|  |  |
| --- | --- |
| 学号 | 10214518409 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 数据分析的学习与实践-蔡徐坤粉丝数及转发数据真假状况分析 |
| 专 业 | 软件工程 |
| 班 级 | 软件1182 |
| 姓 名 | 章华鼎 |
| 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、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 |

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.分析目标 1](#_Toc62213750)

[2.分析流程 1](#_Toc62213751)

[3.分析实施 1](#_Toc62213752)

[3.1数据获取 1](#_Toc62213753)

[3.2数据预处理 4](#_Toc62213754)

[4.数据分析 4](#_Toc62213755)

[4.1. 蔡徐坤的微博转发是否存在假流量的分析 4](#_Toc62213756)

[4.2. 探究真假流量所占的比例分别是多少 5](#_Toc62213757)

[4.3. 假流量粉丝是如何产生的（对data\_fake进行讨论） 7](#_Toc62213758)

[4.4 真流量粉的粉丝画像 10](#_Toc62213759)

[5.思考 15](#_Toc62213760)

[6.总结 15](#_Toc62213761)

# 分析目标

* 1. 蔡徐坤的微博转发是否存在假流量？
  2. 真假流量所占的比例各有多少？
  3. 假流量粉丝是如何生产出来的？
  4. 真流量粉的粉丝画像是怎样的？

# 2.分析流程

①数据获取:找出需要用作数据分析的数据并进行获取。

②数据预处理:对获取数据进行去重、缺失值处理等一系列数据清洗操作。

③数据可视化:文字不如数据，数据不如图片，可视化是对于数据最好的展现方式，让人一目了然。

④深入分析:对于数据深入分析，辨别蔡徐坤真假粉丝实际占比是多少。

# 3.分析实施

## 3.1数据获取

因为数据由老师提供，这里仅放爬虫的代码，代码如下:

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

# Scrapy settings for CaiXuKun project

#

# For simplicity, this file contains only settings considered important or

# commonly used. You can find more settings consulting the documentation:

#

# http://doc.scrapy.org/en/latest/topics/settings.html

# http://scrapy.readthedocs.org/en/latest/topics/downloader-middleware.html

# http://scrapy.readthedocs.org/en/latest/topics/spider-middleware.html

import os

BOT\_NAME = 'CaiXuKun'

SPIDER\_MODULES = ['CaiXuKun.spiders']

NEWSPIDER\_MODULE = 'CaiXuKun.spiders'

# Crawl responsibly by identifying yourself (and your website) on the user-agent

#USER\_AGENT = 'CaiXuKun (+http://www.yourdomain.com)'

# Obey robots.txt rules

ROBOTSTXT\_OBEY = False

# Configure maximum concurrent requests performed by Scrapy (default: 16)

CONCURRENT\_REQUESTS = 32

# Configure a delay for requests for the same website (default: 0)

# See http://scrapy.readthedocs.org/en/latest/topics/settings.html#download-delay

# See also autothrottle settings and docs

# DOWNLOAD\_DELAY = 2

# The download delay setting will honor only one of:

#CONCURRENT\_REQUESTS\_PER\_DOMAIN = 16

#CONCURRENT\_REQUESTS\_PER\_IP = 16

# Disable cookies (enabled by default)

#COOKIES\_ENABLED = False

# Disable Telnet Console (enabled by default)

#TELNETCONSOLE\_ENABLED = False

# Override the default request headers:

DEFAULT\_REQUEST\_HEADERS = {

'Accept': 'application/json, text/plain, \*/\*',

'MWeibo-Pwa': '1',

'Referer': 'https://m.weibo.cn/status/Hk1YczUdT',

'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_13\_5) AppleWebKit/537.36 (KHTML, like Gecko)'

' Chrome/72.0.3626.119 Safari/537.36',

'X-Requested-With': 'XMLHttpRequest'

}

# Enable or disable spider middlewares

# See http://scrapy.readthedocs.org/en/latest/topics/spider-middleware.html

#SPIDER\_MIDDLEWARES = {

# 'CaiXuKun.middlewares.CaixukunSpiderMiddleware': 543,

#}

# Enable or disable downloader middlewares

# See http://scrapy.readthedocs.org/en/latest/topics/downloader-middleware.html

DOWNLOADER\_MIDDLEWARES = {

'CaiXuKun.middlewares.TooManyRequestsRetryMiddleware': 543,

}

RETRY\_HTTP\_CODES = [418]

