
Final Project: The Stories Behind Danmaku Data

2019/12/16

Runzhe Zhan (MB95543)
Group G

Intro

#What is Danmaku ?

- Danmaku (弹幕) allows viewers to generate scrolling marquee comments on videos
- Associated with the key frame in the videos
- Anonymously, Short, Highly noisy



Source: Bilibili - av76601298, 花花与三猫CatLive

#Data Overview

- **Data Source:** Bilibili
- **Data domain:** Available anime produced by Kyoani (Kyoto Animation)
- **Data Content:** Danmaku, detailed information about each anime
- **External Data:** Bangumi.tv, MyAnimelist
- **Data Amount:** 203 MB (Mongodb dump) + 2.96G (Raw data before cleaning)



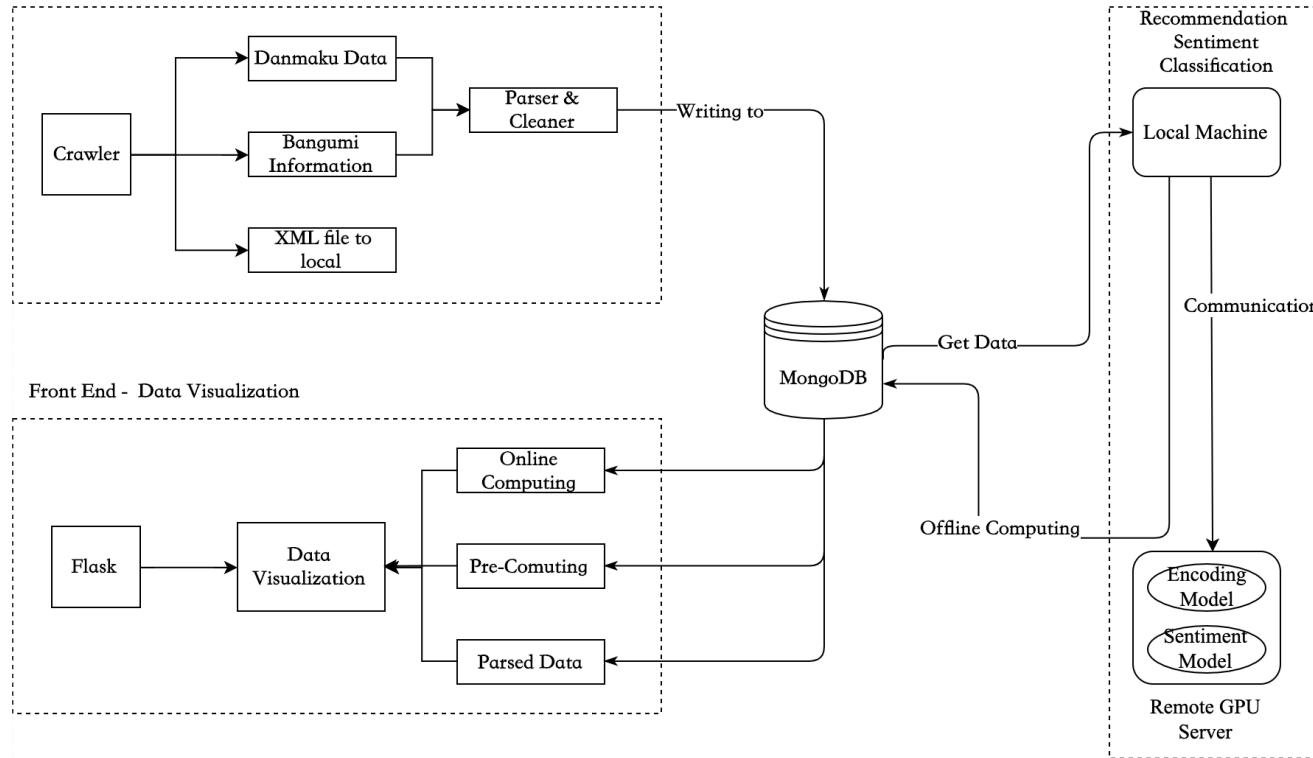
#Project Overview

- Python 3.6
- Extra Libraries:
Flask, requests, BeautifulSoups, pymongo, NLTK, tensorflow, bert-as-service
- Main contribution:
 - Collecting & Cleaning Data
 - NOSQL via MongoDB
 - Building a flask website to visualize the data
 - Sentiment Classification
 - Tag-based Recommendation
 - Analysis & Case Study

