"Bullet Screen" Analysis on Bilibili

bilibili

CAPP 30122 Project by
Yilun Dai
Ling Dai
Fangfang Wan

Introduction

Goal: To identify internet language patterns of people of different interests through "bullet screen" sent in videos.

bilibili

"Bullet screen": Comments that fly through the screen viewers send when watching a video. (dynamical loading)

苏州话好评　　　　　表白长恨歌　这个好　　　　　　这个

为我蛙蛙打call　　　闽南语吧　　　表白歌表白背景

妹子的外公我会永远记住他！！

没学过听着不知不觉背下来了

蛙蛙耶

敢不敢用普通话唱？

杨家有女初长成

Data Collection

Web Crawler:

- Bilibili: main site → categories → videos (sorted by popularity) → "bullet screens"

Information:

Name of the video, unique id, associated tags, bullet screens, etc.

Alternative:

Twitch: collect Twitch chats using Twitch IRC

Analysis

1. Aggregate texts
2. Data cleaning
3. Word segmentation / tokenization
4. Filter common words
5. Data visualization (words / n-grams)

Potential Additional Analysis (depending on additional features)

Final Product

- Interactive word cloud generator: Generate word clouds based on the user inputs (category/tags). Words with higher frequency will yield larger clouds.

Possible Additional features

- An algorithm that recommends Bilibili videos based on users' sent bullet screens/comments. (Word frequency distribution → PCA → calculate similarity and recommend nearest videos)
- An algorithm that searches relevant Bilibili videos based on the n-gram frequencies in "bullet screens" .

bilibili