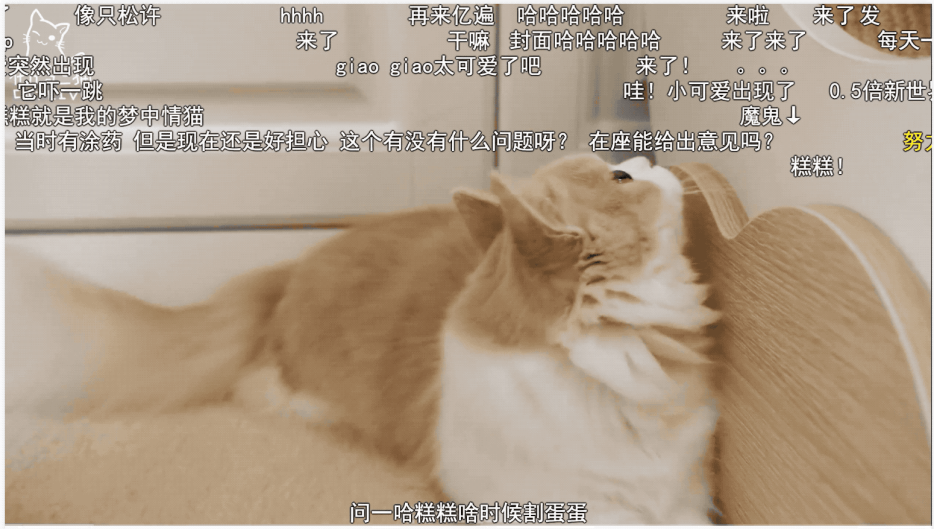
Final Project:
The Stories Behind Danmaku Data

2019/12/16

Runzhe Zhan (MB95543)
Group G

#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

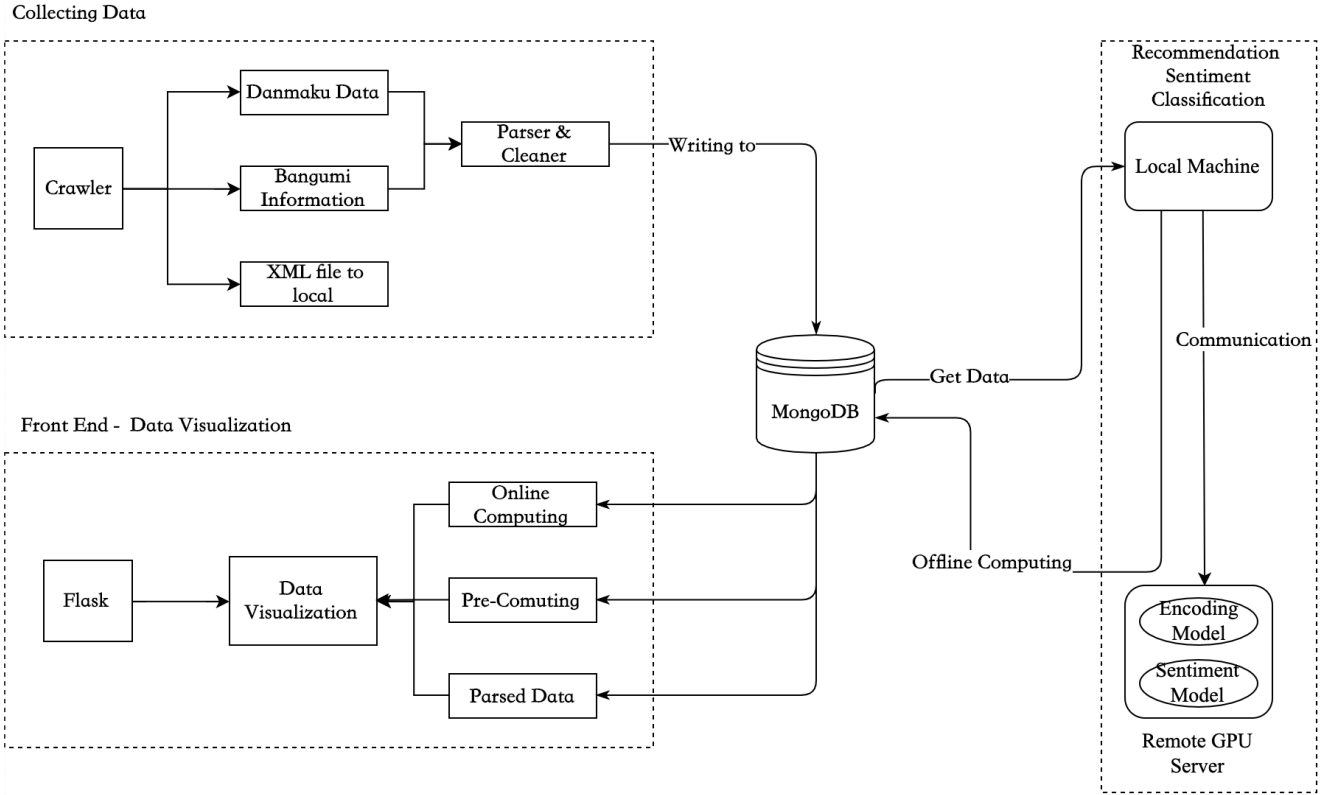
- **Data Source:** Bilibili
- **Data domain:** Available anime produced by Kyoani (Kyoto Animation)
- **Data Content:** Danmaku, detailed information about each anime
- **External Data:** Bangumi.tv, MyAnimelist
- **Data Amount:** 203 MB (Mongodb dump) + 2.96G (Raw data before cleaning)

