|  |  |
| --- | --- |
| 学号 | 10214518230 |



课 程 设 计

课程名称 Python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 吴亦凡微博热点分析 |
| 专 业 | \_\_\_\_软件工程\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 班 级 | \_\_\_\_\_软件1182\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 姓 名 | \_\_\_\_\_罗新茹\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 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 对数据的读取、数据清洗、数据分析、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18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（jupyter notebook）、课程设计报告每人一份、答辩PPT每小组一份，按小组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18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。

每个班每个小组题目不可重复。如果各小组发现更有意思的题目，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、对数据的处理和分析能力、以及答辩的表达能力等。

**三、学时分配进度安排**

|  |  |  |
| --- | --- | --- |
| 序号 | 设计内容 | 所用时间 |
|  | 下发任务书， | 1天 |
|  | 学生查阅资料，组建团队，完成选题 | 1天 |
|  | 各小组准备好数据，确定分析目标 | 2天 |
|  | 完成数据清洗以及数据分析 | 3天 |
|  | 形成初步的报告书和ppt | 1天 |
|  | 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**

整个过程要求使用jupyter notebook完成开发。

**3.评分标准**

|  |  |
| --- | --- |
| **评分依据** | **评分成绩** |
| 1．设计方案正确，具有可行性、创新性 | 25分 |
| 2．系统完成的情况以及对开发工具运用的熟练程度 | 25分 |
| 3．态度认真、刻苦钻研、遵守纪律以及课堂考勤 | 10分 |
| 4．设计报告规范、课程设计报告质量高 | 20分 |
| 5．课程设计答辩逻辑清晰，内容正确 | 10分 |
| 6. 团队的协作能力 | 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](#_Toc7486)

[2. 开发环境 2](#_Toc15400)

[3. 数据清洗 2](#_Toc23320)

[4. 数据分析 4](#_Toc1194)

[5. 结论和建议 12](#_Toc18562)

[6. 课设总结 13](#_Toc68)

[参 考 文 献 15](#_Toc16463)

[课程设计成绩评定表 20](#_Toc25029)

# 

# 设计题目

吴亦凡微博的热点分析

## 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，选取吴亦凡和蔡徐坤的微博中100万+转发的微博，探求是否存在假流量？假流量所占比重有多少？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）和stopwords.txt（ 停用词表）

## 分析目标

本次课设将从以下五个方面对数据进行分析：

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

软件环境：Microsoft Windows 10

Python版本：python3.8

其他第三方库版本：使用的主流Notebook环境是jupyter Notebook.需要安装的包有：pyecharts，pandas，numpy，matplotlib， seaborn，snownlp

# 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，代码如下：

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']

in\_columns

data = data[in\_columns]

data

data.info()

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

data.to\_csv

清洗后的结果如下图所示：



图3.1数据清洗图（1）

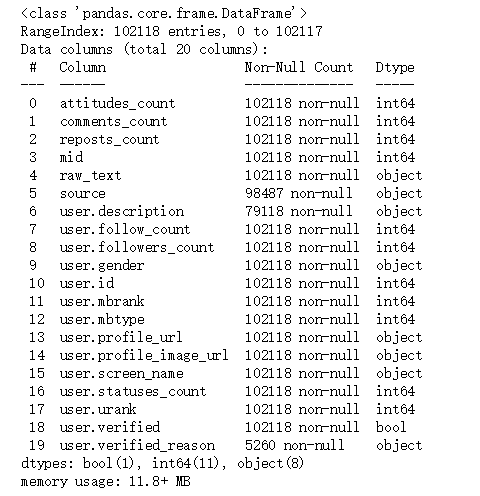


图3.2数据清洗图（2）

# 数据分析

## 吴亦凡该微博的转发是否存在假流量？分析

先通过看看吴亦凡的粉丝性别比例，代码如下

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

fans\_num

通过柱形图来展示男女粉丝比例：

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

如下图所示：

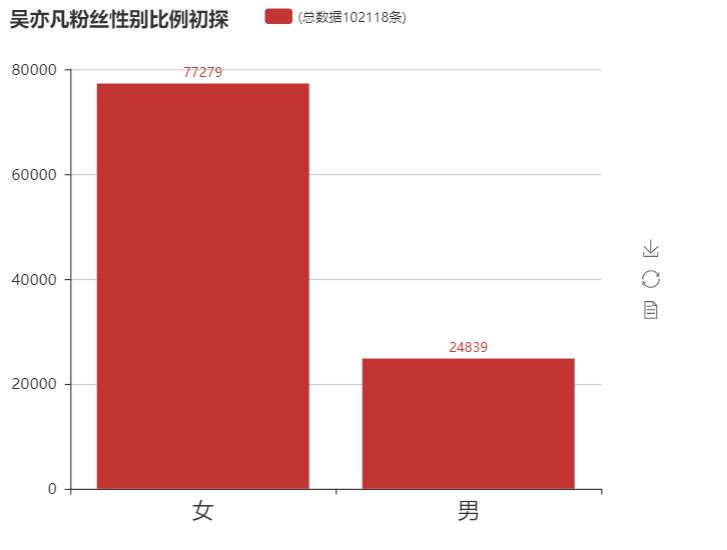


图4.1吴亦凡粉丝性别比例初探

然后，进行真假流量的统计，代码如下：

#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，代码如下：

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

data\_fake2\_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

#取出真粉的转发

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

data\_true

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)))