RETRY\_TIMES = 2

# Enable or disable extensions

# See http://scrapy.readthedocs.org/en/latest/topics/extensions.html

#EXTENSIONS = {

# 'scrapy.extensions.telnet.TelnetConsole': None,

#}

# Configure item pipelines

# See http://scrapy.readthedocs.org/en/latest/topics/item-pipeline.html

ITEM\_PIPELINES = {

'CaiXuKun.pipelines.CaixukunPipeline': 300,

}

# Enable and configure the AutoThrottle extension (disabled by default)

# See http://doc.scrapy.org/en/latest/topics/autothrottle.html

#AUTOTHROTTLE\_ENABLED = True

# The initial download delay

#AUTOTHROTTLE\_START\_DELAY = 5

# The maximum download delay to be set in case of high latencies

#AUTOTHROTTLE\_MAX\_DELAY = 60

# The average number of requests Scrapy should be sending in parallel to

# each remote server

#AUTOTHROTTLE\_TARGET\_CONCURRENCY = 1.0

# Enable showing throttling stats for every response received:

#AUTOTHROTTLE\_DEBUG = False

# Enable and configure HTTP caching (disabled by default)

# See http://scrapy.readthedocs.org/en/latest/topics/downloader-middleware.html#httpcache-middleware-settings

#HTTPCACHE\_ENABLED = True

#HTTPCACHE\_EXPIRATION\_SECS = 0

#HTTPCACHE\_DIR = 'httpcache'

#HTTPCACHE\_IGNORE\_HTTP\_CODES = []

#HTTPCACHE\_STORAGE = 'scrapy.extensions.httpcache.FilesystemCacheStorage'

MONGODB\_HOST = os.environ.get('MONGODB\_HOST', '127.0.0.1') #本地数据库

MONGODB\_PORT = os.environ.get('MONGODB\_PORT', '27017') #数据库端口

MONGODB\_URI = 'mongodb://{}:{}'.format(MONGODB\_HOST, MONGODB\_PORT)