#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

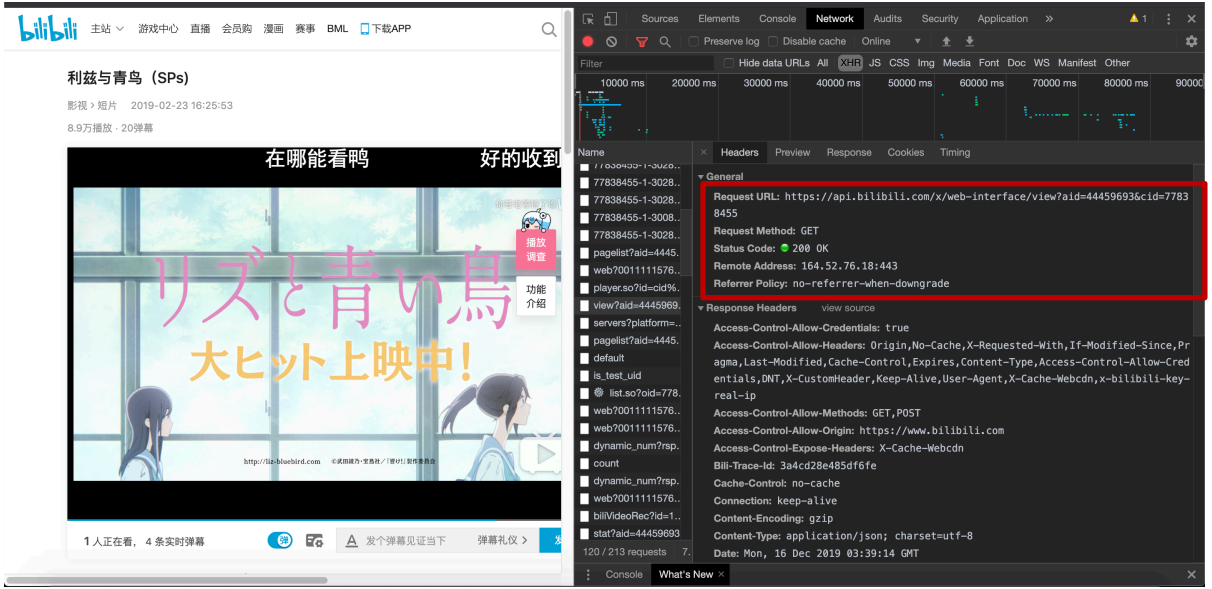
#Project Structure



Main workflow

Collecting Data

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



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

    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:

    awsl = 啊我死了 (I'm dead) (Not surface meaning…)

Main workflow

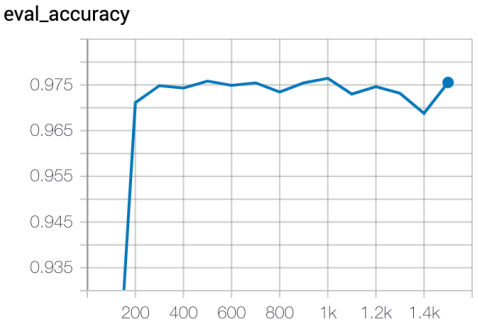
Sentiment Analysis

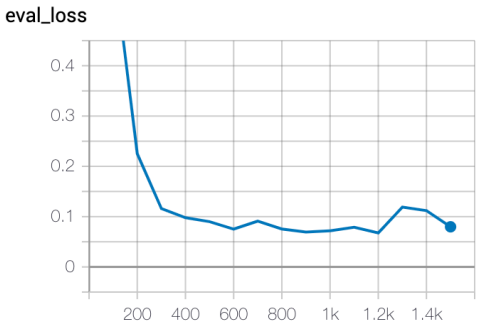
- Don't have well-labeled in-domain data.
- Existing API service is expensive while processing large dataset:

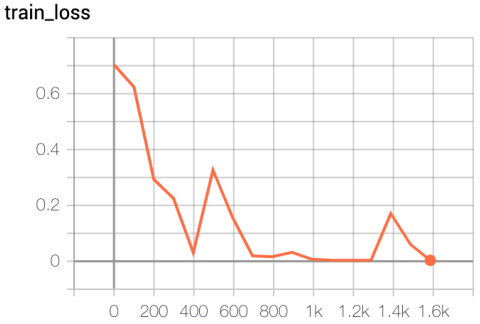
  Alibaba Cloud API: 1000000 times / CNY¥ 1800

- 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)



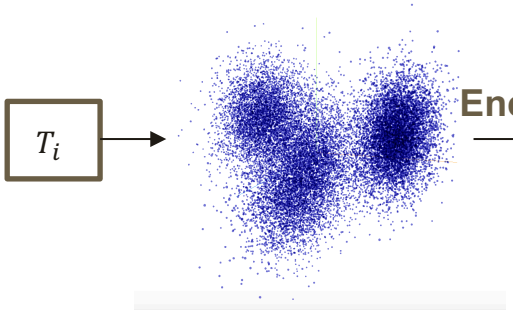




Main workflow

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

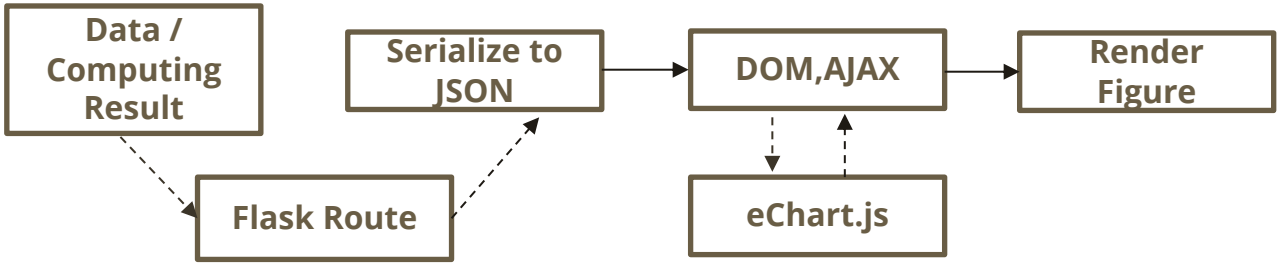


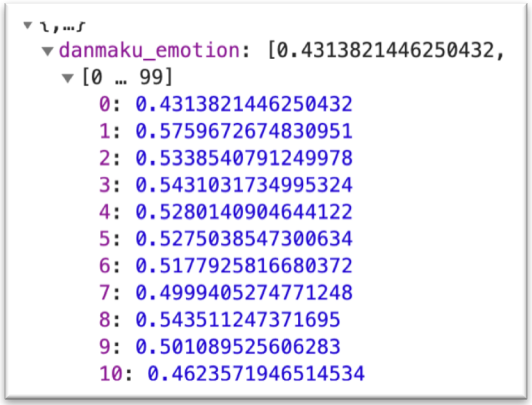
$$A = \frac{[\![\Sigma_{i=1}^{n} Embed(T)]\!]_i}{n}$$