Intro

#Project Structure

Collecting Data



Main workflow

Collecting Data

- 1. Crawled by finding their API – Google Chrome, Wireshark is helpful

The screenshot shows a Bilibili video player for a short film titled "利益与青鸟 (SPs)". The video thumbnail features two characters looking out of a window with Japanese text overlaid: "リズと青い鳥 大ヒット上映中!". Below the video, there's a link to "http://liz-bluebird.com". The Google Chrome Network tab is open, displaying a list of requests. A specific request for the API endpoint `https://api.bilibili.com/x/web-interface/view?aid=44459693&cld=77838455-1-3028..` is highlighted with a red box. The "General" section of the request details is visible, showing the Request URL, Request Method (GET), Status Code (200 OK), and Remote Address (164.52.76.18:443). The "Response Headers" section shows various HTTP headers including Access-Control-Allow-Methods, Access-Control-Allow-Origin, Access-Control-Expose-Headers, and Cache-Control.

Request information

- 2. Simulate the request (Fake headers, Copy-cat Cookies)
- 3. Python is all you need !
- 4. Store to MongoDB for the subsequent phase

Main workflow

Cleaning Data

- Highly noisy data
- For example (Real cases)
 - When you do danmaku counting...

Remove Chinese punctuations: (Sound! Euphonium)

吹响吧！上低音号 吹响吧，上低音号 吹响吧，上低音号！ 吹响吧！上低音号！

Integrate repeat characters: (Hahaha)

2333 233333 23333333 哈哈哈 哈哈哈哈哈哈哈 哈哈哈哈哈哈哈

Cross-lingual cases: (Unsolved)

我很好奇 vs. わたし、気になります(I'm curious)

滑滑蛋 vs. ふわふわtime (cozy time)

- When you do sentiment classification...

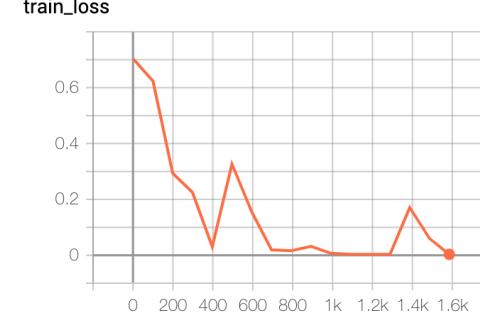
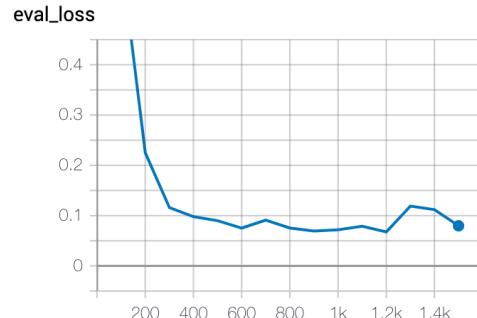
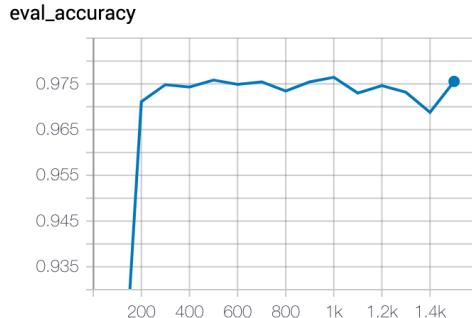
Rules for Network-specific language:

awsI = 啊我死了 (I'm dead) (Not surface meaning...)

