|  |  |
| --- | --- |
| 学号 | 20514118220 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 吴亦凡微博热点的分析 |
| 专 业 | \_\_\_\_\_\_\_软件工程\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 班 级 | \_\_\_\_\_\_\_软件1182\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 姓 名 | \_\_\_\_\_\_\_伍丽丽\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 成 绩 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 指 导 老 师 | \_\_\_\_\_\_\_\_\_\_\_徐勇\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

2021 年 1 月 11 日至 2021 年 1 月 22 日

武汉华夏理工学院信息工程学院

**课 程 设 计 任 务 书**

课程名称：python程序设计课程设计 指导教师：徐勇

班级名称： 软件1181-1182 开课院、系：计算机与网络工程系

**一、课程设计目的与任务**

“python程序设计课程设计”是一个综合性的学习实践型实验教学环节，将在“python程序设计”课程的授课基础上，对python的基础语法、python 对文件的操作、python 对数据的操作、数据处理基础、数据可视化等若干个知识点进行综合运用。

python是一种具有天然开源基因的编程语言，了解开源社区和广泛的使用开源工具，也是Python学习实践的重要环节。因此在本次课程设计中，要求掌握jupyter notebook、Git等常用工具，以及github等重要开源社区的使用。

**二、课程设计的内容与基本要求**

数据分析的基本技术和方法在“python程序设计”课程中已经进行完整的讲授，由于数据分析本身是基于业务场景的，因此本次课程设计环节更加偏重于实际的业务场景的实践。通过对近期互联网热点的调查，准备了15个具体的业务场景，用于本次课程设计的具体任务场景。具体包括如下：

1. 北上广深租房状况分析；
2. 蔡某坤粉丝数及转发数据真假状况分析；
3. 地震的数据分析；
4. 英文名字的数据分析；
5. 外籍英文老师收入虚高情况数据分析；
6. 我国城市空气污染和烟花燃放的关系分析；
7. 针对996工作，程序员群体的看法的分析；
8. 吴某凡微博热点的分析；
9. 节假日长假景点人满为患的数据分析；
10. 针对荔枝的品种、销售地等维度，进行价格数据分析；
11. 分析芒果TV《我是大侦探》的观众评论数据；
12. 针对当前儿科医生的缺乏，对相关数据进行分析；
13. 著名网游《绝地求生》的数据分析；
14. 实习岗位状况的数据分析；
15. 电影《流浪地球》的观众评价的数据分析。

每三个学生组建一个课程设计小组，最后的任务输出包括代码每小组一份、课程设计报告每人一份、答辩ppt每小组一份，并进行课程设计成果答辩。小组成员均参与前述工作，但是每个人的侧重点不同。

每个课程设计小组可以从上述15个场景中选取一个作为课程设计的选题，选题中提供了待分析的数据，和现有的分析方法。各小组，通过学习和实践现有的分析方法，理解实战分析的思维过程并锻炼实际动手能力，再此基础上可以扩展更多维度的分析和数据展现形式。

每个班每个选题最多只能被两个小组选中，先选先得。同时，如果各小组发现更有意思的场景，并能够获取到相关待分析的数据，也可以申请作为选题方向。

本次课程设计的目标是培养学生的团队协作能力、对python知识点的综合运用、对实际场景的理解和适应能力、针对答辩的表达能力等。注重过程，期待成果，但不强求结果的尽善尽美。

**三、学时分配进度安排**

|  |  |  |
| --- | --- | --- |
| 序号 | 设计内容 | 所用时间 |
| 1 | 下发任务书，学生查阅资料 | 1天 |
| 2 | 组建团队，并进行选题和团队匹配 | 1天 |
| 3 | 各小组对自己的选题开始进行研究 | 2天 |
| 4 | 各小组开始准备数据并形成初步处理意见 | 1天 |
| 5 | 代码和实现分析的动手实践 | 2天 |
| 6 | 形成初步的报告书和ppt | 1天 |
| 7 | 答辩并完成报告书 | 2天 |
| 合 计 | | 2周 |

**四、课程设计考核及评分标准**

**1.设计报告要求**

课程设计报告要求逻辑清晰、层次分明、书写整洁。课程设计报告为每人一份，同一个小组的各成员的整体报告内容结构一致，但个人侧重点不同，个人着重撰写自己的工作内容，其他人的内容只要体现文档结构的完整性即可。

课程设计考核将综合考虑学生考勤和参与度、团队协作能力，过程管理能力、成果达成情况等。

**2.过程要求**

整个过程要求通过github来进行过程化发布，即阶段性地在github上提交过程结果。

**3.评分标准**

|  |  |
| --- | --- |
| **评分依据** | **评分成绩** |
| 1．团队协作能力 | 25分 |
| 2．python综合运用能力 | 25分 |
| 3．态度认真、刻苦钻研、遵守纪律 | 10分 |
| 4．过程完成、对工具的使用、对github的运用 | 20分 |
| 5．课程设计答辩逻辑清晰，内容正确 | 10分 |
| 6. 课程设计期间的课堂考勤、创新能力 | 10分 |
| 总分 | 100分 |