MONGODB\_DATABASE = os.environ.get('MONGODB\_DATABASE', 'CaiXuKun') #数据库名字

## 3.2数据预处理

对抓取完的数据，首先进行的是数据清洗，然后进行有效字段的选取，最后只留下唯一的数据。对caixukun.csv文件的数据清洗如图3.1所示。

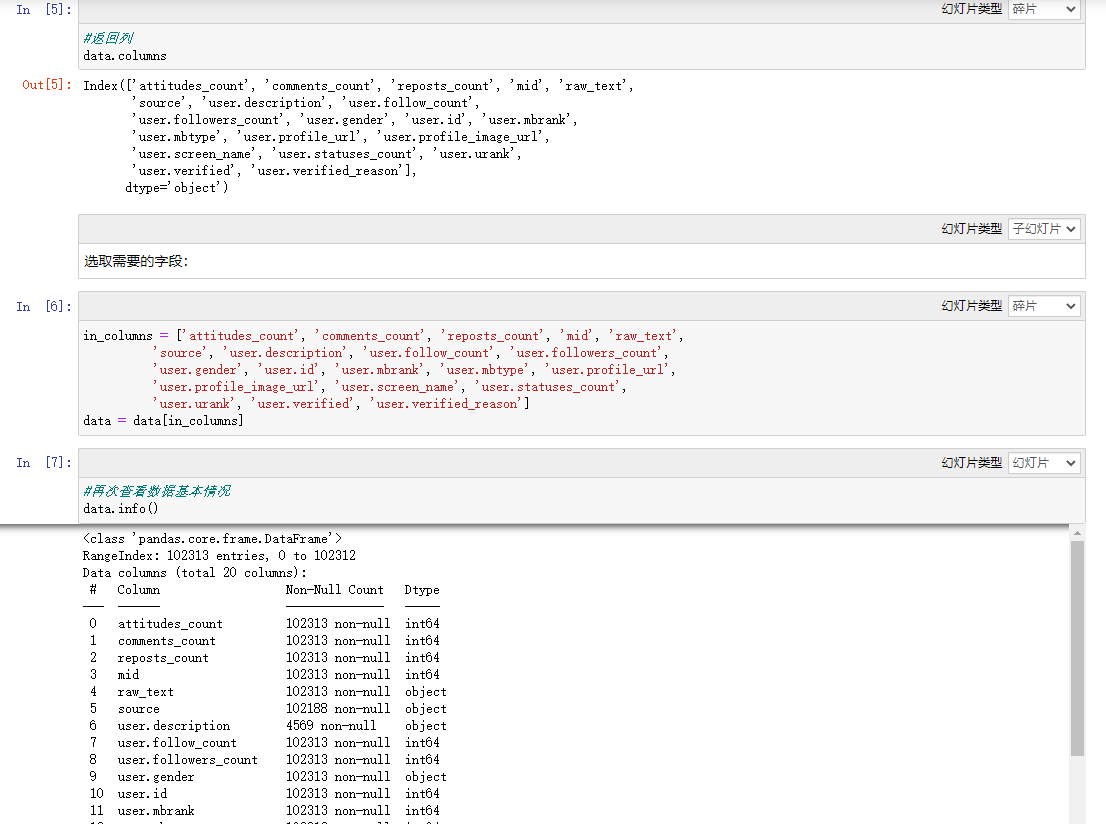


图3.1 清洗数据

# 4.数据分析

## 4.1. 蔡徐坤的微博转发是否存在假流量的分析

蔡徐坤粉丝的性别比例初探情况如下图4.1所示：

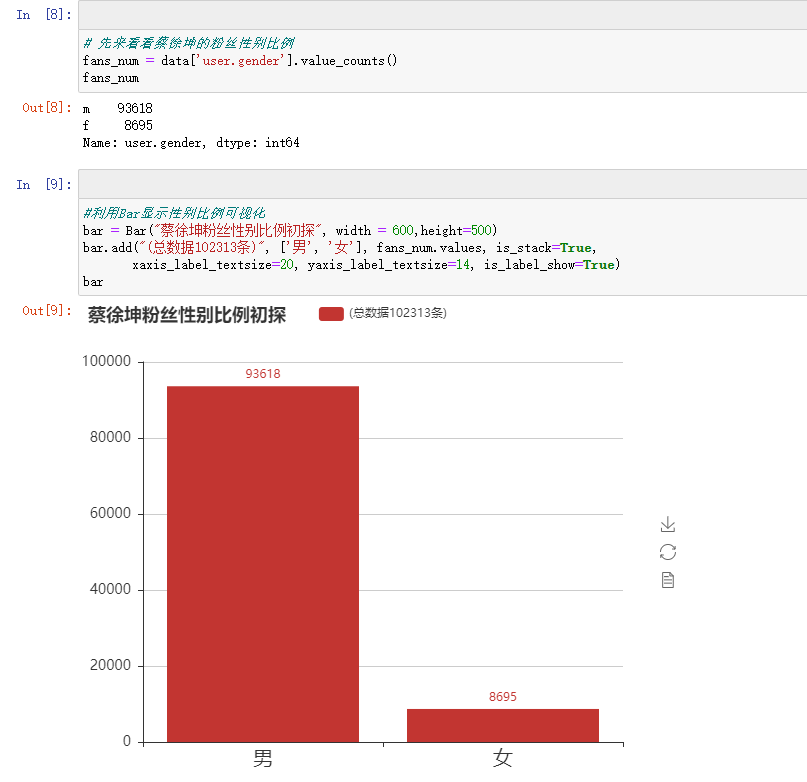


图4.1 蔡徐坤粉丝性别比例初探

从上图中我们可以初步看到蔡徐坤的男女粉丝比例大致是91.5：8.5，这有点不合乎情理，我们可以大致判断其中应该有不少假流量粉丝。

## 4.2. 探究真假流量所占的比例分别是多少

通过一番探索分析，我们把转发数据中转发者的关注或者粉丝数少于等于5、筛选出用户描述为‘wu’，发表评论、转发之后被点赞数评论数再转发数都为0、微博会员等级为0级的数据，以及转发者的关注这部分数据，便是我们所说的假流量。