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

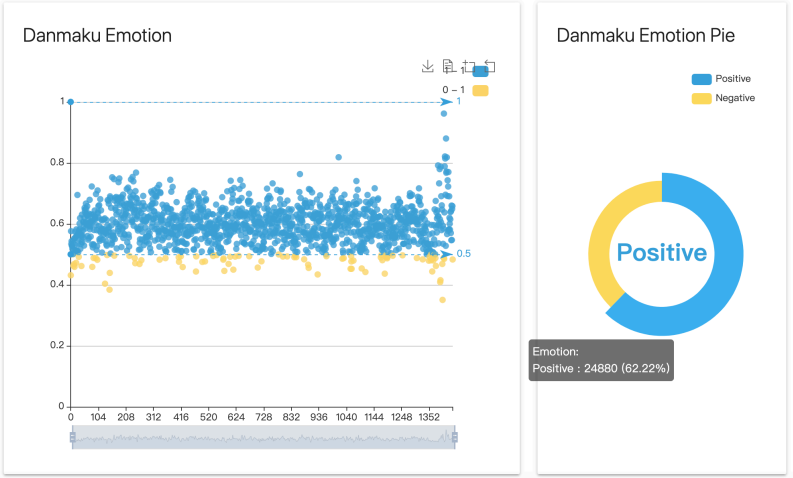
Data Visualization

Visualization





JSON as data carrier

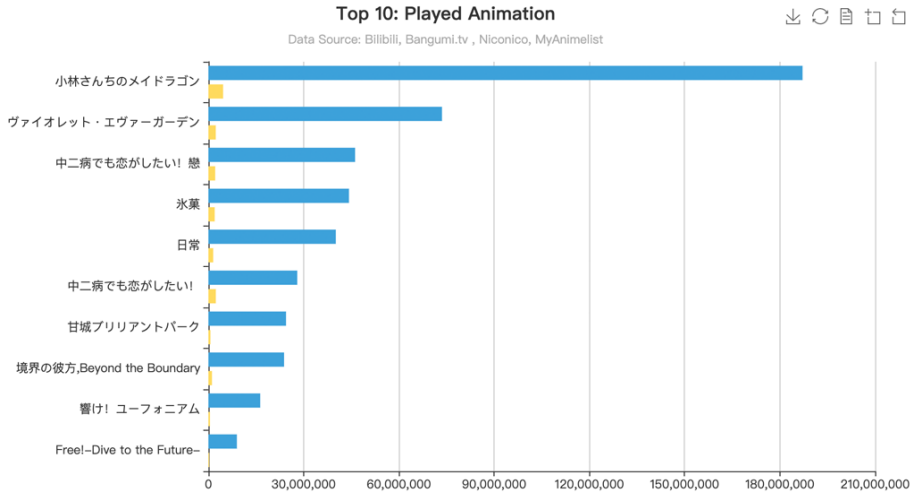


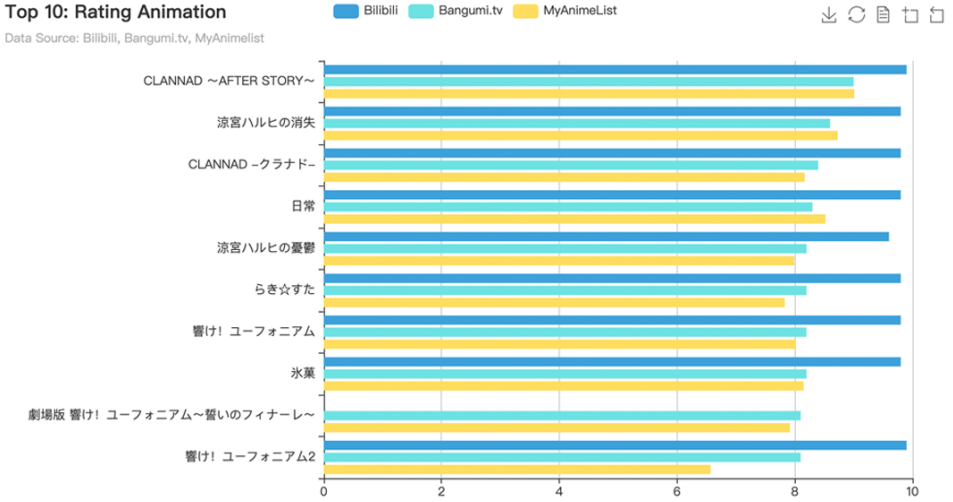
Async Render via eChats.js

