has been mentioned in the previous section.

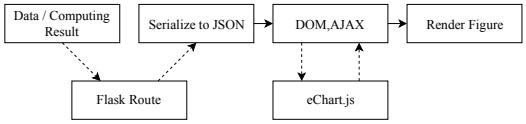

Figure 3: The flow the data visualization

[5] Alibaba Cloud Instance Type ecs.gn5-c4g1.xlarge

5 Data Analysis

5.1 Macro Perspective

From the macro perspective, the first thing is there is no strong relationship between rating[6] and the number of viewing. Although the new anime in recent years have high popularity, they haven't received high ratings. For example, the anime whose the number of play and the number of Danmaku ranked in the top-three even not enter the top-ten list of the ratings. Besides, there are some differences between the rating given by Chinese (mainly from Mainland) and western audiences. For example, "響け！ユーフォニアム 2" owns a higher rating over eight points among Chinese viewers but only got 6.57 points from western viewers. Especially, for users from Bilibili, they tend to give higher ratings than other sites as shown in Table 3. Generally speaking, the anime produced by Kyoto animation appear polarized comments vary from works to works due to their exploration in a different genre. Nevertheless, "CLANNAD ～ After story～ " is a widely praised anime worldwide which would be discussed in section 5.3.

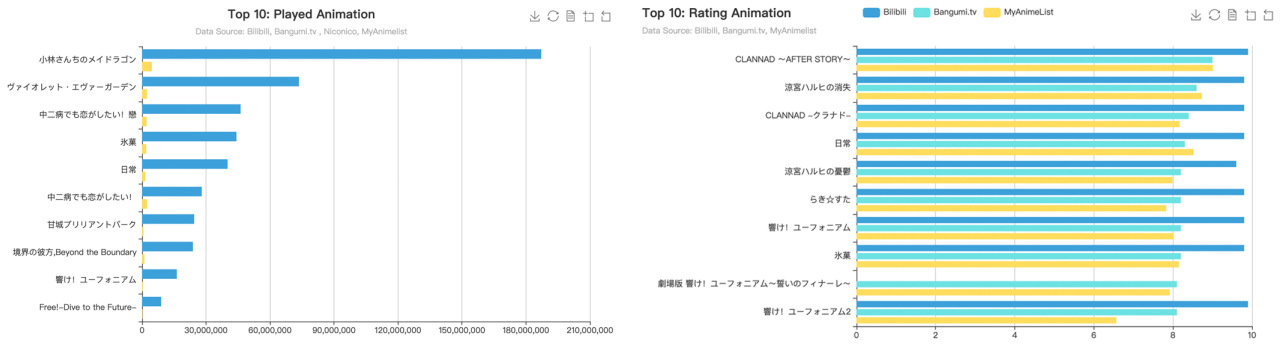

Figure 4: The top-10 popular anime vs. top-10 rating anime

Websites	Mean	Std.	Median
Bilibili	9.66	0.20	9.7
Bangumi.tv	7.45	0.76	7.4
MyAnimelist	7.86	0.49	7.86

Table 3: Rating statistical result - Kyoto Animation

5.2 Social Factors

This section may try to explore how will the Danmaku change under the influence of social factors. It is easy to get from the Figure 5(a) picture that the sending time of Danmaku is uneven and most of Danmaku sent in some specific time intervals. Given that the vast majority of the audience is students or young people, the number of Danmaku sent during the working or studying hours is far less than their leisure time. It could easily predict that which time the server would face the state of high concurrent requests. At the same time, Kyoto Animation encountered the arson attack few months ago. After this incident, the number of Danmaku of anime produced by Kyoto Animation grew sharply (as shown in Figure 6(b)), indicating that the social topics may have the leading effect of the increment of Danmaku. More specifically, if we visualize the Danmaku containing "京阿尼"(Kyoani, refers to Kyoto Animation) using wordcloud (as shown in Figure 6(a)) within three days of this accident, the content of Danmaku is also highly related to the event.

