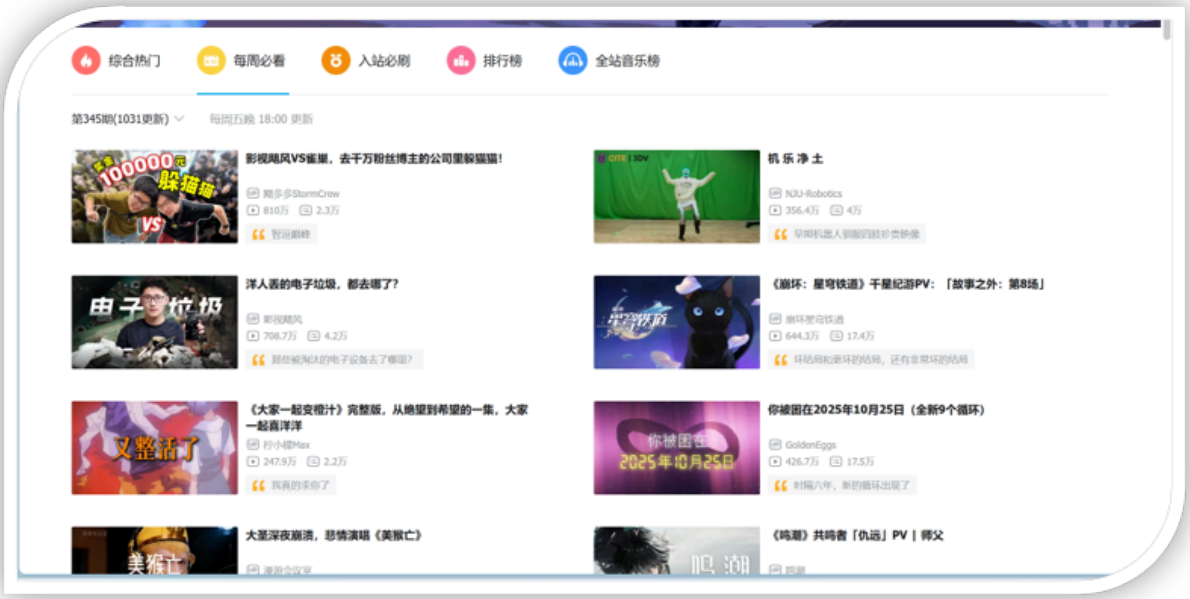
# 二、B站每周必看视频弹幕数据获取

## 1. 项目目标

获取B站「每周必看」系列视频的弹幕数据，为后续数据分析提供基础数据源。

## 2. 选择「每周必看」的原因

- 具备完整的历史数据，每一期对应一周时间，可直接映射到时间轴，便于时序分析；
- 视频经过平台筛选，均为对应时间段的热门内容，具有较强的用户代表性和话题性。



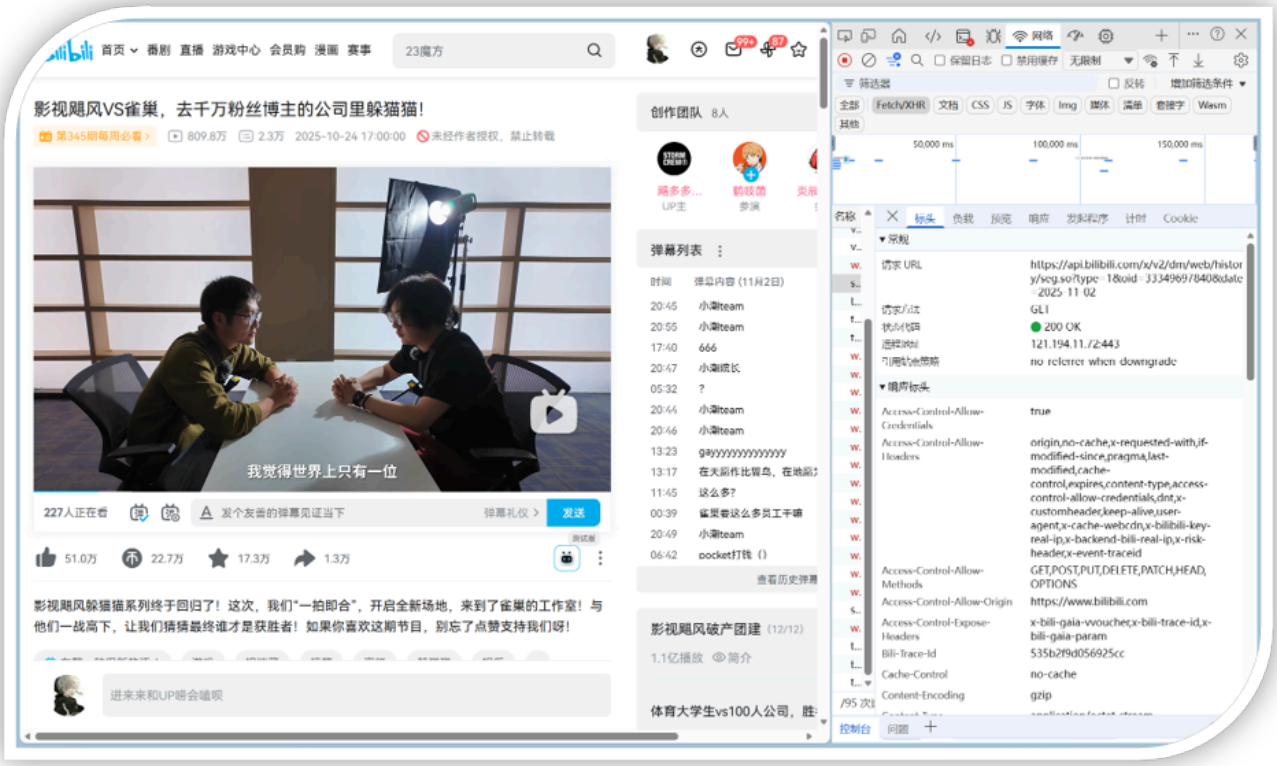

# 3. 数据获取原理

## 3.1 弹幕数据存储规律

B站网页端支持「查看历史弹幕」功能，通过F12开发人员工具可捕获弹幕数据的请求URL，其结构如下：

```
https://api.bilibili.com/x/v2/dm/web/history/seg.so?type=1&oid=XXX&date=YYYY-MM-DD