注：按上述六项分别记分后求和，根据小组成员贡献率综合评定，记载个人最后成绩。

成绩等级：优（90分—100分）、良（80分—89分）、中（70分—79分）、及格（60分—69分）、60分以下为不及格。

**五、指导地点与时间**

本课程设计将安排在第20-21周，采用腾讯会议和QQ群的方式，以在线形式进行。具体安排如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 周次 | 星期一 | 星期二 | 星期三 | 星期四 | 星期五 |
| 第20周 | 第5-8节 | 第5-8节 |  | 第5-8节 |  |
| 第21周 |  | 第5-8节 |  | 第5-8节 |  |

执笔:徐 勇 日期：2021-1-8

审阅:钱小红 日期：2021-1-8

**目 录**

[1. 设计题目 1](#_Toc7418)

[1.1. 项目背景 1](#_Toc5803)

[1.2. 数据来源 1](#_Toc13114)

[1.3. 分析目标 1](#_Toc9938)

[2. 开发环境 2](#_Toc32173)

[3. 数据读取 2](#_Toc31718)

[4. 数据清洗 3](#_Toc29134)

[5. 数据分析 3](#_Toc5003)

[5.1. 吴亦凡该微博的转发是否存在假流量 3](#_Toc25701)

[5.2. 大家对于《大碗宽面》怎么看 6](#_Toc20055)

[5.3. 有多少人拿吴亦凡跟蔡徐坤做对比 8](#_Toc16114)

[5.4. 有多少人开始路转粉了 9](#_Toc22018)

[5.5. 评论的词云图 10](#_Toc14747)

[6. 结论 12](#_Toc10788)

[7. 课设总结 12](#_Toc1406)

[7.1. 问题与解决方法 12](#_Toc12220)

[7.2. 设计体会 13](#_Toc29329)

[参 考 文 献 14](#_Toc16918)

[附 录 15](#_Toc13681)

# 设计题目

吴亦凡微博热点的分析

## 项目背景

微博已经成为追星一族必备软件，各家粉丝为了让自己的爱豆有更大的流量，更高的人气，在各种榜单中有一席之地，纷纷转发、点赞和评论爱豆的微博。可见，微博转发量等数据越来越成为了评判明星是否火的方式，于是出现了经纪公司或者忠实的粉丝自费购买虚假转发的现象。于是我们怀着好奇心，选取吴亦凡发表《大碗宽面》的微博，探求是否存在假流量？假流量所占比重有多少？所以本次课程设计通过对微博热点文件（kriswu.csv)中的粉丝数量，流量数据等的统计，综合运用文件读取、pandas数据处理和matplotlib绘图功能等，全面掌握Python程序设计与数据处理方法，从而使读者具备大数据处理的基本能力。

## 数据来源

随机抓取吴亦凡发表《大碗宽面》微博的微博转发数据10万条，并分析大家对于这首歌的情感倾向如何（大碗宽面 VS 律师函警告，情感分析吴亦凡自黑式圈粉！），项目主要随机抓取吴亦凡发表《大碗宽面》微博的微博转发数据10万条，并分析该条微博的真假转发比例，以及大家对于这首歌的情感倾向如何。

用到的数据包有：

kriswu.csv （吴亦凡微博热点数据统计）

stopwords.txt（ 停用词表）

## 分析目标

1. 吴亦凡该微博的转发是否存在假流量？

2.大家对于《大碗宽面》怎么看？

3.有多少人拿吴亦凡跟蔡徐坤做对比？

4. 有多少人开始路转粉了？

5.评论的词云图

# 开发环境

操作系统：Windows10

选题使用的主流Notebook环境：jupyter Notebook.

运行环境：python3.8

需要安装的工具：

Pyecharts0.5.6

Pandas1.0.2

Numpy1.19.5

Matplotlib3.3.3

Seaborn0.11.1

Snownlp0.12.3

# 数据读取

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from pyecharts.charts.bar import Bar

%matplotlib inline

plt.style.use('ggplot')

import matplotlib as mpl

mpl.rcParams['font.sans-serif'] = ['SimHei'] #解决seaborn中文字体显示问题

plt.rc('figure', figsize=(10, 10)) #把plt默认的图片size调大一点

plt.rcParams["figure.dpi"] =mpl.rcParams['axes.unicode\_minus'] = False # 解决保存图像是负号'-'显示为方块的问题

data = pd.read\_csv('kriswu.csv', encoding='UTF-8')

#打印，以确保数据运行正常

Data

data.info()

data.sample(5)

# 数据清洗

由于数据入库的时候没有进行清洗，所以数据多出了很多没用的字段，需要先清洗掉。

首先打印数据列表：

print(list(data.columns))

筛选出KrisWu中需要用到的数据列：

in\_columns = ['attitudes\_count', 'comments\_count', 'reposts\_count', 'mid', 'raw\_text','source','user.description','user.follow\_count','user.followers\_count','user.gender', 'user.id', 'user.mbrank', 'user.mbtype', 'user.profile\_url', 'user.profile\_image\_url','user.screen\_name', 'user.statuses\_count', 'user.urank', 'user.verified', 'user.verified\_reason']

data = data[in\_columns]

data.info()

data.to\_csv('kriswu.csv', index=False)

筛选出的数据在分析下面的问题时将会用到。

# 数据分析

## 吴亦凡该微博的转发是否存在假流量

# 吴亦凡的粉丝性别比例

fans\_num = data['user.gender'].value\_counts()

fans\_num

from pyecharts import Bar

bar = Bar("吴亦凡粉丝性别比例初探", width = 600,height=500)

bar.add("(总数据102118条)", ['女', '男'], fans\_num.values, is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True)