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

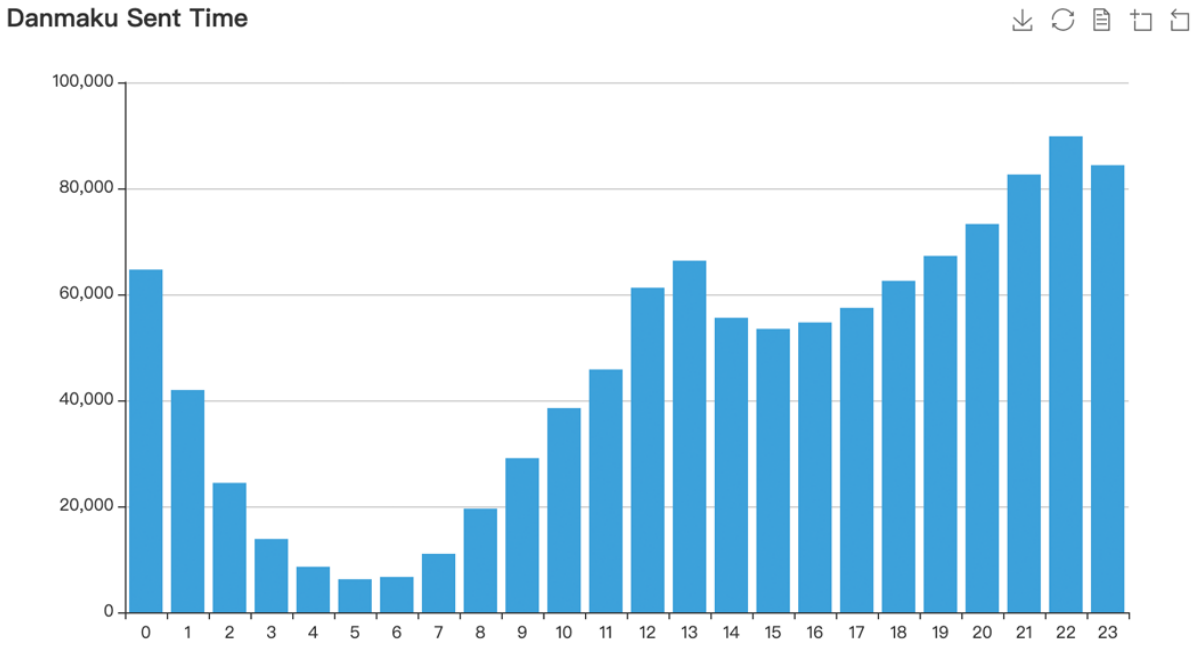




Data Analysis

Macro perspective

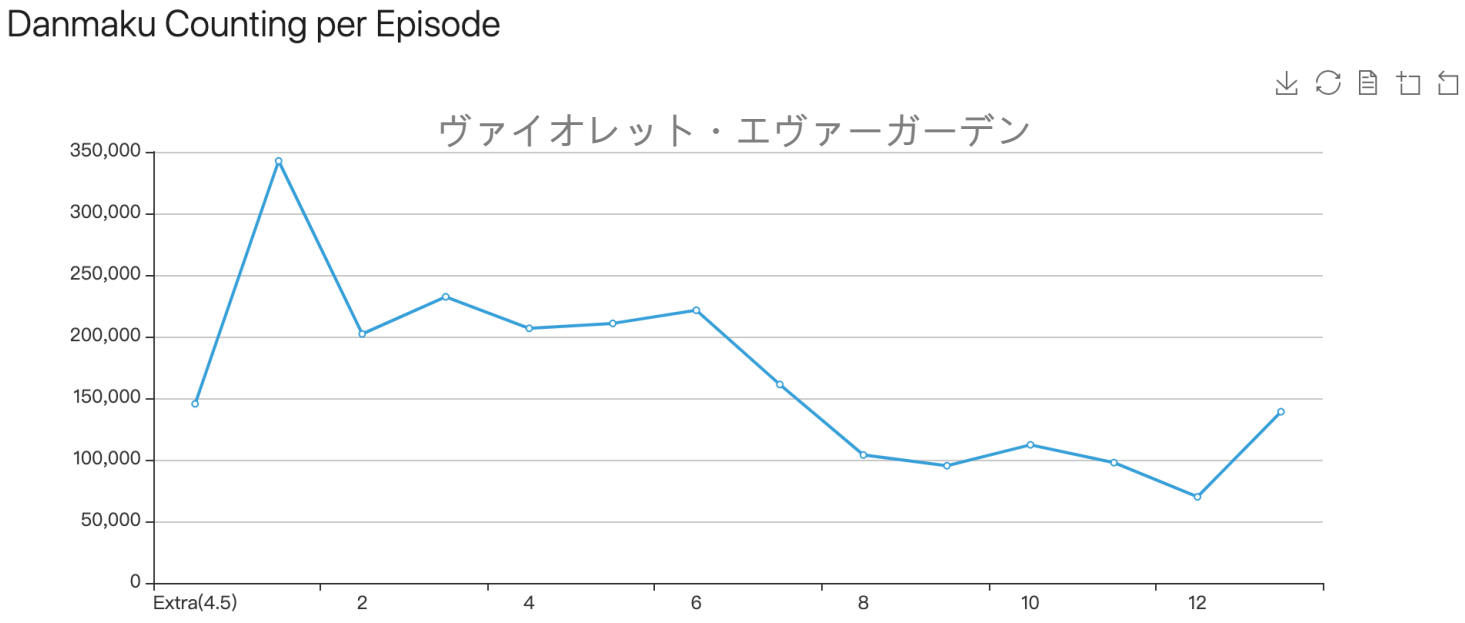
- The users mainly sent danmaku in 21:00 ~ 00:00
- Maybe not so many people are slacking off during working hours ...