得出：真粉丝转发数占总转发数的99.58%

假粉丝转发数占总转发数的0.42%

再通过柱状图显示：

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

如下图所示：

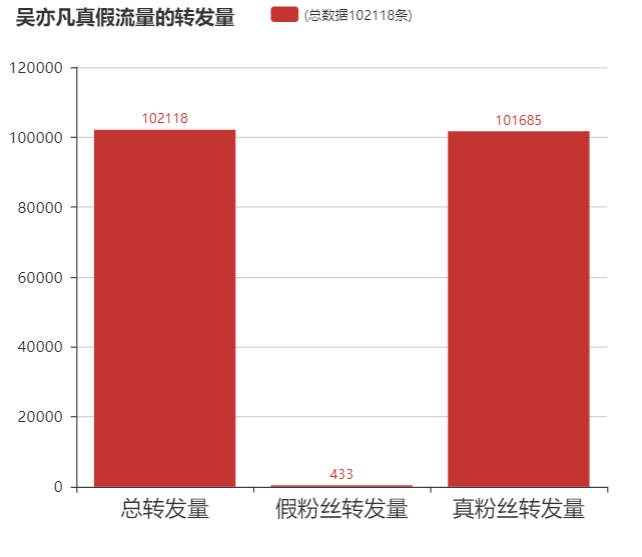


图4.2吴亦凡真假流量的转发量

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

如下图所示：

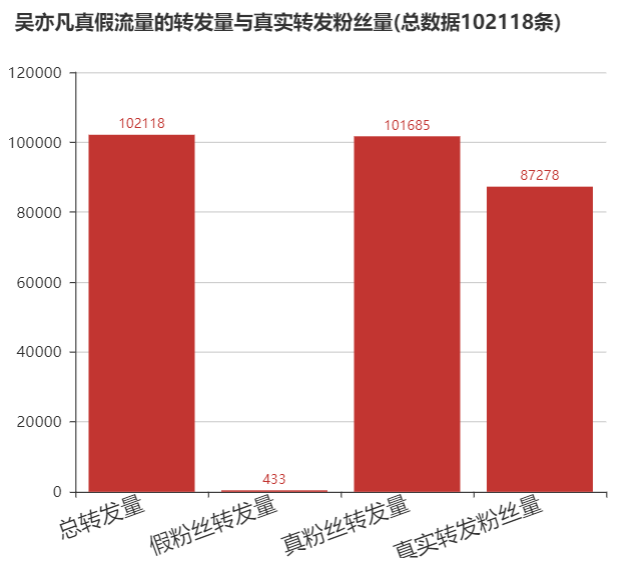


图4.3 吴亦凡真假流量的转发量与真是转发粉丝量

最后得出结论：根据柱状图显示，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。

## 大家对于《大碗宽面》怎么看？分析

首先通过情绪分析大众对于《大碗宽面》的看法，用到snownlp情绪分析法：

from snownlp import SnowNLP

def get\_sent\_snownlp(data):

s = SnowNLP(data)