Main workflow

Sentiment Analysis

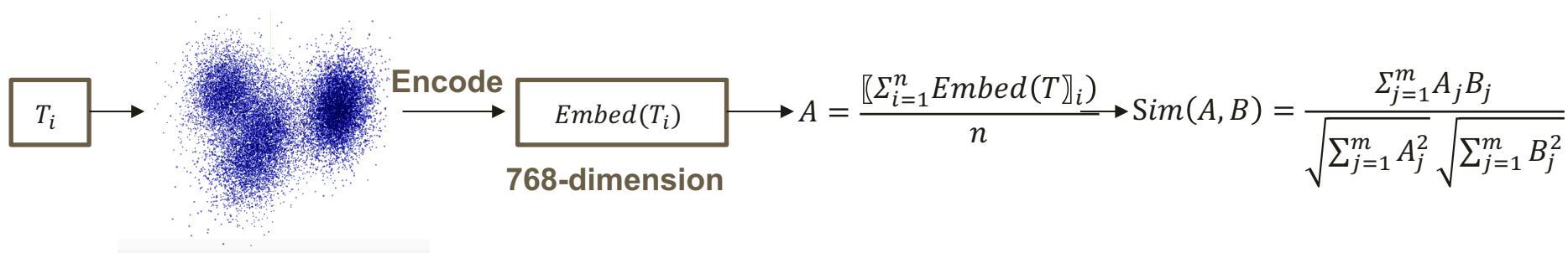
- Don't have well-labeled in-domain data.
- Existing API service is expensive while processing large dataset:
Alibaba Cloud API: 1000000 times / CNY¥ 1800
- My solution:
 - BERT-series **Pre-trained model**: NEZHA released by HUAWEI NOAH Lab
 - Use **related-domain** dataset: Weibo Sentiment
(Relationship: short text, SNS community)
 - Fine-tuning in NEZHA model (**eval_accuracy = 0.9783315, eval_loss = 0.068682075**)



Main workflow

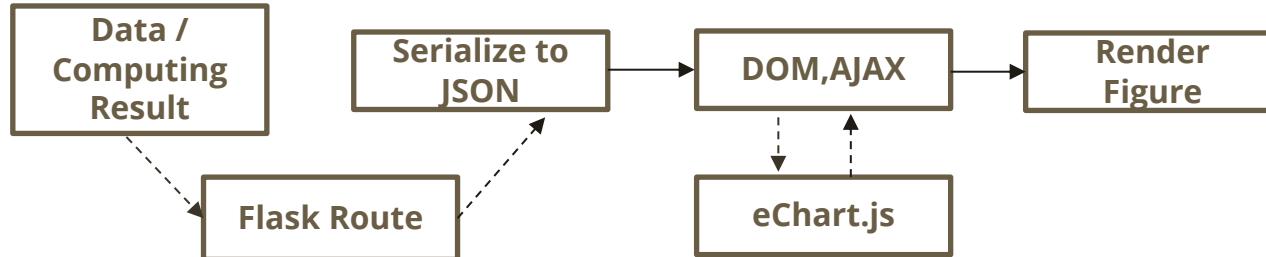
Tag-based Recommendation

- The user IDs have been **hashed** in the Danmaku records
- The staff from Bilibili have added **tags** for most anime
- My solution: tag-based recommendation
 - **Encode** each tag T_i using NEZHA pre-trained model (not fine-tuning)
 - Calculate the anime **embedding** from the tag vectors
 - Construct the **cosine similarity** matrix between the anime (except itself)
 - **Rank** the similarity then **select** to recommend



Data Visualization

Visualization



```
▼ ...  
  ▼ danmaku_emotion: [0.4313821446250432,  
    ▼ [0 ... 99]  
      0: 0.4313821446250432  
      1: 0.5759672674830951  
      2: 0.5338540791249978  
      3: 0.5431031734995324  
      4: 0.5280140904644122  
      5: 0.5275038547300634  
      6: 0.5177925816680372  
      7: 0.4999405274771248  
      8: 0.543511247371695  
      9: 0.501089525606283  
     10: 0.4623571946514534
```