Data Analysis

Macro perspective

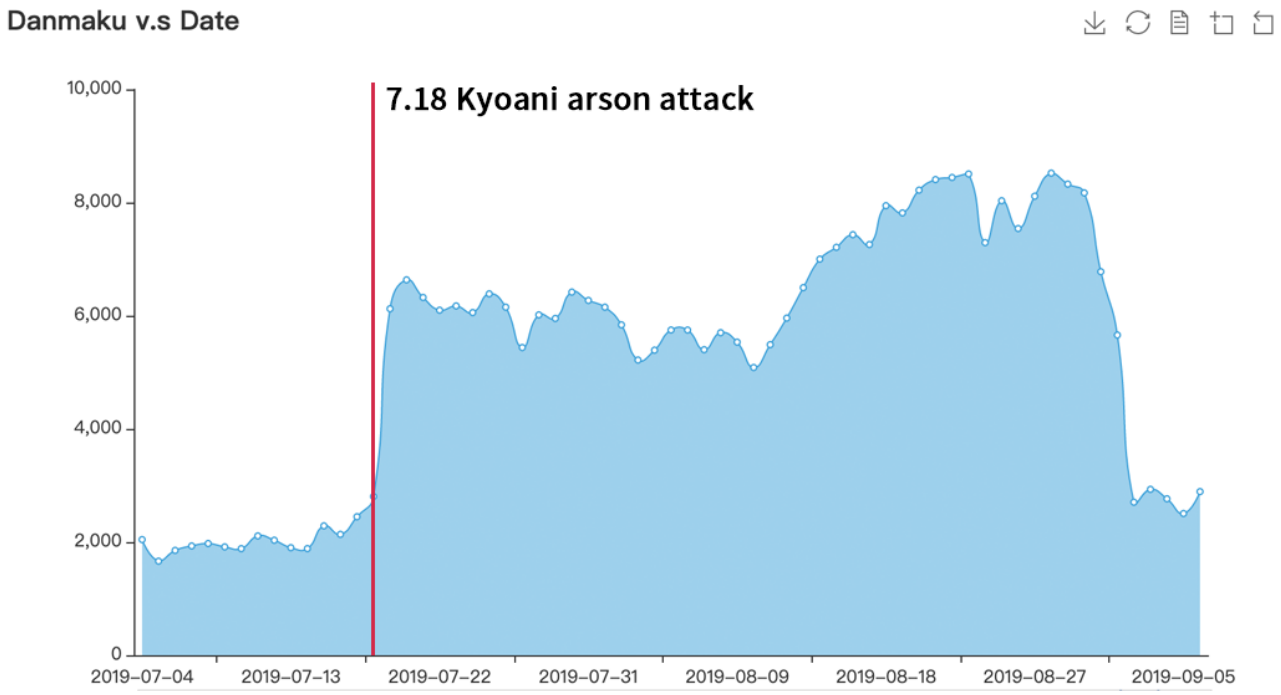
- **Generally speaking, the number of danmaku decreases along with the number of episodes but would increase in the last few episodes**