Bar

将吴亦凡粉丝分为男女两部分并计算数量绘图，吴亦凡粉丝性别比例初探图如图5.1所示。

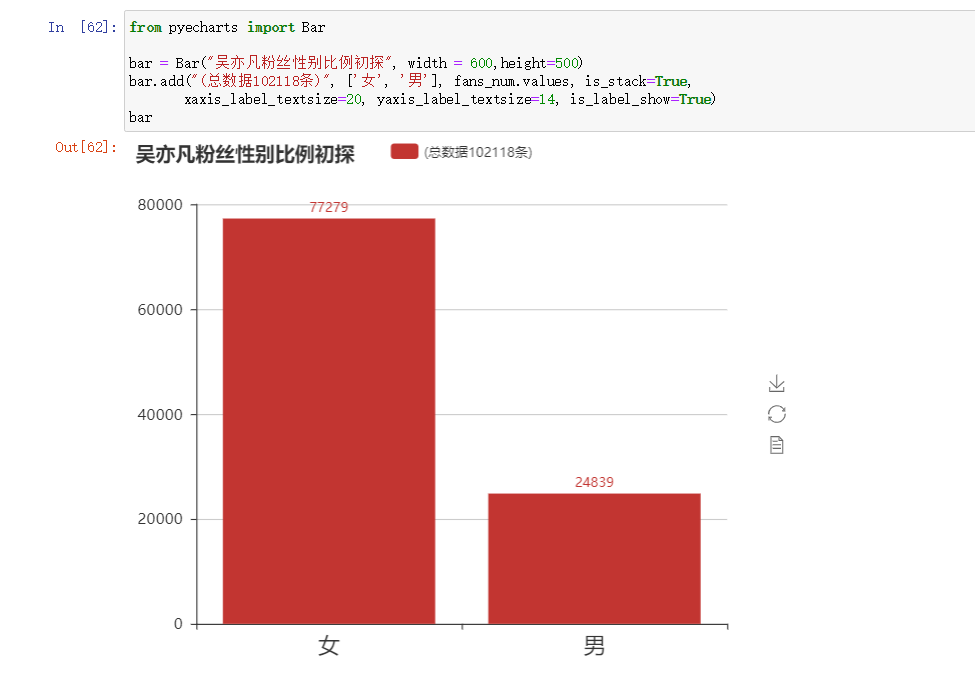


图5.1 吴亦凡粉丝性别比例初探图

np.round(fans\_num/fans\_num.sum()\*100, 2)

data[data['user.gender']=='f'].sample(5)

data[data['user.gender']=='m'].sample(5)

#粉丝数大于等于5但昵称包含“用户”的这样的数据抽取出来。基本上可以断定是假粉丝

data\_fake2\_index = data[(data['user.follow\_count']>5)&

(data['user.followers\_count']>5)&

(data['user.screen\_name'].str.contains('用户'))].index

data\_fake2\_index

data\_fake2\_index.shape

# 昵称里包含“用户”的，基本上可以断定是假粉丝

data\_fake2\_index = data[(data['user.follow\_count']>5)&

(data['user.followers\_count']>5)&

(data['user.screen\_name'].str.contains('用户'))].index

# 把假的流量粉丝转发组合起来

data\_fake = pd.concat([ data.iloc[data\_fake2\_index]])

data\_fake.shape

# 取出真粉的转发

data\_true = data.drop(data\_fake.index)

data\_true.shape

print('真粉丝转发数占总转发数的{}%'.format(np.round(data\_true.shape[0]/data.shape[0]\*100, 2)))

print('假粉丝转发数占总转发数的{}%'.format(np.round(data\_fake.shape[0]/data.shape[0]\*100, 2)))

bar = Bar("吴亦凡真假流量的转发量", width = 600,height=500)

bar.add("(总数据102118条)", ['总转发量', '假粉丝转发量', '真粉丝转发量'],

[data.shape[0], data\_fake.shape[0], data\_true.shape[0]], is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True)

Bar

将转发中的总转发量、假粉丝转发量和真粉丝转发量计算出来并绘图，吴亦凡真假流量的转发量如图5.2所示。

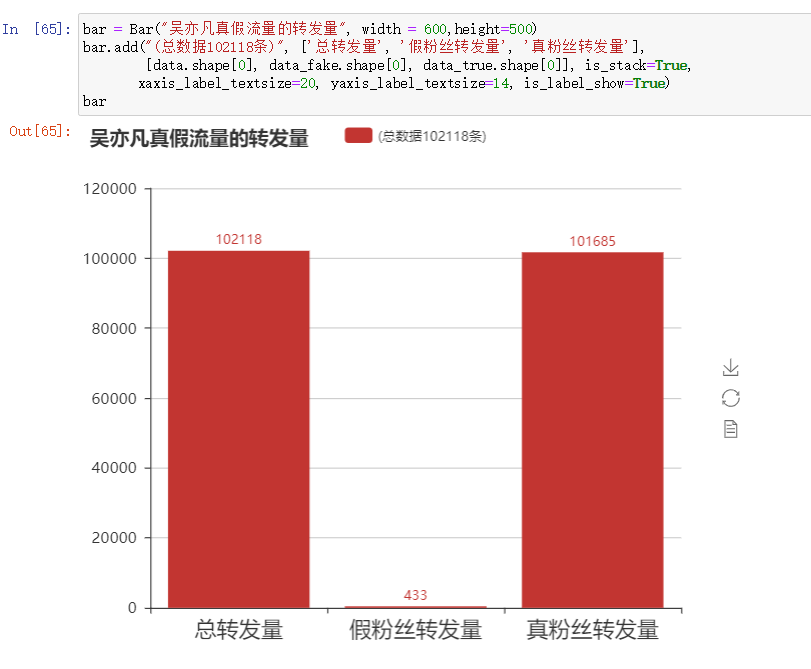


图5.2 吴亦凡真假流量的转发量

real\_fans\_num = data\_true.drop\_duplicates(subset='user.id').shape[0]