```

- 核心可变参数：
  - `oid`：视频唯一标识符，与视频的`cid`一一对应；
  - `date`：弹幕发送日期，格式为「年-月-日」。
- 只要获取目标视频的`cid`和目标日期范围，即可通过循环请求实现弹幕批量爬取。

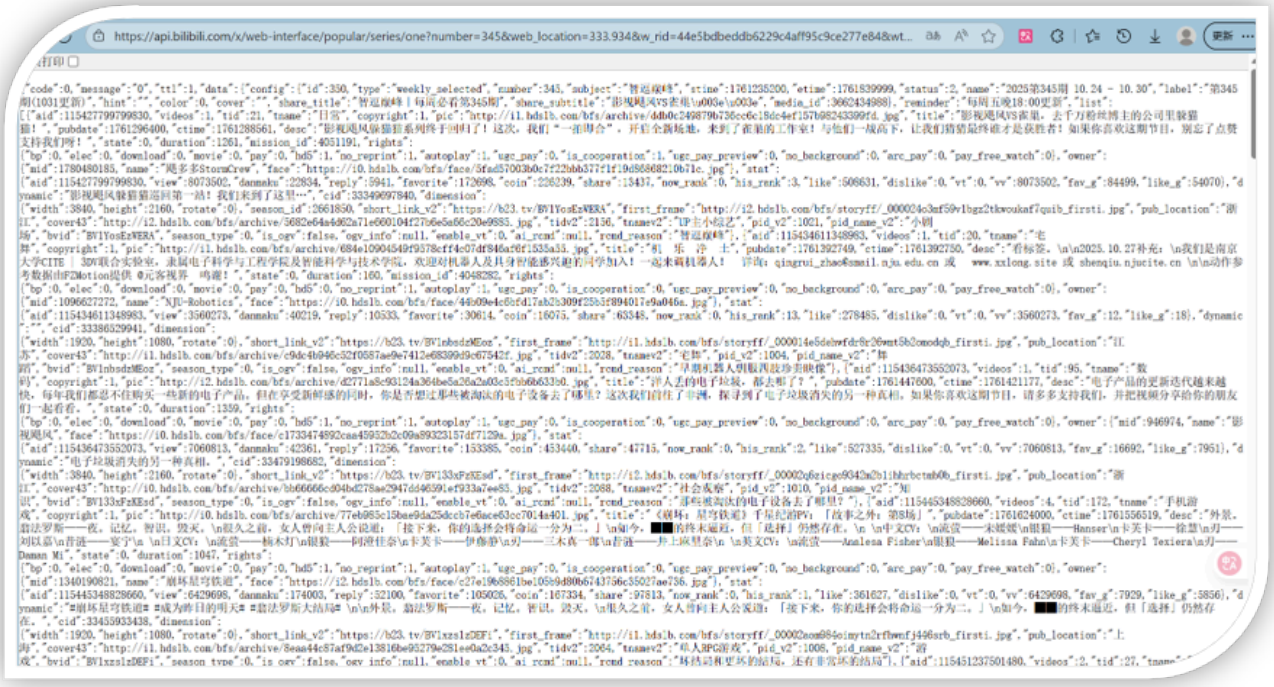



## 3.2 「每周必看」视频信息获取规律

通过F12工具捕获「每周必看」单期视频列表的请求URL，返回数据为JSON格式，可直接提取视频的`cid`（即弹幕URL中的`oid`）：

```
https://api.bilibili.com/x/web-interface/popular/series/one?
number=XXX&web_location=333.934&w_rid=XXX&wts=XXX
```