[6]Bangumi.tv is the representative anime review website in Mainland China whereas MyAnimelist could represent the views from western countries.

If we change the target to all one million Danmaku, the composition of the Danmaku is mainly divided into two parts: one is the hot Internet-specific words such as “awsl”, “2333” and “xswl”, the rest are mostly representative symbols or lines from the characters in the anime (like “我很好奇” from 氷菓 ,“k-on” from K-ON! and “我不高兴” from 境界の彼方). This proves that the composition of anime Danmaku is not only influenced by social factors, but also by the network culture and the content of anime itself.

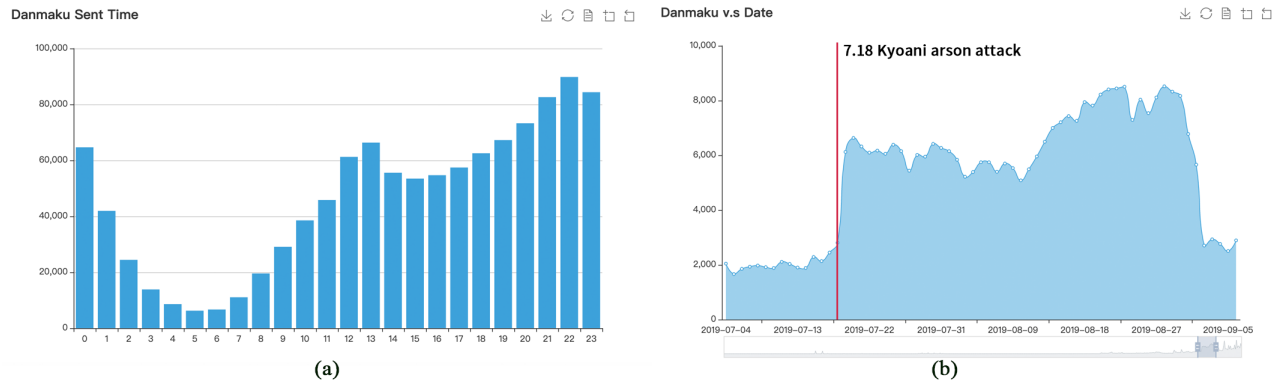

Figure 5: (a) The time distribution of sending Danmaku. (b) The number of Danmaku influenced by arson attack.

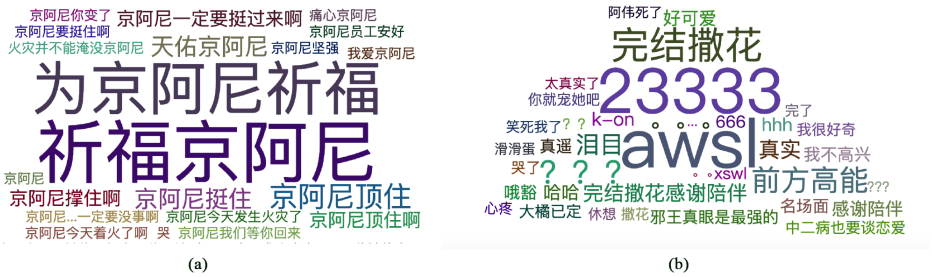

Figure 6: (a) Danmaku contains “Kyoani” from 18 July to 20 July, 2019. (b) Recent Danmaku wordcloud.

5.3 Anime Case Study

5.3.1 Interaction effects

A good story or script that would universally receive high ratings may not neglect the control of the emotions of a wide range of audiences. The effect of emotional control may be directly reflected through the Danmaku. The analysis in this part would be conducted by analyzing the visual results of the sentiment[7] and the number of Danmaku at each time point (all episodes) of the anime “CLANNAD ∼ After story ∼”.

It could be observed that the audience’s emotions are in a positive state most of the time, but the negative emotion still distributes in different time from the beginning to the end. However, there are few points whose value is extreme which proves that the producers don’t lose control of the audience’s emotions. For this anime itself, it’s a deeply moving anime with both comedy and tragedy which can’t completely prevent the audience from generating negative emotions. So from the perspective of the audience’s emotions, the producer well controlled the subtle balance between positive and negative emotions.