JSON as data carrier

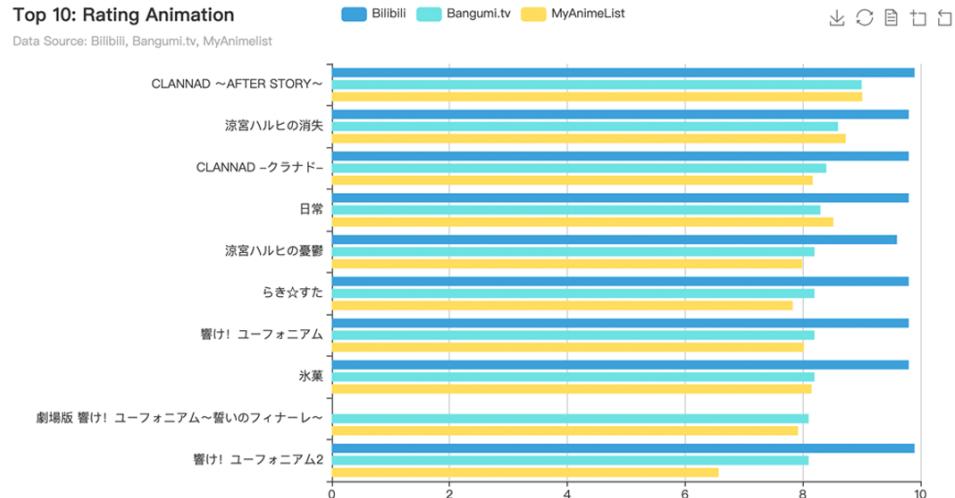
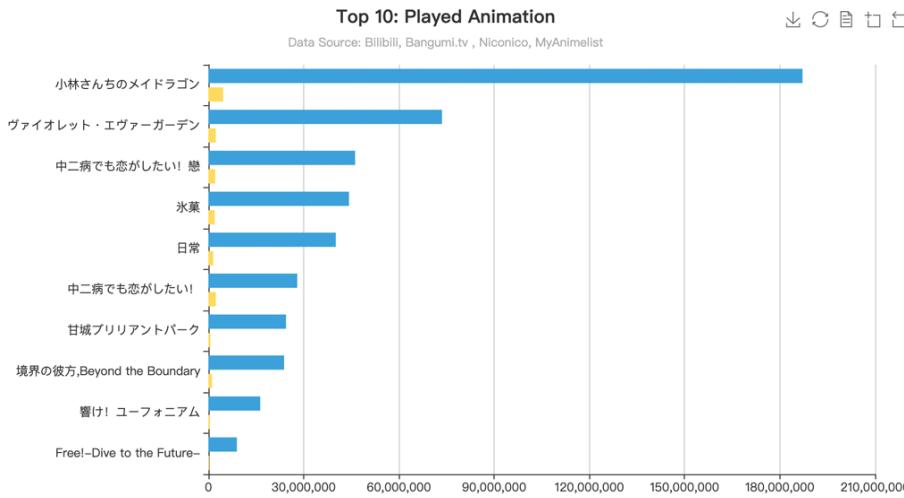


Async Render via eCharts.js

Data Analysis

Macro perspective

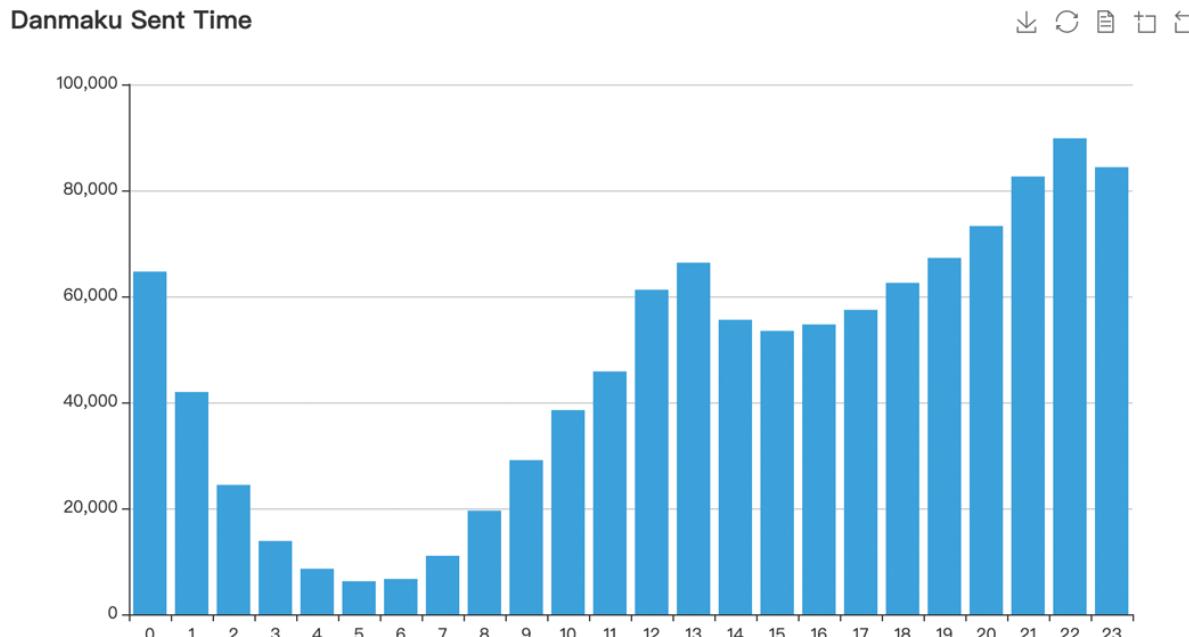
- Recent popular anime may not get good reviews
- There are **differences in preferences** between overseas and Chinese audiences
- The users from **Bilibili** likely to give **higher rating** (Top 10 anime received 9.8+ score)
- **CLANNAD ~After Story~** is widely praised worldwide



Data Analysis

Macro perspective

- The users mainly sent danmaku in 21:00 ~ 00:00
- Maybe not so many people are slacking off during working hours ...

