

# Python程序设计报告

设计题目： Python爬取B站数据并可视化展示

学生姓名：

专 业： 计算机科学与技术

班 级：

学 号：

指导教师： 马学森

完成日期： 2021.01.03

# 摘要

B站，既bilibili视频弹幕网（https://www.bilibili.com）是年轻人喜爱的视频网站。每天都有大量的观众进行视频观看活动，并留下弹幕和评论。这些数据被B站转换为固定类型的数据进行储存。本次大作业我的工作就是读取这些数据，并存在csv文件中，并使用numpy、matplotlib等python常用库对数据进行可视化的分析和展示。

**目录**

[摘要 1](#_Toc9696)

[一、 程序设计背景说明 2](#_Toc19985)

[1. 应用的使用背景 2](#_Toc22371)

[2. 程序设计的背景 2](#_Toc6509)

[二、 设计的核心算法 2](#_Toc20030)

[1. 爬取数据的算法 2](#_Toc16432)

[2. 转换数据的代码 3](#_Toc30108)

[3. 图形化展示的代码 3](#_Toc15935)

[三、 代码展示 4](#_Toc18833)

[1. get\_data.py 4](#_Toc22555)

[2. to\_graph.py 6](#_Toc14955)

[四、 数据的可视化展示 8](#_Toc19299)

[1. 热度和播放量评论 8](#_Toc2989)

[2. 弹幕数量的折线图 8](#_Toc24682)

[五、 思考与总结 9](#_Toc5077)

# 程序设计背景说明

## 应用的使用背景

获取排行榜数据可以很好的使用可视化工具等对其进行处理，这样，不仅有利于更好、更加直观的理解数据的具体内容，而且可以对数据进行具体的分析，并进行展示，还具有很高的效率和实用性。

## 程序设计的背景

参考到了群里的示例文件，觉得可视化方向是python的一个很好的课题，又不知道应该如何获取数据，获取何处的数据，最后决定从平时使用的B站中获得数据，并进行数据的处理和使用，这不仅符合我平时的生活习惯，也可以很好的为其他使用b站的小伙伴提供一个小“玩具”，姑且算是一种程序设计再实际生活中的应用了。

# 设计的核心算法

## 爬取数据的算法

f **=** **open(**'rank\_list\_of\_bilibili.csv'**,** mode**=**'a'**,** encoding**=**'utf-8-sig'**,** newline**=**''**)**

csv\_writer **=** csv**.**DictWriter**(**f**,** fieldnames**=[**'标题'**,** '播放量'**,** '弹幕量'**,** '作者'**,** '综合得分'**,** '视频地址'**])**

csv\_writer**.**writeheader**()**

# get data from url of bilibili

url **=** 'https://www.bilibili.com/v/popular/rank/all?spm\_id\_from=333.851.b\_7072696d61727950616765546162.3'

headers **=** **{**

'user-agent'**:** 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.138 Safari/537.36'

**}**

# get response

response **=** requests**.**get**(**url**=** url**,** headers **=** headers**)**

selector **=** parsel**.**Selector**(**response**.**text**)**

lis **=** selector**.**css**(**'.rank-list li'**)**

使用了csv的读取和写入，并且使用了简单的爬取方式，发送头，并获得反馈

## 转换数据的代码

**def** change\_num**(**number**):**

flag **=** '万'

**if** flag **in** number**:**

# change wan to 10000

number **=** re**.**sub**(**'万'**,**''**,**number**)**

**return** **float(**number**)** **\*** 10000

**else:**

**return** **float(**number**)**

这一部分的主要作用就是将爬取的数据中，含有汉字的部分进行处理，并转换为float新数据

## 图形化展示的代码

# you can chenge the width of the bar(single)

bar\_width **=** 0.4

'''

print the graph :using matplotib

'''

plt**.**title**(**'B站视频top20'**)**

plt**.**rcParams**[**'font.sans-serif'**]** **=** **[**'SimHei'**]** # 显示汉字

plt**.**xticks**(**rotation**=-**90**)**

plt**.**xlabel**(**'播放量（百万）'**)**

plt**.**ylabel**(**'视频名称'**)**

x **=** np**.**arange**(len(**lists**[**"title"**]))**

# plt.plot(x,lists["bf\_info"],marker = 'o',markersize = 3)

# for a, b in zip(x , lists["bf\_info"]):

# plt.text(a,b,b,ha = 'center', va='bottom', fontsize=10)

plt**.**barh**(**x **,**lists**[**"bf\_info"**]** **,**bar\_width **,**tick\_label **=** lists**[**"title"**]** **,**label **=** '热度'**)**