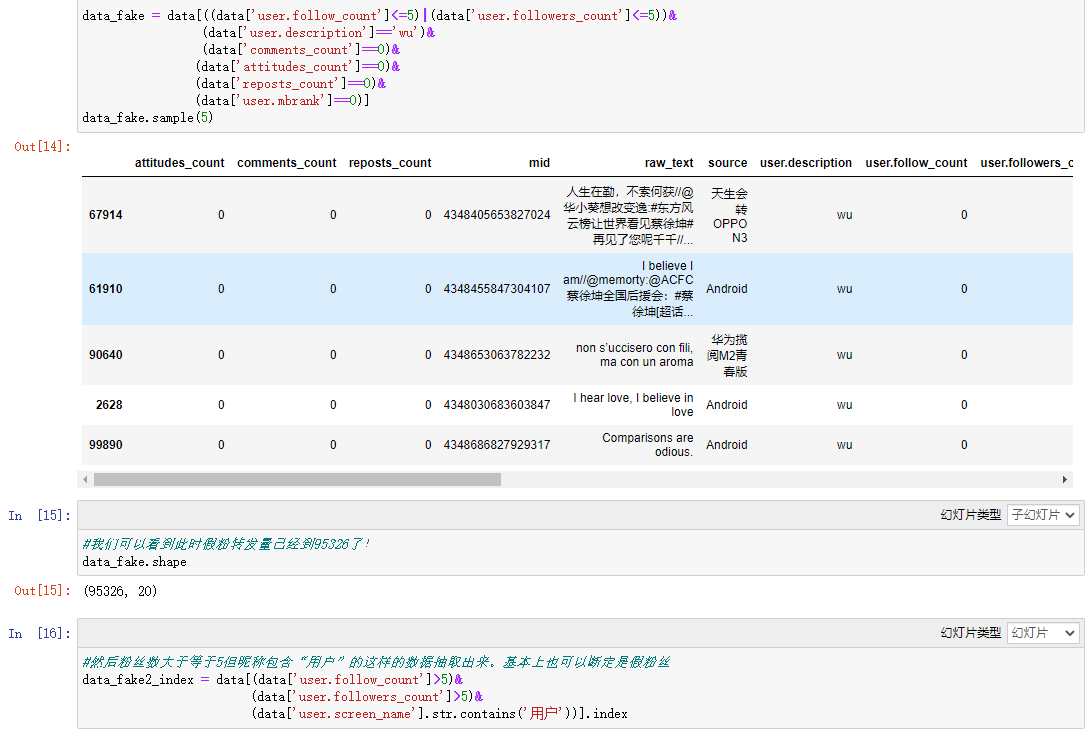


图4.2 蔡徐坤部分假粉丝数据信息

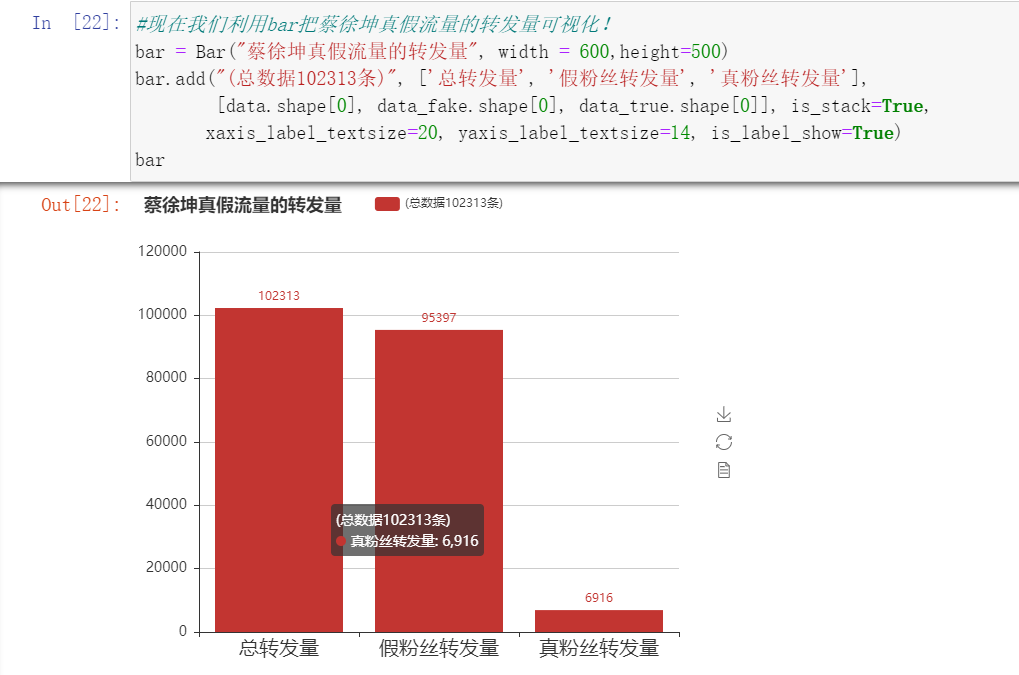


图4.3 蔡徐坤真假浏览转发量

从上图我们可以看到，在102313条转发数据中，有95397条是由假粉丝转发的，占了总转发的93.24%，只有6916条是由真粉丝转发的，占6.76%。

那么6919条真粉丝转发的数据中，除去重复转发刷榜的数量，里面一共有多少个真粉丝在转发呢？我们把这部分数据按照粉丝微博ID进行去重。



图4.4 蔡徐坤真假粉丝转发量

从上图可以看出，真实转发粉丝量仅占总转发数的3.84%！