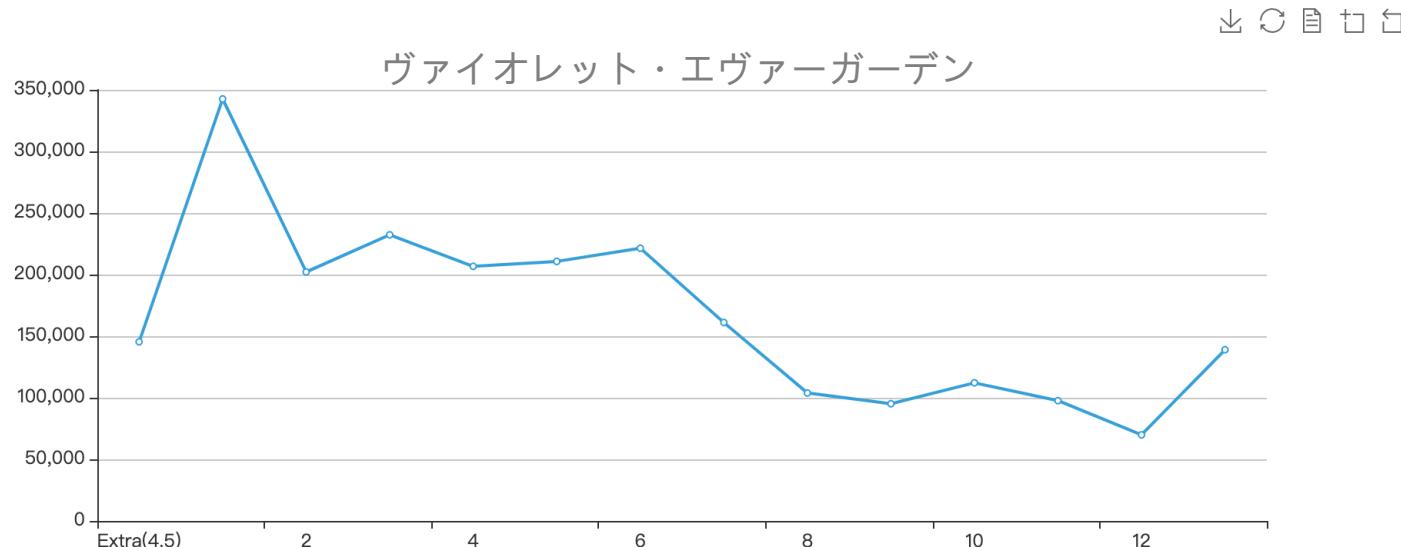


Data Analysis

Macro perspective

- Generally speaking, the number of danmaku decreases along with the number of episodes but would increase in the last few episodes