bar = Bar("吴亦凡真假流量的转发量与真实转发粉丝量(总数据102118条)", width = 600,height=500)

bar.add('', ['总转发量', '假粉丝转发量', '真粉丝转发量', '真实转发粉丝量'],

[data.shape[0], data\_fake.shape[0], data\_true.shape[0], real\_fans\_num], is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True, xaxis\_rotate=20)

Bar

将转发中的总转发量、假粉丝转发量、真粉丝转发量和真实转发粉丝量计算出来并绘图，吴亦凡真假流量的转发量与真实转发粉丝量如图5.3所示。

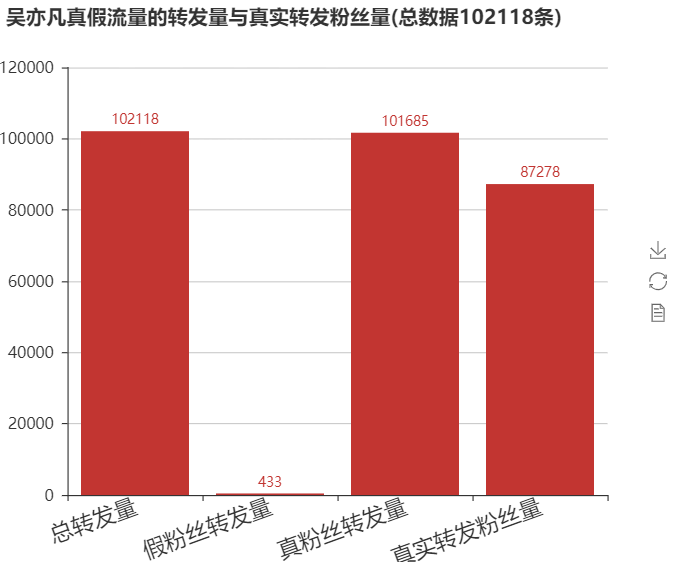


图5.3 吴亦凡真假流量的转发量与真实转发粉丝量

print('真实转发粉丝量占总转发数的{}%'.format(np.round(real\_fans\_num/data.shape[0]\*100, 2)))

输出真实粉丝转发量与总转发量之比，为85.47%。

## 大家对于《大碗宽面》怎么看

data.info()

这里用到了snownlp库，主要可以进行中文分词、词性标注、情感分析、文本分类、转换拼音、繁体转简体、提取文本关键词、提取摘要、分割句子、文本相似。

from snownlp import SnowNLP

def get\_sent\_snownlp(data):

s = SnowNLP(data)

return s.sentiments

data\_true['clean\_text'] = data\_true['raw\_text'].str.split("//", expand=True)[0]

ata\_true['sent\_nlp'] = data\_true.loc[data\_true['clean\_text']!='', 'clean\_text'].apply(get\_sent\_snownlp)

data\_true[['clean\_text', 'sent\_nlp']].sample(5)

提取'clean\_text和clean\_text两列的数据并输出部分内容，用来展示微博评论中对《大碗宽面》的评价，《大碗宽面》的评价如图5.4所示。

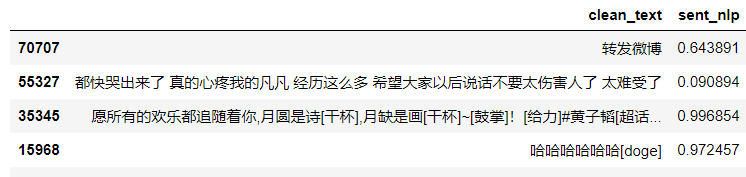


图5.4 《大碗宽面》的评价

data\_true.loc[-data\_true['clean\_text'].isin(['转发微博', 'repost', '轉發微博']), 'sent\_nlp'].mean()

from pyecharts.charts.gauge import Gauge

from pyecharts.charts.geo import Geo

g = Gauge()

g.add('', ['评论对《大碗宽面》\n的平均评分'], [68.6])