return s.sentiments

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

结果如图所示：

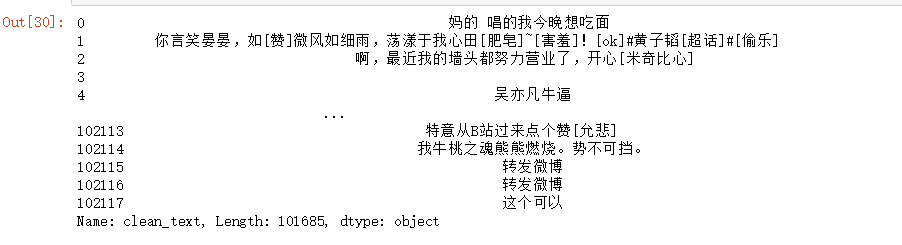


图4.4 《大碗宽面》评价

再通过扇形图展示：

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

如图所示：



图4.5 《大碗宽面》平均分

其中前五的评价如下：

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

如图所示：

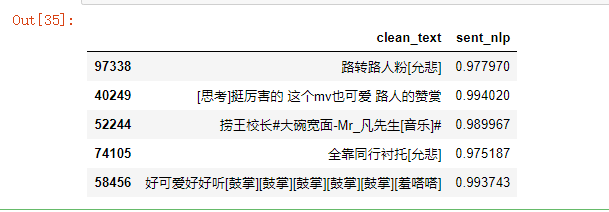


图4.6 《大碗宽面》前五评价

得出结论：

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%。

## 有多少人拿吴亦凡跟蔡徐坤做对比？ 分析

通过读取的数据中我们抽取了前十的对比评价：

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

结果如下图所示：

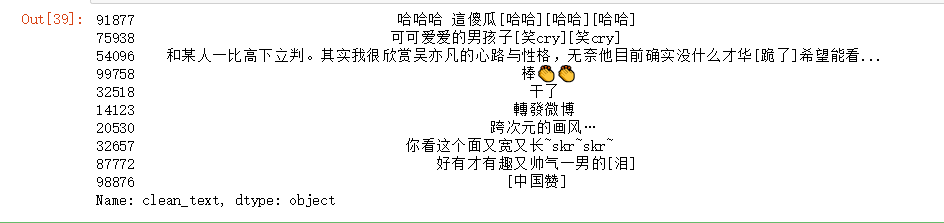


图4.7 前十条对比

再通过sum进行计算总共有多少人拿吴亦凡与蔡徐坤做对比：

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

得到：有6388人

然后查看大家对蔡徐坤的评价:

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']]

如下图所示：

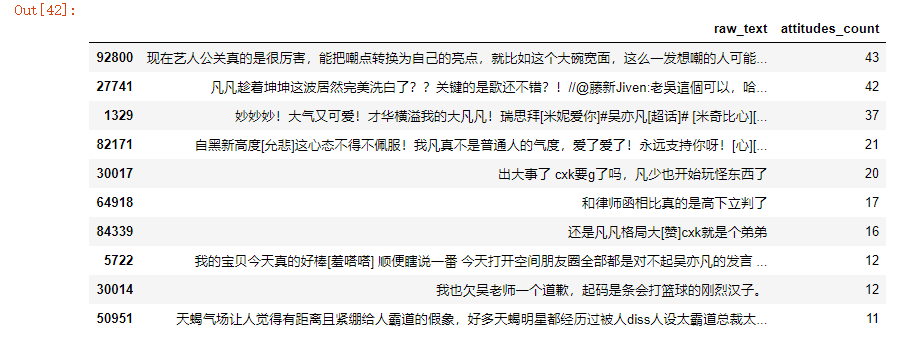


图4.8 蔡徐坤评价

得出结论：根据提取的原始文本数据知道有6388的人拿吴亦凡跟蔡徐坤作比较。

## 有多少人开始路转粉了？分析

通过sum函数统计路转粉人数：

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

再提取前五条路转粉内容：

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)

如下图所示：

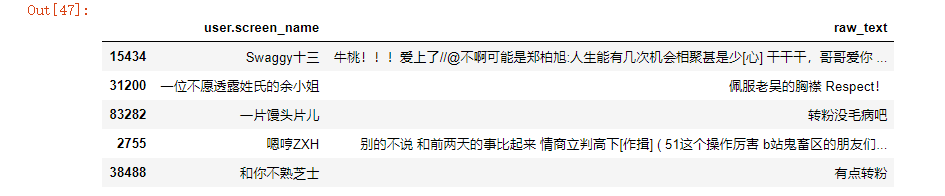


图4.9 前五条数据

最后通过柱形图展示路转粉男女比例：

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

如下图所示：

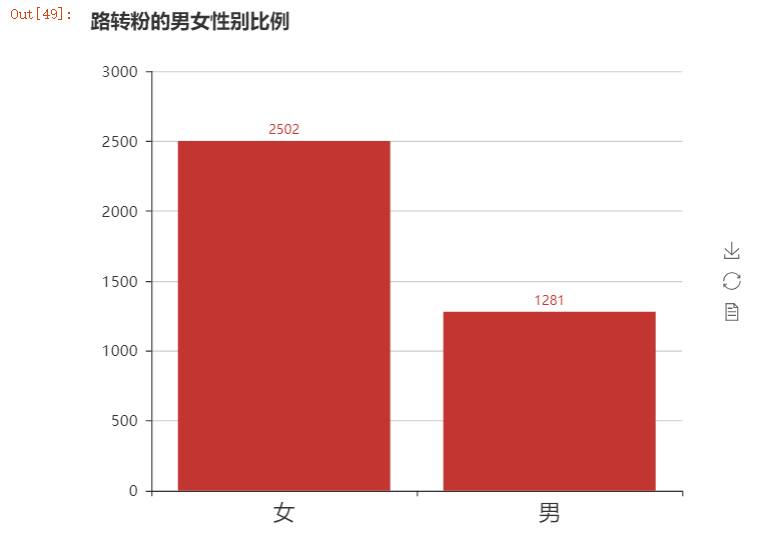


图4.10 路转粉的男女比例

得出结论：通过提取数据显示有3783人开始路转粉，路转粉的男生是1281人次，女生是2502人次。

## 词云图

通过jieba分析吴亦凡评论的分词模块，再通过wordcloud更加直观和艺术的展示文本

# 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)

如下图所示：



图4.11 词云图

# 结论和建议

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.路转粉的人也挺多的，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