plt**.**barh**(**x **+** bar\_width **,**lists**[**"score"**]** **,**bar\_width **,**tick\_label **=** lists**[**"title"**]** **,**label **=** '分数'**)**

plt**.**legend**()**

plt**.**show**()**

这一部分是使用matplotlib进行绘图，不解释

# 代码展示

## get\_data.py

#-\*- coding utf-8 -\*-

**import** requests

**import** parsel

**import** csv

f **=** **open(**'rank\_list\_of\_bilibili.csv'**,** mode**=**'a'**,** encoding**=**'utf-8-sig'**,** newline**=**''**)**

csv\_writer **=** csv**.**DictWriter**(**f**,** fieldnames**=[**'标题'**,** '播放量'**,** '弹幕量'**,** '作者'**,** '综合得分'**,** '视频地址'**])**

csv\_writer**.**writeheader**()**

# get data from url of bilibili

url **=** 'https://www.bilibili.com/v/popular/rank/all?spm\_id\_from=333.851.b\_7072696d61727950616765546162.3'

headers **=** **{**

'user-agent'**:** 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.138 Safari/537.36'

**}**

# get response

response **=** requests**.**get**(**url**=** url**,** headers **=** headers**)**

selector **=** parsel**.**Selector**(**response**.**text**)**

lis **=** selector**.**css**(**'.rank-list li'**)**

flag **=** 0

# make dit of data

dit **=** **{}**

**for** li **in** lis**:**

# 标题

title **=** li**.**css**(**'.info a::text'**).**get**()**

# 播放量

bf\_info **=** li**.**css**(**'div.content > div.info > div.detail > span:nth-child(1)::text'**).**get**().**strip**()**

# 弹幕量

dm\_info **=** li**.**css**(**'div.content > div.info > div.detail > span:nth-child(2)::text'**).**get**().**strip**()**

# 作者

bq\_info **=** li**.**css**(**'div.content > div.info > div.detail > a > span::text'**).**get**().**strip**()**

# 综合得分

score **=** li**.**css**(**'.pts div::text'**).**get**()**

# 视频地址

page\_url **=** li**.**css**(**'.img a::attr(href)'**).**get**()**

# put data input dit

dit **=** **{**

'标题'**:** title**,**

'播放量'**:** bf\_info**,**

'弹幕量'**:** dm\_info**,**

'作者'**:** bq\_info**,**

'综合得分'**:** score**,**

'视频地址'**:** page\_url**,**

**}**

# flag is use to control the number of list

**if** flag **==** 25**:**

**break**

csv\_writer**.**writerow**(**dit**)**

flag **+=** 1

# to show the ditail of data

**print(**dit**)**

## to\_graph.py

#-\*- coding utf-8 -\*-

**import** csv

**import** numpy **as** np

**import** matplotlib**.**pyplot **as** plt

**import** re

'''

to solve the chinese in num\_data get from bilibili

input : string

output : floatS

'''

**def** change\_num**(**number**):**

flag **=** '万'

**if** flag **in** number**:**

# change wan to 10000

number **=** re**.**sub**(**'万'**,**''**,**number**)**

**return** **float(**number**)** **\*** 10000

**else:**

**return** **float(**number**)**

# get data from \*.csv

# encoding UTF-8

**with** **open(**r"rank\_list\_of\_bilibili.csv"**,**encoding**=**'UTF-8'**)** **as** bili**:**

read\_csv **=** csv**.**reader**(**bili**)**

# prepare list

title **,**bf\_info **,**dm\_info**,** bq\_info **,**score **,**page\_url **=** **[],[],[],[],[],[]**

index **=** 0**;**

**for** i **in** read\_csv**:**

**if** index **!=** 0**:**

title**.**append**(**i**[**0**])**

bf\_info**.**append**(**change\_num**(**i**[**1**]))**

dm\_info**.**append**(**change\_num**(**i**[**2**]))**

bq\_info**.**append**(**i**[**3**])**

score**.**append**(**change\_num**(**i**[**4**]))**

page\_url**.**append**(**i**[**5**])**

index **+=** 1

# set title

title\_list **=** **[**'title'**,**'bf\_info'**,**'dm\_info'**,**'bq\_info'**,**'score'**,**'page\_url'**]**

lists **=** **{}**

lists**[**"title"**],**lists**[**"bf\_info"**],**lists**[**"dm\_info"**],**lists**[**"bq\_info"**]** **,**lists**[**"score"**]** **,**lists**[**"page\_url"**]** **=** title **,**bf\_info **,**dm\_info **,**bq\_info **,**score **,**page\_url