g

利用pyecharts，将大碗宽面的平均分（68.6分）以图表的方式打印出来，平均评分如图5.5所示。

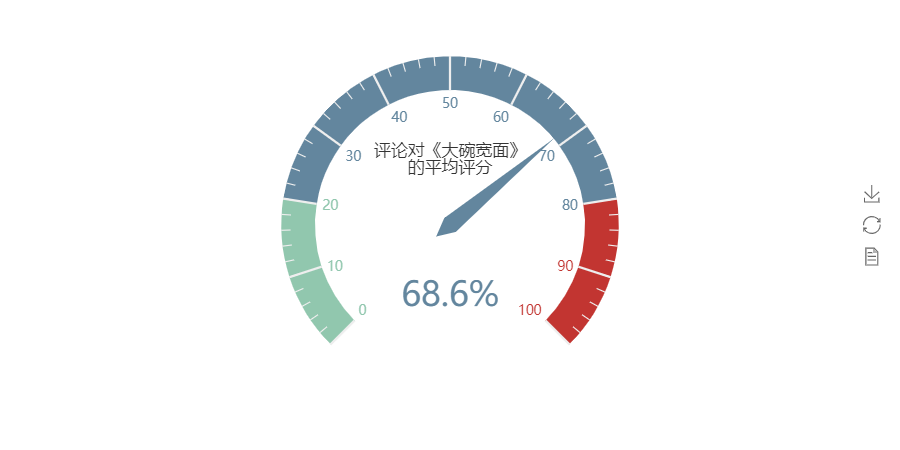


图5.5 平均评分

data\_true.loc[data\_true['sent\_nlp']>0.9, ['clean\_text', 'sent\_nlp']].sample(5)

选取打分超过九十分的评论提取出来。超过九十分的评价如图5.6所示。

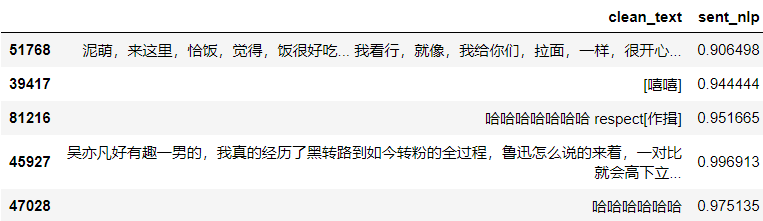


图5.6 超过九十分的评价

data\_true.loc[data\_true['attitudes\_count'].sort\_values(ascending=False)[:10].index, ['user.screen\_name', 'clean\_text', 'attitudes\_count']]

data\_true.loc[data\_true['attitudes\_count'].sort\_values(ascending=False)[:100].index, ['user.screen\_name', 'clean\_text', 'attitudes\_count']].sample(5)

将热评提取出来。热评图如图5.7所示。

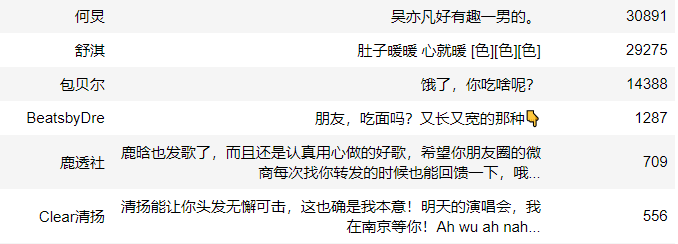


图5.7 热评

## 有多少人拿吴亦凡跟蔡徐坤做对比

data\_true.info()

data\_true['clean\_text'].sample(10)

data\_true['clean\_text'].sample(10)

提取

data\_true['raw\_text'].str.contains('kun|坤|律师|球|函|cxk|比|弟|CXK|胸|格局|气度|衬托').sum()

计算含有关键字“kun|坤|律师|球|函|cxk|比|弟|CXK|胸|格局|气度|衬托”的数量，得出含有这些关键字的评论有6388条。

data\_kun = data\_true.loc[data\_true['raw\_text'].str.contains('kun|坤|律师|球|函|cxk|比|弟|CXK|胸|格局|气度|衬托'),

['user.screen\_name', 'raw\_text', 'attitudes\_count']]

data\_kun.loc[list(data\_kun['attitudes\_count'].nlargest(10).index), ['raw\_text', 'attitudes\_count']]

提取部分与蔡徐坤做对比的评论并打印出来。对比评论如图5.8所示。

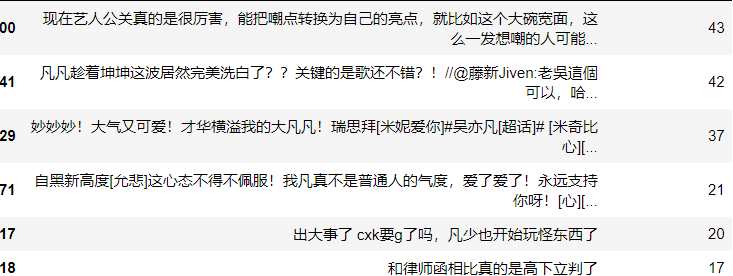


图5.8 对比评论

data\_true[data\_true['raw\_text']!=''].shape

6229/96018

## 有多少人开始路转粉了

data\_true['clean\_text'].sample(10)

data\_true['raw\_text'].str.contains('转粉|爱上|重新|路|圈粉|espect|瑞思拜').sum()

计算评论中含有关键词“转粉|爱上|重新|路|圈粉|espect|瑞思拜”的评论数量，得出含有这些关键词的评论有3783条。

data\_true.loc[data\_true['raw\_text'].str.contains('转粉|爱上|重新|路|圈粉|espect|瑞思拜'),

['user.screen\_name', 'raw\_text']].sample(5)

提取部分含有这些关键词的评论并打印出来。路转粉粉丝评论如图5.9所示。

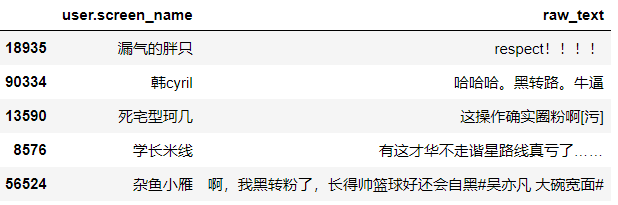


图5.9 路转粉粉丝评论

fans = data\_true.loc[data\_true['raw\_text'].str.contains('转粉|爱上|重新|路|圈粉|espect|瑞思拜'),