Data Analysis

Society Factors – 7.18 Kyoto Animation arson attack

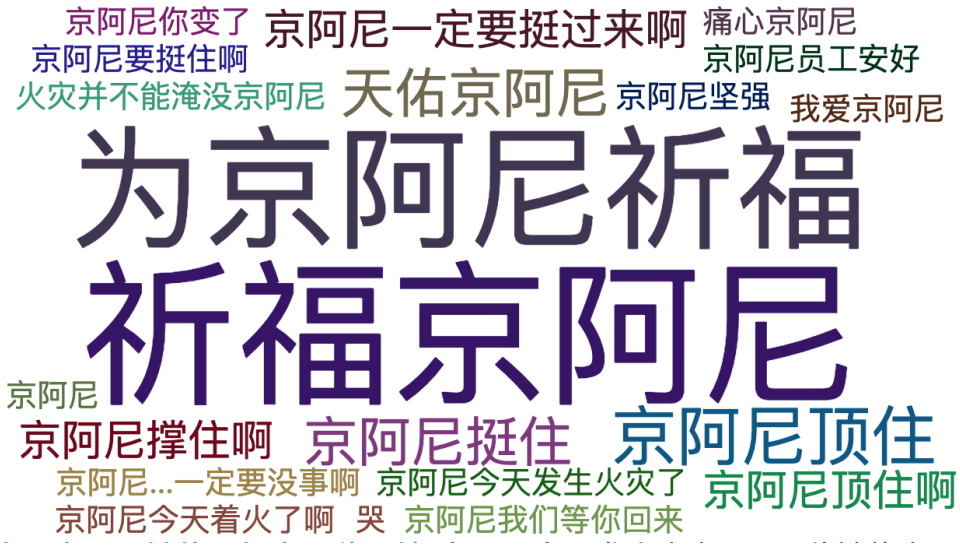
- The number of danmaku grows sharply around 18 July



Data Analysis

Social Factors – 7.18 Kyoto Animation arson attack

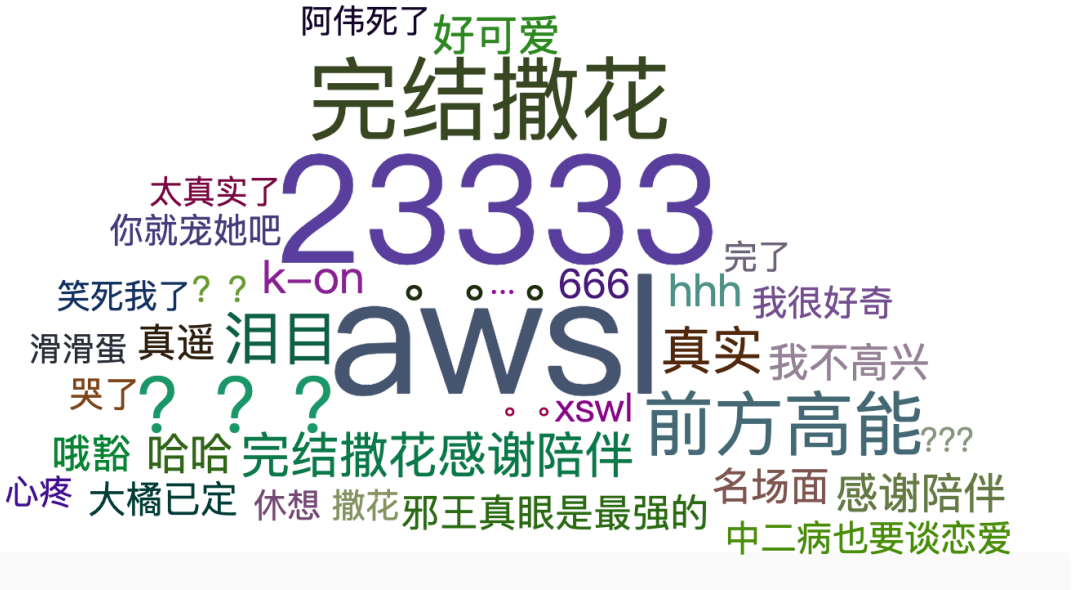
- The number of danmaku grows sharply around 18 July
- Word Cloud