Danmaku Counting per Episode

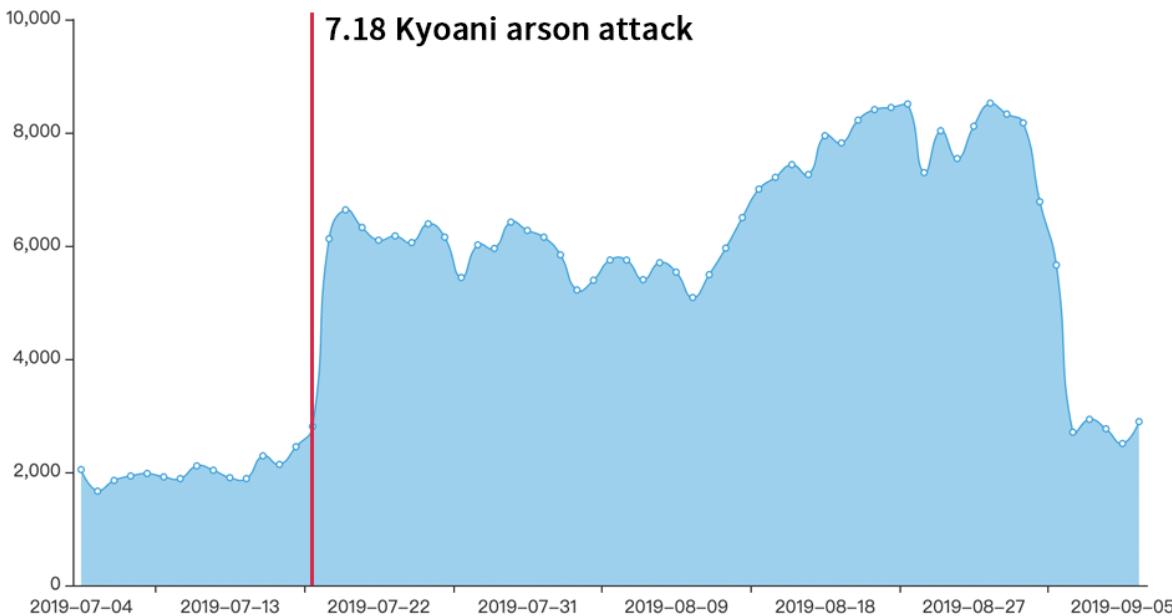


Data Analysis

Society Factors – 7.18 Kyoto Animation arson attack

- The number of danmaku grows sharply around 18 July

Danmaku v.s Date



Data Analysis

Social Factors – 7.18 Kyoto Animation arson attack

- The number of danmaku grows sharply around 18 July
- Word Cloud

为京阿尼祈福
祈福京阿尼

京阿尼
京阿尼...一定要没事啊 京阿尼今天发生火灾了 京阿尼顶住啊
京阿尼今天着火了啊 哭 京阿尼我们等你回来
京阿尼要挺住啊 京阿尼一定要挺过来啊 痛心京阿尼 京阿尼员工安好
火灾并不能淹没京阿尼 天佑京阿尼 京阿尼坚强 我爱京阿尼

Data Analysis

Top Danmaku in anime

- 'awsl' is the most frequently used danmaku in 2019
- Other hot danmaku mainly come from actor's lines

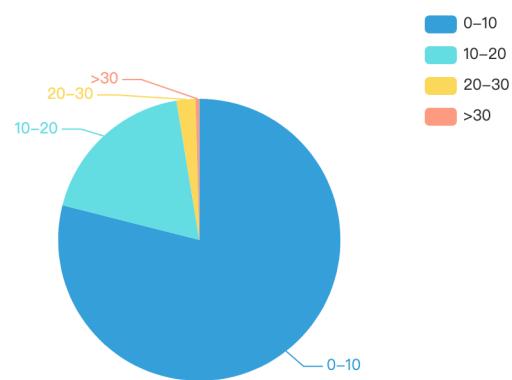
阿伟死了 好可爱
完结撒花
23333
太真实了
你就宠她吧
笑死我了? ?
滑滑蛋 真遥 泪目
哭了?
哦豁 哈哈 完结撒花 感谢陪伴
心疼 大橘已定 休想 撒花 邪王真眼是最强的
名场面 感谢陪伴
中二病也要谈恋爱
完了
666 hhh 我很好奇
真实 我不高兴
前方高能???

Data Analysis

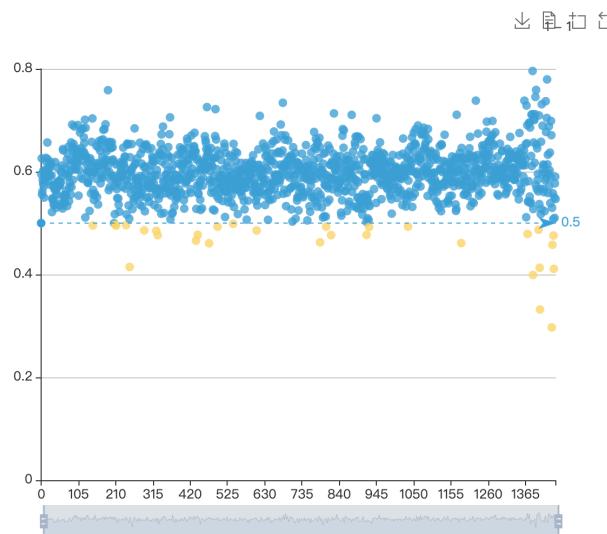
Case Study: CLANNAD ~After Story~

- 'Short' length is the feature of danmaku
- Well-controlled the balance between positive and negative emotion