### 4.3. 假流量粉丝是如何产生的（对data\_fake进行讨论）

先来看看构成蔡徐坤假流量粉丝的性别比例



图4.5 蔡徐坤假粉丝男女比例

从上图可见在95397条假转发中，有40838个假粉丝。其中男性的比例高达95.42%！

再来看看蔡徐坤假粉丝使用设备TOP10

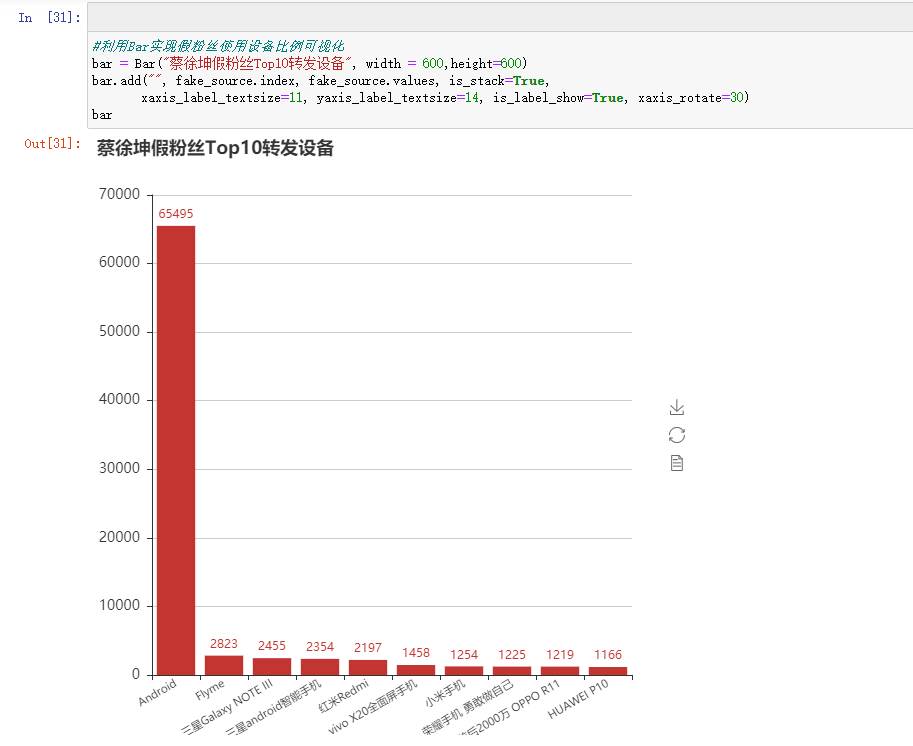


图4.6 蔡徐坤假粉丝使用设备TOP10

在假粉丝使用的Top10转发设备中，Android高居榜首，这也再次证明了这些粉丝是假粉丝。

下面是假粉丝的一些信息分析情况：



图4.7 蔡徐坤假粉丝相关信息

## 4.4 真流量粉的粉丝画像

先来看看真粉丝的性别比例。可见3926个真粉丝中，女生（f）占了绝大部分，这才是符合逻辑的粉丝比例嘛。