[7]The value of emotion is the average value of all Danmaku in each second.

From the perspective of sustained effect, maintaining the sustained effect is helpful to make the audiences sticky. This effect in an anime that 24-minutes episodes are that constantly create a bridge plot for the audience which is a kind of stimulation. This effect could be reflected in the number of Danmaku at different time points. As shown in the Figure 7(a), the number of Danmaku would appear a peak value in a certain period of time and then decline. The stimulation to audiences acts like a sine function that could be clearly observed from the change of the number of Danmaku at the minute level(as shown in Figure 7(b)). At the same time, another phenomenon that can reflect the user's preference is that the number of Danmaku at the beginning and the ending is far higher than the middle part of the story narration, which means that only a small proportion of the audience is willing to participate in the Danmaku discussion. In other words, this phenomenon of "shy" viewers may be observed in Figure 4 that there is a significant difference between the amount of playing (blue bar) and the number of Danmaku (yellow bar). Although BiliBili is one of the first video websites owing Danmaku function in Mainland China, only a few users feel free to join in the Danmaku discussion.

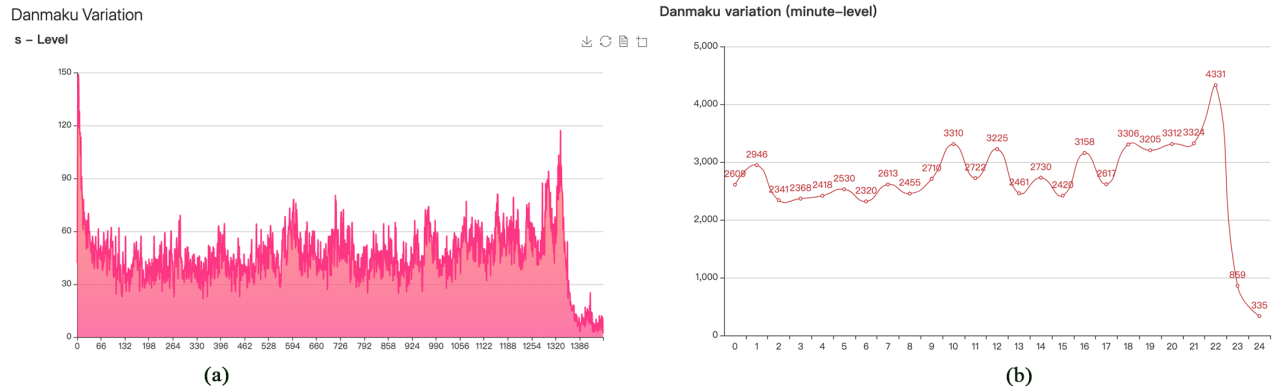

Figure 7: (a)The Danmaku variation of "CLANNAD After Story" (second level). (b)The Danmaku variation of "CLANNAD After Story" (minute level).

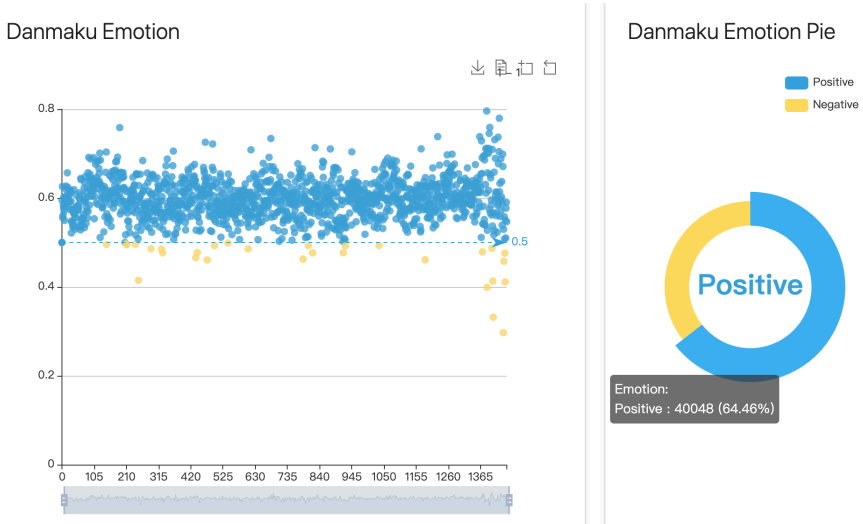

Figure 8: The emotion distribution of Danmaku of "CLANNAD After Story"