Danmaku Length Distribution



Danmaku Emotion



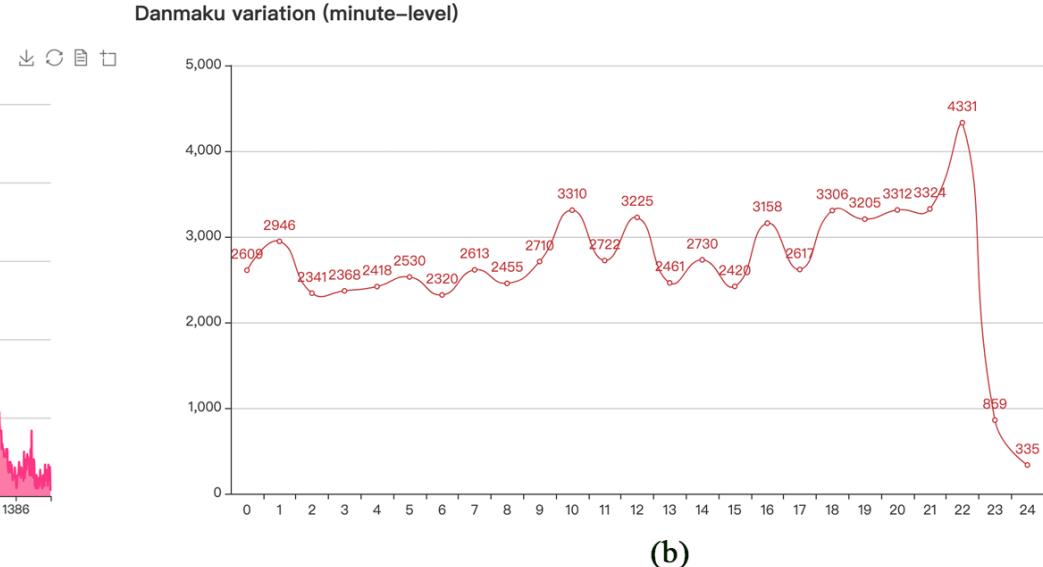
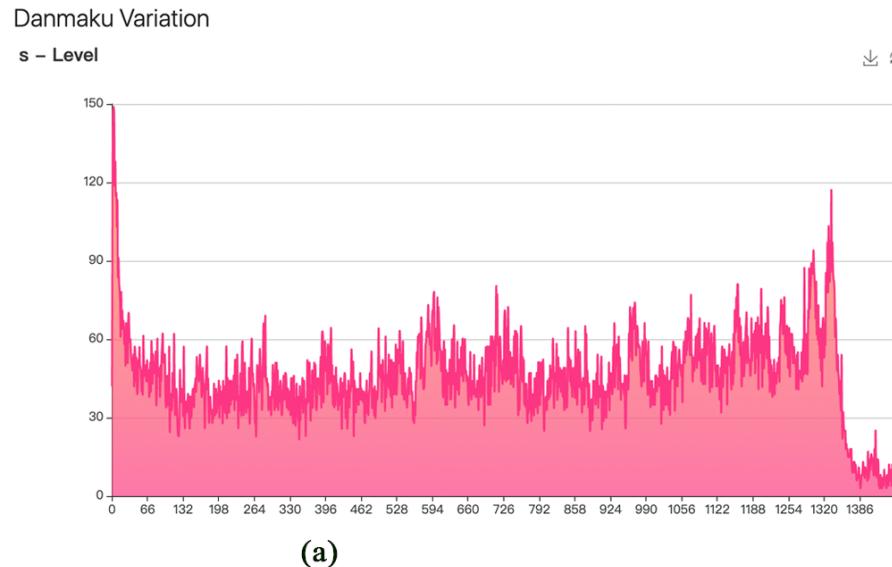
Danmaku Emotion Pie



Data Analysis

Case Study: CLANNAD ~After Story~

- Good stories will constantly stimulate the audience's feelings in each episode
- The stimulation acts like the "SIN" function



Data Analysis

Case Study: CLANNAD ~After Story~

- Use the danmaku data collected by myself many years ago
- The episode people pay more attention is different compared to previous years

To compress the size of file, Please refer to:

<https://drive.google.com/open?id=1F6SZixRKOrpaISXuLddLpF6tGGuuqsqR>



Explore your story...

bangumi.imzhezhe.com