'user.gender'].value\_counts()

from pyecharts.charts.bar import Bar

bar = Bar("路转粉的男女性别比例", width = 600,height=500)

bar.add("", ['女', '男'], fans.values, is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True)

Bar

提取含有“转粉|爱上|重新|路|圈粉|espect|瑞思拜”的评论，通过性别分成两部分，并计算出对应的数量并绘图。路转粉的男女性别比例如图5.10所示。

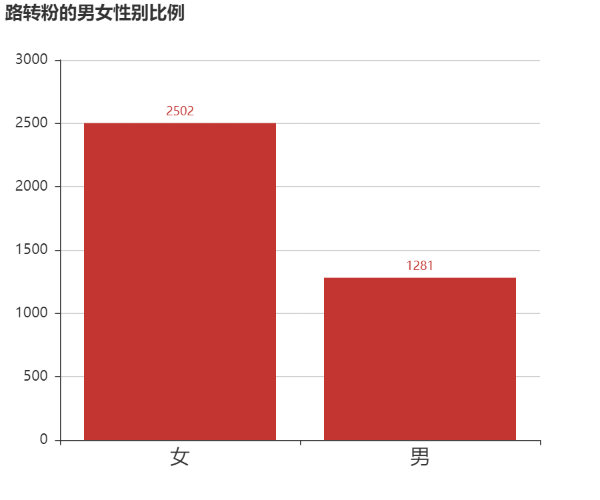


图5.10 路转粉的男女性别比例

fans/fans.sum()

路转粉的女粉丝数量：路转粉的男粉丝数量约为6:4。

## 评论的词云图

# 绘制评论词云图

import jieba

from collections import Counter

from pyecharts.charts.wordcloud import WordCloud

jieba.add\_word('吴亦凡')

swords = [x.strip() for x in open ('stopwords.txt', encoding='UTF-8')]

def plot\_word\_cloud(data, swords):

text = ''.join(data)

words = list(jieba.cut(text))

ex\_sw\_words = []

for word in words:

if len(word)>1 and (word not in swords):

ex\_sw\_words.append(word)

c = Counter()

c = Counter(ex\_sw\_words)

wc\_data = pd.DataFrame({'word':list(c.keys()), 'counts':list(c.values())}).sort\_values(by='counts', ascending=False).head(100)

wordcloud = WordCloud(width=1300, height=620)

wordcloud.add("", wc\_data['word'], wc\_data['counts'], word\_size\_range=[12, 60])

return wordcloud

plot\_word\_cloud(data=data\_true.loc[-data\_true['clean\_text'].str.contains("转发"), 'clean\_text'], swords=swords)

将评论内容绘制词云图，词云图如图5.11所示。



图5.11 评论词云图

# 结论

通过对吴亦凡微博热点数据进行分析，我们得出以下结论：

1.女粉丝所占比例比男粉丝多，存在假流量，但是在所有转发中，绝大部分是真实粉丝转发，只有少量的假粉丝转发量。

2.大部分的人还是喜欢听《大碗宽面》的，对《大碗宽面》的评论平均分是68.6%。

3.拿吴亦凡和蔡徐坤做对比的人数根据提取数据前十条显示有6388人次。

4.路转粉的人数比较多，其中男女比例大概是男：女=3：7。

5.吴亦凡挺受大家欢迎的！！！

# 课设总结

## 问题与解决方法

课设过程中遇到的问题有以及解决方法：

1. 数据清洗模块中，使用columns函数后运行出现错误，原因是输入KrisWu.csv文件的列名有误。
2. 绘图板块出现白板，没有图片，是pyecharts的版本不适配的原因，卸载原先装好的pyecharts，安装Pyecharts0.5.6。
3. 代码运行到Snownlp进行情绪分析时出现运行缓慢，再运行之前已经运行过的代码时也运行不出来，其实是因为数据量过大，只需等待一段时间便可继续。

## 设计体会

这次课程设计，我选的题目是吴亦凡微博热点的分析。老师给出的文件中有多个问题需要我们去解决，我在做该设计的过程中也遇到了很多问题，有的是知识存储不足，有的是考虑不够周全，之所以能够顺利实现基本功功能，离不开老师和其他小组成员的鼎力相助，充分体现出了团队的力量。  
 在此次课设的过程中我学到了许多新的知识比如Numpy、Snownlp、Seaborn等一些工具的运用，是一个让我把书本上的理论知识运用于实践中的好机会，还锻炼了我其他方面的潜力，提高了我的综合素质，提高了独立思考问题、自我动手操作的潜力，在工作的过程中，复习了以前学习过的知识，并掌握了一些应用知识的技巧等也使我更加有团队精神。

我们不可能做到面面俱到，但一定要做到步步扎实，作为一个学习程序编程的人，要保持清醒的头脑，以现实为依据，让自己的每一行代码都能实现自己的意义。在编写代码时，由于自己技术知识水平不广，常常遇到技术难题。在制作PPT的过程中，为了在答辩过程中展示出更好的效果，不断尝试添加一些图片和调色，也花费了较长的时间。