5.3.2 Genre variation

As an animation studio with a high reputation, Kyoto animation has always been an industry benchmark and continues to explore different subjects of the anime. The Figure 9 shows the variation of the genre by pie charts calculating from the anime tags. This section also could be considered as the supplementary part echoes to section 5.1. From 2003∼2008, Kyoto Animation had a close cooperative relationship with Key studio, and the anime-series adapted from the game which is called "Key Trilogy" made great success and laid a solid foundation for future development of Kyoto Animation. Thanks to the excellent and touching story of the game itself and the existing foundation of users, the anime which received high ratings mainly produced in this stage. With the launch of the anime 涼宮ハルヒの憂鬱 (The Melancholy of Haruhi Suzumiya) which is the first anime based on adaptation from light novel, the Kyoto Animation stepped to the new phase. The works in this stage are basically adapted from light novels, and the theme is closer to daily life and campus life. These five years'(2009∼2014) works not only make Kyoto animation gradually gain a large amount of popularity, but also have a significant impact on the entire animation industry. In the past five years, Kyoto animation began to launch a series of exploratory original anime based on the summary of past experience and production flows. The form is becoming diverse which not limited to light novel adaptation. However, some works are not fully recognized by viewers in the process of exploration, resulting in a huge standard deviation of the score as mentioned in the previous section 5.1.

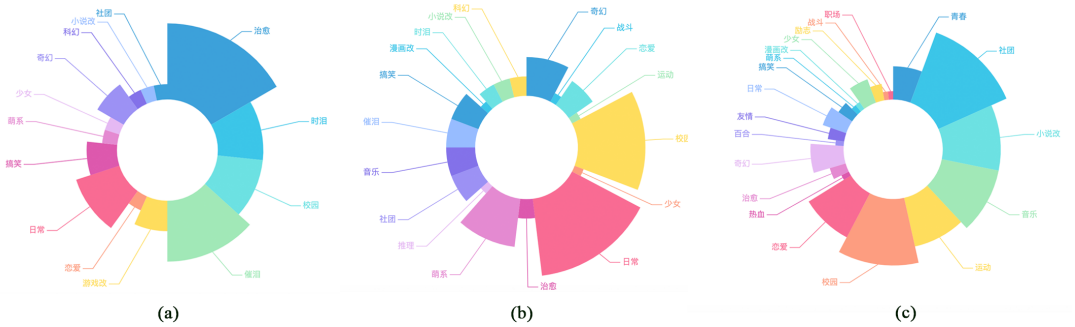

Figure 9: (a)Tag information from 2003∼2008. (b)Tag information from 2009∼2014. (c)Tag information from 2015∼2019

6 Conclusions, Future Works and Implications

This project driven by interest mainly explores the feasibility of processing and visualizing the Danmaku data, and tries to conduct a case study through the Danmaku and additional animation data. Since the sentiment classifier model trained in this project is still imperfect, the future works would keep exploring how to build the sentiment classifier for this domain lacking well-labeled data. There's no denying that the data of this project may not be as valuable as that of some datasets with strong practical significance. But it still has a certain significance for some specific groups.

For the anime content providers in Mainland China, the platform like Bilibili must purchase the broadcast right from anime production companies due to copyright protection. So it is worth considering how to selectively purchase the anime would gain high popularity based on analyzing the preference of the user group on their own site. What's more, the Danmaku could reflect the user's response to the video content and can directly obtain detailed frame-based feedback, which is a valuable resource for animation companies or video producers.