**print(**lists**)**

# you can chenge the width of the bar(single)

bar\_width **=** 0.4

'''

print the graph :using matplotib

'''

plt**.**title**(**'B站视频top20'**)**

plt**.**rcParams**[**'font.sans-serif'**]** **=** **[**'SimHei'**]** # 显示汉字

plt**.**xticks**(**rotation**=-**90**)**

plt**.**xlabel**(**'播放量（百万）'**)**

plt**.**ylabel**(**'视频名称'**)**

x **=** np**.**arange**(len(**lists**[**"title"**]))**

# plt.plot(x,lists["bf\_info"],marker = 'o',markersize = 3)

# for a, b in zip(x , lists["bf\_info"]):

# plt.text(a,b,b,ha = 'center', va='bottom', fontsize=10)

plt**.**barh**(**x **,**lists**[**"bf\_info"**]** **,**bar\_width **,**tick\_label **=** lists**[**"title"**]** **,**label **=** '热度'**)**

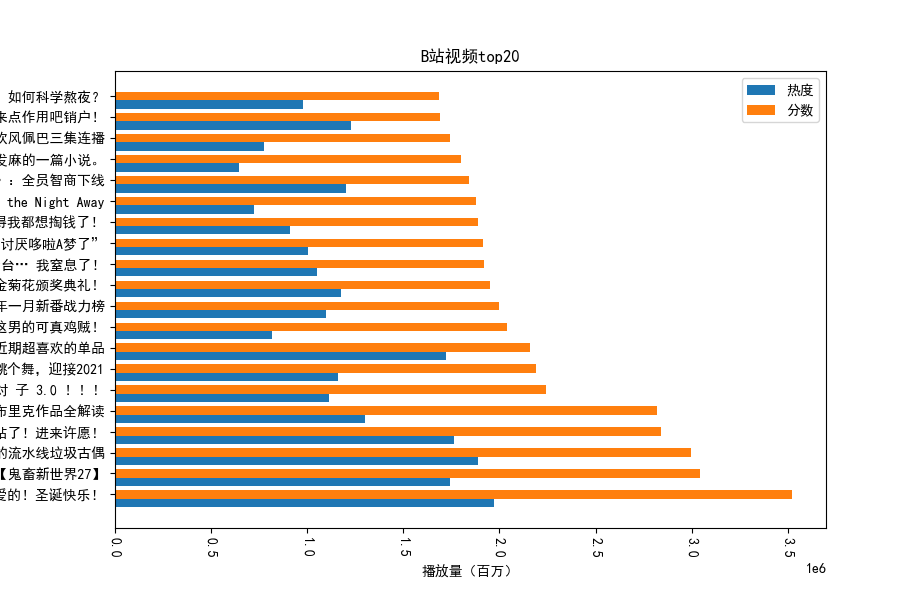
plt**.**barh**(**x **+** bar\_width **,**lists**[**"score"**]** **,**bar\_width **,**tick\_label **=** lists**[**"title"**]** **,**label **=** '分数'**)**

plt**.**legend**()**

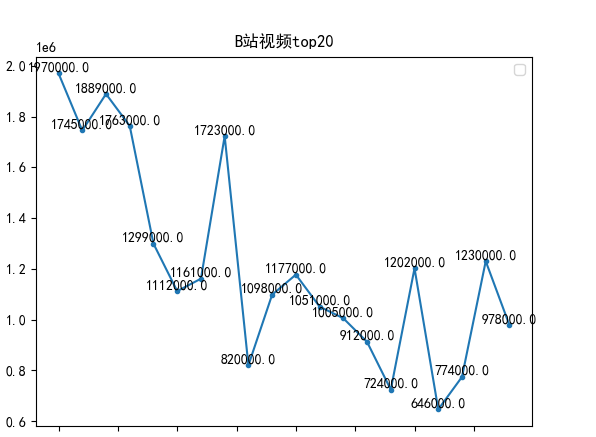
plt**.**show**()**

# 数据的可视化展示

## 热度和播放量评论



## 弹幕数量的折线图



# 思考与总结

本次大作业，我觉得是一次很好的锻炼。不仅很好的练习了如何使用python的各种库文件，并且更加熟悉了python的基本操作。更加重要的是寻求和获得解决问题的方法。学习语言本身就是一种进步的过程，遇到不会的问题怎么去解决，我再解决问题的过程中学到的经验比什么都重要。

另一个方面，历经坎坷才更加珍贵。虽然我的大作业只是算是一个很简陋的小工具，但是也经历了多次调试和更正。