这次课程设计让我意识到了自己做为计算机软件工程专业的学生，要想在以后的职业中崭露头角，除了要有过硬的理论知识，健康的体魄外，还必须具备良好的心理素质，使自己在以后的途中无论经历什么样的困难，都立于不败之地。“纸上得来终觉浅，绝知此事要躬行!” 在这短短的时间里，让我深深的感觉到自己在实际应用中所学专业知识的匮乏。让我真 真领悟到“学无止境”这句话的涵义。而所学的，都是课本.上没有而对我们又非常实用的东西，这又给我们的实训增加了浓墨淡采的光辉。我懂得了实际生活中，专业知识是怎样应用与实践的。我学习到的不仅仅是课程上的知识得到实际应用，还有编程的习惯和进行数据分析时应注意方方面面，收获颇多。

**参 考 文 献**

[1] 张若愚 Python 科学计算[M].北京：清华大学出版社.2012.

[2] 刘腾红，孙细明.信息系统分析与设计[M]. 北京：科学出版社，2003.

[3] 麦金尼，利用Python进行数据分析[M].北京：[机械工业出版社](https://baike.sogou.com/lemma/ShowInnerLink.htm?lemmaId=4852466&ss_c=ssc.citiao.link" \t "https://baike.sogou.com/_blank)，2013.

[4] 马林，数据重现:文件系统原理精解与数据恢复最佳实践[M].科学出版社，2009.

[5] A Python script for adaptive layout optimization of trusses[R]. Geneva:Robert Cimman, 2013.

附 录

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from pyecharts.charts.bar import Bar

%matplotlib inline

plt.style.use('ggplot')

import matplotlib as mpl

mpl.rcParams['font.sans-serif'] = ['SimHei'] #解决seaborn中文字体显示问题

plt.rc('figure', figsize=(10, 10)) #把plt默认的图片size调大一点

plt.rcParams["figure.dpi"] =mpl.rcParams['axes.unicode\_minus'] = False # 解决保存图像是负号'-'显示为方块的问题

data = pd.read\_csv('kriswu.csv', encoding='UTF-8')

#打印，以确保数据运行正常

Data

data.info()

data.sample(5)

#数据清洗

print(list(data.columns))

in\_columns = ['attitudes\_count', 'comments\_count', 'reposts\_count', 'mid', 'raw\_text','source','user.description','user.follow\_count','user.followers\_count','user.gender', 'user.id', 'user.mbrank', 'user.mbtype', 'user.profile\_url', 'user.profile\_image\_url','user.screen\_name', 'user.statuses\_count', 'user.urank', 'user.verified', 'user.verified\_reason']

data = data[in\_columns]

data.info()

data.to\_csv('kriswu.csv', index=False)

#数据分析

#吴亦凡该微博的转发是否存在假流量

# 吴亦凡的粉丝性别比例

fans\_num = data['user.gender'].value\_counts()

fans\_num

from pyecharts import Bar

bar = Bar("吴亦凡粉丝性别比例初探", width = 600,height=500)

bar.add("(总数据102118条)", ['女', '男'], fans\_num.values, is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True)

Bar

np.round(fans\_num/fans\_num.sum()\*100, 2)

data[data['user.gender']=='f'].sample(5)

data[data['user.gender']=='m'].sample(5)

#粉丝数大于等于5但昵称包含“用户”的这样的数据抽取出来。基本上可以断定是假粉丝

data\_fake2\_index = data[(data['user.follow\_count']>5)&

(data['user.followers\_count']>5)&

(data['user.screen\_name'].str.contains('用户'))].index

data\_fake2\_index

data\_fake2\_index.shape

# 昵称里包含“用户”的，基本上可以断定是假粉丝

data\_fake2\_index = data[(data['user.follow\_count']>5)&

(data['user.followers\_count']>5)&

(data['user.screen\_name'].str.contains('用户'))].index

# 把假的流量粉丝转发组合起来

data\_fake = pd.concat([ data.iloc[data\_fake2\_index]])

data\_fake.shape

# 取出真粉的转发

data\_true = data.drop(data\_fake.index)

data\_true.shape

print('真粉丝转发数占总转发数的{}%'.format(np.round(data\_true.shape[0]/data.shape[0]\*100, 2)))

print('假粉丝转发数占总转发数的{}%'.format(np.round(data\_fake.shape[0]/data.shape[0]\*100, 2)))

bar = Bar("吴亦凡真假流量的转发量", width = 600,height=500)

bar.add("(总数据102118条)", ['总转发量', '假粉丝转发量', '真粉丝转发量'],

[data.shape[0], data\_fake.shape[0], data\_true.shape[0]], is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True)

Bar

real\_fans\_num = data\_true.drop\_duplicates(subset='user.id').shape[0]

bar = Bar("吴亦凡真假流量的转发量与真实转发粉丝量(总数据102118条)", width = 600,height=500)

bar.add('', ['总转发量', '假粉丝转发量', '真粉丝转发量', '真实转发粉丝量'],

[data.shape[0], data\_fake.shape[0], data\_true.shape[0], real\_fans\_num], is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True, xaxis\_rotate=20)

Bar

# 吴亦凡真假流量的转发量与真实转发粉丝量

print('真实转发粉丝量占总转发数的{}%'.format(np.round(real\_fans\_num/data.shape[0]\*100, 2)))