- 核心可变参数：`number`：「每周必看」的期数（如345即第345期）。

# 4. 具体实现步骤

## 4.1 环境准备

依赖库：`json`、`pandas`、`urllib`、`random`、`time`、`requests`、`re`

## 4.2 第一步：爬取「每周必看」视频基础信息（含cid）

### 4.2.1 核心函数定义

```
import json
import pandas as pd
from urllib import request, parse
import random
import time

def get_popular(url, headers):
```

```
    req = request.Request(url, headers=headers)
    response = request.urlopen(req)
    html = response.read()
    string = html.decode('utf8')
    time.sleep(random.random()*10)  # 设置随机间隔，避免过度查询被限制
    d = json.loads(string)
    df1 = pd.DataFrame()
    l_num = []
    l_title = []
    l_cid = []
    for i in range(len(d['data']['list'])):
        l_num.append(d['data']['config']['number'])  # 期数
        l_title.append(d['data']['list'][i]['title'])  # 视频标题
        l_cid.append(d['data']['list'][i]['cid'])  # 视频cid（对应弹幕oid）
    df1['number'] = l_num
    df1['title'] = l_title
    df1['cid'] = l_cid
    return df1
```

### 4.2.2 请求头配置（伪装真实用户）

```
headers = {
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/142.0.0.0 Safari/537.36 Edg/142.0.0.0',
    'cookie': "buvid3-AADA300E-DD22-E5F1-D078-981FE489901882947infoc;..."
}
```

### 4.2.3 批量爬取并保存数据

```
df_popular = pd.DataFrame()
# 爬取第1期到第348期（排除第105期，可能存在数据异常）
for i in range(1, 349):
    if i != 105:
        url = f'https://api.bilibili.com/x/web-interface/popular/series/one?number={i}&web_location=333.934&w_rid=44e5bdbeddb6229c4aff95c9ce277e84&wts=17'
        df1 = get_popular(url, headers)
        df_popular = pd.concat([df_popular, df1], ignore_index=True)
        print('第', i, '期已获取')

# 提取期数和cid，保存到Excel（避免重复爬取）
df_popular_cid = df_popular[['number', 'cid']]
df_popular_cid.to_excel('D:/O桌面文件/课程任务/DSAI/每周必看数据.xlsx', index=False)
```

## 4.3 第二步：基于cid爬取弹幕数据

### 4.3.1 单个视频单日弹幕爬取函数

```
import requests
import re
import random
import time
```

```
def danmu_get1(date, url0, headers):
    content_list = []
    url = f'{url0}{date}'
    response = requests.get(url=url, headers=headers)
    response.encoding = 'utf-8'
    # 提取中文弹幕内容（可根据需求调整正则表达式）
    temp_list = re.findall('[\u4e00-\u9fa5]+', response.text)
    content_list.extend(temp_list)
    time.sleep(random.random()*20)  # 延长间隔，降低请求频率
    return content_list
```

### 4.3.2 批量爬取配置与执行

- 日期选择：每期视频对应周的倒数第二天开始；
- 数据限制：B站历史弹幕最早保存到**2021年7月7日**，120期之前的视频从该日期开始爬取；
- 循环终止条件：单个视频弹幕数≥1000条 或 追溯天数≥100天；
- 保存格式：按「期数」分类，每一期的弹幕保存为独立txt文档，存放于同一文件夹。

```
url1 = f'https://api.bilibili.com/x/v2/dm/web/history/seg.so?type=1&oid='
url2 = f'&date='
headers = {
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/142.0.0.0 Safari/537.36 Edg/142.0.0.0',
    'cookie': "buvid3=AADA300E-DD22-E5F1-D078-981FE489901B82947infoc; ..."
}

for i in list(set(df_p['number']))[345:]: #中括号内根据期数调整
    df1=df_p[df_p['number']==i]
    danmu_l0=[]
    for j in range(len(df1)):
        oid=df1.iloc[j,]['cid']
        url0=f'{url1}{oid}{url2}'
        if i < 120:
            date0='2021-07-01'
        else:
            target_date = datetime.datetime.strptime('2021-07-07', '%Y-%m-%d') +
datetime.timedelta(days=(i - 120) * 7)
            date0=target_date.strftime('%Y-%m-%d')

        danmu_l=[]
        cycle=0
        while len(danmu_l)<1000 and cycle<100:
            danmu_l.extend(danmu_get1(date0,url0,headers))
            date0 = (datetime.datetime.strptime(date0, '%Y-%m-%d') +
datetime.timedelta(days=1)).strftime('%Y-%m-%d')
            cycle+=1
        danmu_l0.extend(danmu_l)
        if len(danmu_l)!=200:
            print('第',i,'期','第',j+1,'个视频弹幕已获取','共有',len(danmu_l),'条')
        elif len(danmu_l) == 200:
```

```
            print('第',i,'期','第',j+1,'个视频出现问题')
            raise Exception("已被B站反爬，停止运行")
        elif len(danmu_l) == 100:
            print('第',i,'期','第',j+1,'个视频出现问题')
            raise Exception("账号未登录，停止运行")

    content = '\n'.join(danmu_l0)
    txtname=f"D:/0桌面文件/课程任务/DSAI/每期弹幕/第{i}期.txt"
    with open(txtname, mode='w', encoding='utf-8') as f:
        f.write(content)
    print('第',i,"期保存完成",'共有',len(danmu_l0),'条')
```

# 5.结果展示