问题：1.一开始，柱形图和扇形图输出是空白的没有图显示出来

解决：跟pyecharts的版本有关系，一开始下的0.1.9.4的版本不能显示，后来卸载了，改成0.5.11版本，在重新启动内核，就有图显示出来了。

问题：2.在情绪分析的时候，代码运行有错误

解决：要下载snownlp才能运行成功，而且由于数据庞大，运行时间较长。

## 设计体会

经过这次课设，对Python学习以及数据处理有了更多的理解和掌握，学会了使用以下工具：

1 jieba 在分析中是使用最广泛的分词模块。

2 wordcloud 可以更加直观和艺术的展示文本

3 Pandas、numpy、ipython 要处理大型数据集常用的Python模块。

4 Matplotlib 一个Python的图形框架，用于绘制二维图形。

5 pyechars 用来解决数据分析可视化的问题。

6 Seaborn 是基于matplotlib的图形可视化python包。它提供了一种高度交互式界面，便于我们在分析中能够做出多张有吸引力的统计图表。

在设计过程中虽然遇到了一些问题，但经过一次又一次的思考，一遍又一遍的检查终于找出了原因所在，也暴露出了前期我在这方面的知识欠缺和经验不足。实践出真知，通过亲自动手制作，使我们掌握的知识不再是纸上谈兵。在课程设计过程中，我们持续发现错误，持续改正，持续领悟，持续获取。最终的检测调试环节，本身就是在践行“过而能改，善莫  
大焉”的知行观。这次课程设计终于顺利完成了，在设计中遇到了很多问题，最后在老师的指导下，终于游逆而解。在今后社会的发展和学习实践过程中，一定要不懈努力，不能遇到问题就想到要退缩，一定要不厌其烦的发现问题所在，然后一一实行解决，只有这样，才能  
成功的做成想做的事，才能在今后的道路上劈荆斩棘，而不是知难而退，那样永远不可能收获成功，收获喜悦，也永远不可能得到社会及他人对你的认可！  
 回顾起此课程设计，至今我仍感慨颇多，从理论到实践，在这段日子里，能够说得是苦多于甜，但是能够学到很多很多的东西，同时不但能够巩固了以前所学过的知识，而且学到了很多在书本上所没有学到过的知识。通过这次课程设计使我懂得了理论与实际相结合是很  
重要的，只有理论知识是远远不够的，只有把所学的理论知识与实践相结合起来，从理论中得出结论，才能真正为社会服务，从而提升自己的实际动手水平和独立思考的水平。在设计的过程中遇到问题，能够说得是困难重重，但可喜的是最终都得到了解决。实验过程中，也对团队精神的实行了考察，让我们在合作起来更加默契，在成功后一起体会喜悦的心情。果然是团结就是力量，只有互相之间默契融洽的配合才能换来最终完美的结果。

**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. 钱程，阳小兰等.基于python的网络爬虫技术【J】.黑龙江科技信息，2016
3. 戚利娜，刘建东.基于python的简单网络爬虫的实现【J】.电脑编程与维护.2017
4. Wesley.J.C,Core Python Programming.2001-9-11

[5] 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.

附 录

# 1.csv文件数据读取

# 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)

2.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']

in\_columns

data = data[in\_columns]

Data

data.info()

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

data.to\_csv

1. 数据分析

# 先来看看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

data\_fake2\_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

data\_fake.shape

# 取出真粉的转发

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

data\_true

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]

real\_fans\_num

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

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)))

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]

data\_true['clean\_text']

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

data\_true['sent\_nlp']

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)

data\_true.info()

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

6338/101685

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()

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)

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()

1. 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 | **明星微博的热点分析** | | |
| 课程设计学生答辩或质疑记录：   1. 为什么要使用jieba？   答：目的是通过jieba分词库统计热评中的高频词汇。   1. 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。   1. 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．设计方案正确，具有可行性、创新性 | | 25分 |  |
| 2．系统完成的情况以及对开发工具运用的熟练程度 | | 25分 |  |
| 3．态度认真、刻苦钻研、遵守纪律以及课堂考勤 | | 10分 |  |
| 4．设计报告规范、课程设计报告质量高 | | 20分 |  |
| 5．课程设计答辩逻辑清晰，内容正确 | | 10分 |  |
| 6. 团队的协作能力 | | 10分 |  |
| 总 分 | | 100分 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 23 日 | | | |