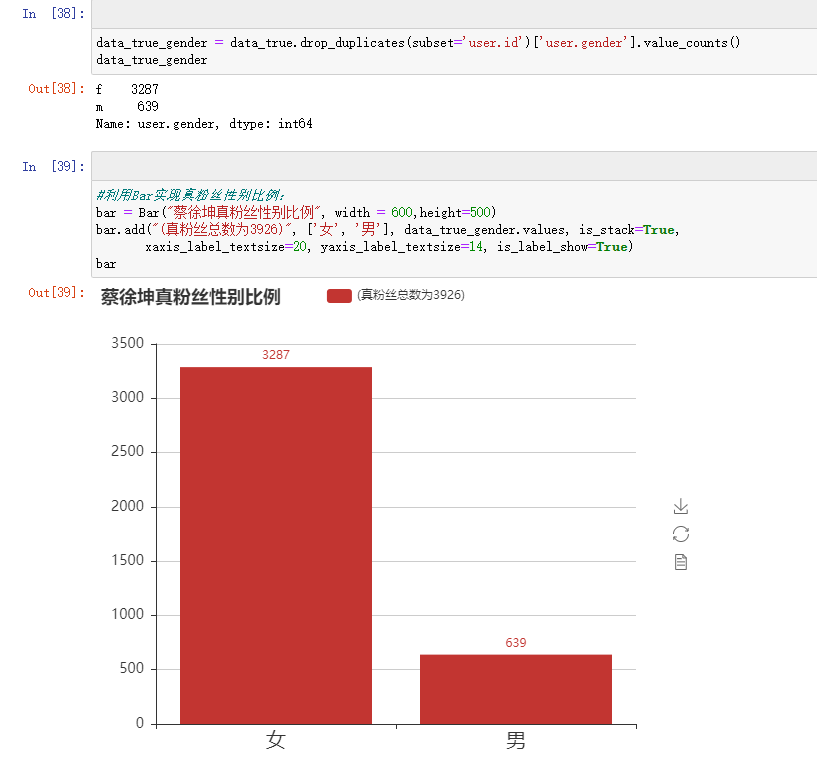


图4.8 蔡徐坤真粉丝男女比例

部分蔡徐坤真粉丝转发微博评论

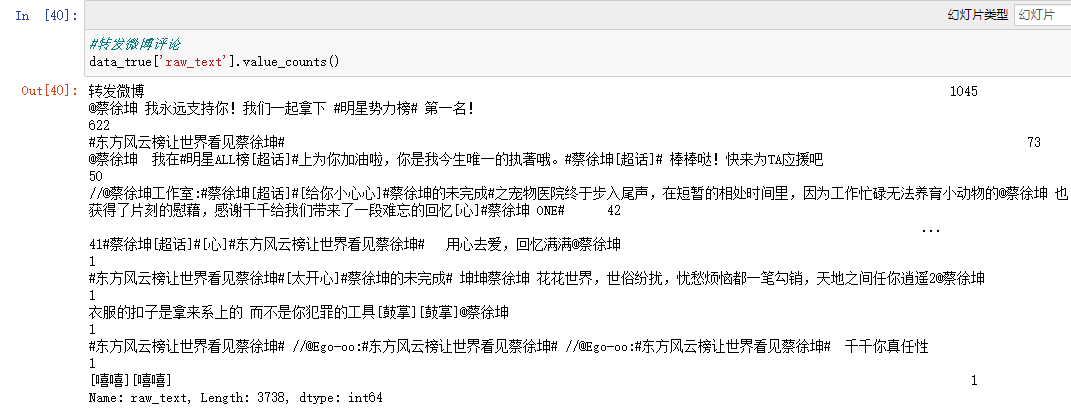


图4.9 部分蔡徐坤真粉丝转发微博评论

在这些粉丝转发所携带的评论中可以看到，很多都是支持蔡徐坤拿下#明星势力榜#或者#东方风云榜#第一名而转发的。

蔡徐坤真粉丝转发设备TOP10情况

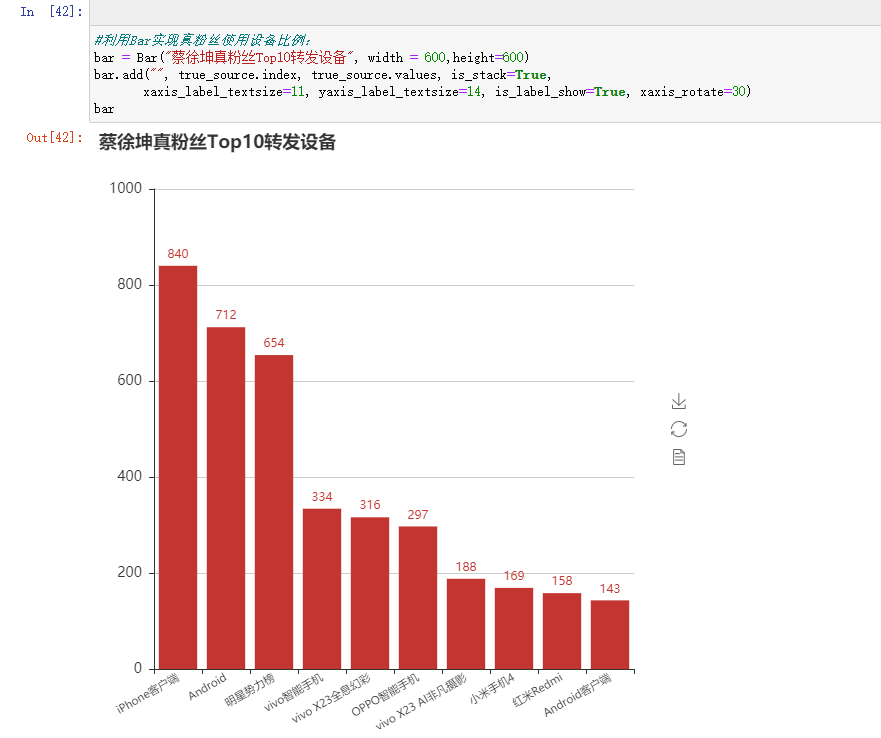


图4.10 蔡徐坤真粉丝TOP10转发设备

可见真粉丝所使用的转发设备，各种设备分布都比较均匀，最受欢迎的是iPhone客户端。