Data Analysis

Top Danmaku in anime

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

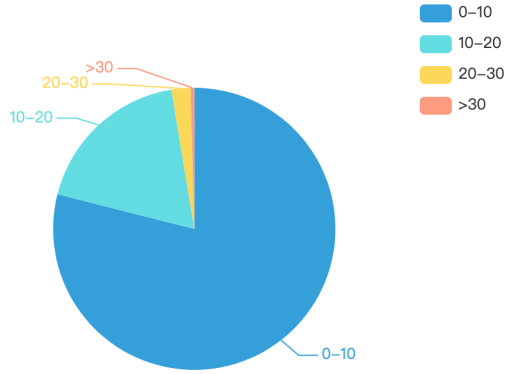


Data Analysis

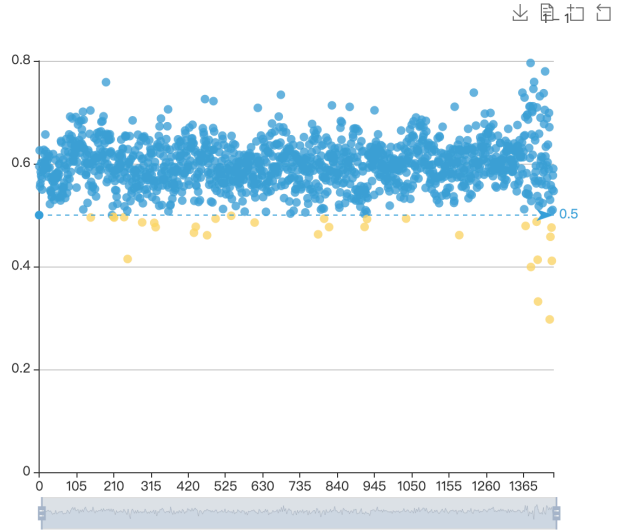
Case Study: CLANNAD ~After Story~

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

Danmaku Length Distribution



Danmaku Emotion

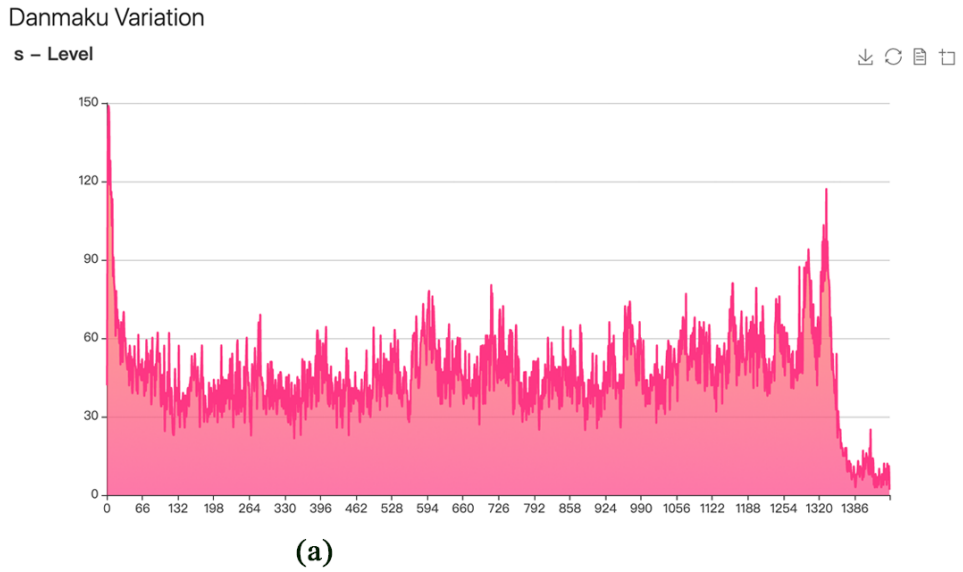


Danmaku Emotion Pie

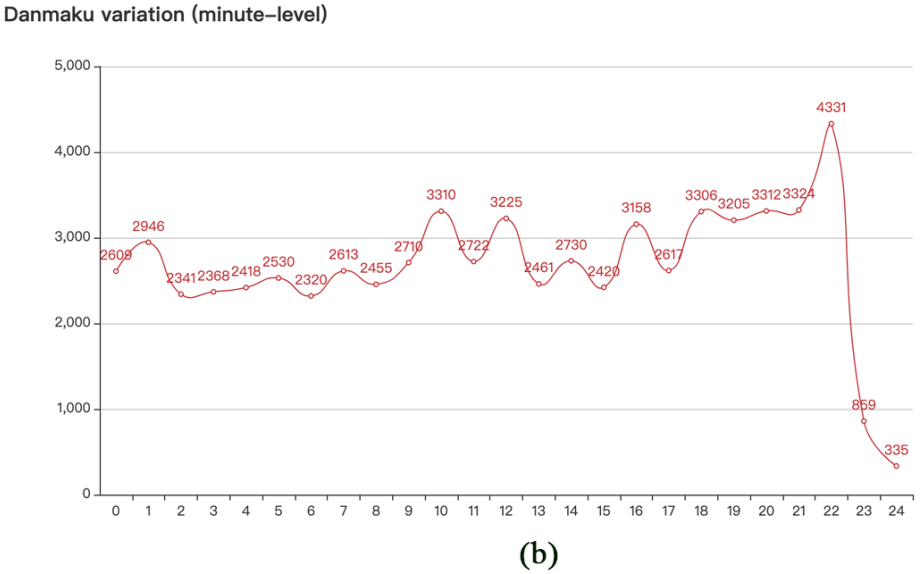
Data Analysis

Case Study: CLANNAD ~After Story~

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



(a)



(b)

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