#大家对于《大碗宽面》怎么看

data.info()

from snownlp import SnowNLP

def get\_sent\_snownlp(data):

s = SnowNLP(data)

return s.sentiments

data\_true['clean\_text'] = data\_true['raw\_text'].str.split("//", expand=True)[0]

ata\_true['sent\_nlp'] = data\_true.loc[data\_true['clean\_text']!='', 'clean\_text'].apply(get\_sent\_snownlp)

data\_true[['clean\_text', 'sent\_nlp']].sample(5)

data\_true.loc[-data\_true['clean\_text'].isin(['转发微博', 'repost', '轉發微博']), 'sent\_nlp'].mean()

from pyecharts.charts.gauge import Gauge

from pyecharts.charts.geo import Geo

g = Gauge()

g.add('', ['评论对《大碗宽面》\n的平均评分'], [68.6])

g

data\_true.loc[data\_true['sent\_nlp']>0.9, ['clean\_text', 'sent\_nlp']].sample(5)

data\_true.loc[data\_true['attitudes\_count'].sort\_values(ascending=False)[:10].index, ['user.screen\_name', 'clean\_text', 'attitudes\_count']]

data\_true.loc[data\_true['attitudes\_count'].sort\_values(ascending=False)[:100].index, ['user.screen\_name', 'clean\_text', 'attitudes\_count']].sample(5)

#有多少人拿吴亦凡跟蔡徐坤做对比

data\_true.info()

data\_true['clean\_text'].sample(10)

data\_true['clean\_text'].sample(10)

data\_true['raw\_text'].str.contains('kun|坤|律师|球|函|cxk|比|弟|CXK|胸|格局|气度|衬托').sum()

data\_kun = data\_true.loc[data\_true['raw\_text'].str.contains('kun|坤|律师|球|函|cxk|比|弟|CXK|胸|格局|气度|衬托'),

['user.screen\_name', 'raw\_text', 'attitudes\_count']]

data\_kun.loc[list(data\_kun['attitudes\_count'].nlargest(10).index), ['raw\_text', 'attitudes\_count']]

data\_true[data\_true['raw\_text']!=''].shape

6229/96018

# 有多少人开始路转粉了

data\_true['clean\_text'].sample(10)

data\_true['raw\_text'].str.contains('转粉|爱上|重新|路|圈粉|espect|瑞思拜').sum()

#计算评论中含有关键词“转粉|爱上|重新|路|圈粉|espect|瑞思拜”的评论数量

data\_true.loc[data\_true['raw\_text'].str.contains('转粉|爱上|重新|路|圈粉|espect|瑞思拜'),

['user.screen\_name', 'raw\_text']].sample(5)

提3取部分含有这些关键词的评论并打印出来

fans = data\_true.loc[data\_true['raw\_text'].str.contains('转粉|爱上|重新|路|圈粉|espect|瑞思拜'),

'user.gender'].value\_counts()

from pyecharts.charts.bar import Bar

bar = Bar("路转粉的男女性别比例", width = 600,height=500)

bar.add("", ['女', '男'], fans.values, is\_stack=True,

xaxis\_label\_textsize=20, yaxis\_label\_textsize=14, is\_label\_show=True)

Bar

fans/fans.sum()

# 绘制评论词云图

import jieba

from collections import Counter

from pyecharts.charts.wordcloud import WordCloud

jieba.add\_word('吴亦凡')

swords = [x.strip() for x in open ('stopwords.txt', encoding='UTF-8')]

def plot\_word\_cloud(data, swords):

text = ''.join(data)

words = list(jieba.cut(text))

ex\_sw\_words = []

for word in words:

if len(word)>1 and (word not in swords):

ex\_sw\_words.append(word)

c = Counter()

c = Counter(ex\_sw\_words)

wc\_data = pd.DataFrame({'word':list(c.keys()), 'counts':list(c.values())}).sort\_values(by='counts', ascending=False).head(100)

wordcloud = WordCloud(width=1300, height=620)

wordcloud.add("", wc\_data['word'], wc\_data['counts'], word\_size\_range=[12, 60])

return wordcloud

plot\_word\_cloud(data=data\_true.loc[-data\_true['clean\_text'].str.contains("转发"), 'clean\_text'], swords=swords)

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | **吴亦凡微博热点的分析** | | |
| 课程设计学生答辩或质疑记录：  1.此次设计运用到了哪些新的工具？  运用到的新的工具有Pyecharts、Numpy、Seaborn、Snownlp。  2.snownlp的作用是什么？ 答：利用snownlp可以进行分词，词性标注，文本摘要提取，文本情感分析。  3.在绘图中怎么进行颜色设置？ 答：调用matplotlib.pyplot.colors()可以得到matplotlib支持的所有颜色，其中，k表示黑色，m表示红色，c表示青色，w表示白色。 | | | |
| **评 分 依 据** | | **分 值** | **评分成绩** |
| 1．团队协作能力 | | 25分 |  |
| 2．python综合运用能力 | | 25分 |  |
| 3．态度认真、刻苦钻研、创新能力 | | 10分 |  |
| 4．过程完成、对工具的使用、对github的运用 | | 20分 |  |
| 5．课程设计答辩逻辑清晰，内容正确 | | 10分 |  |
| 6. 课程设计期间的课堂考勤、遵守纪律 | | 10分 |  |
| 总 分 | | 100分 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 15日 | | | |