下面是一些真粉丝中的信息，真粉丝的平均关注量为222，平均粉丝数是179，同样，很多粉丝名字中喜欢带有“坤”、“蔡”、“葵”、“kun”等字。



图4.11 蔡徐坤真粉丝名字特点

我们把真粉丝的简介做成了词云图。

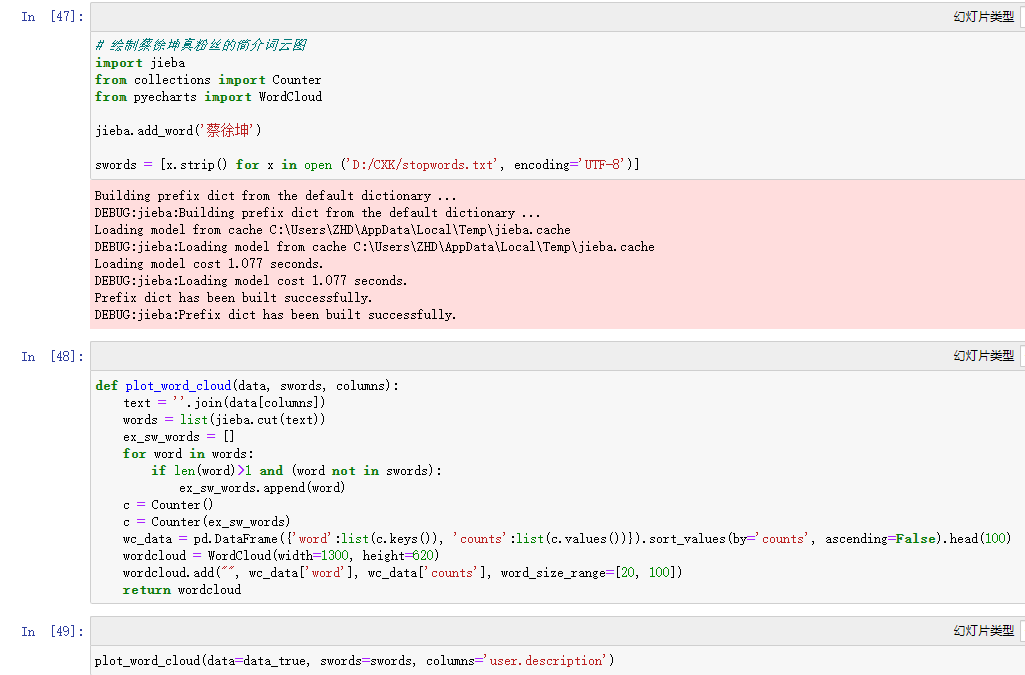


图4.12 蔡徐坤真粉丝简介词云图代码



图4.13 蔡徐坤真粉丝简介词云图

可见，真粉丝们的简介都喜欢带上蔡徐坤的名字，很喜欢蔡徐坤，想要一直陪着他走下去。看着里面的“少年”、“努力”、“自由”“追梦”等字眼，发现这就是我们以前的青春呀。

另外我们还把转发所带的评论做成了词云图。



图4.14 蔡徐坤真粉丝评论词云图

可见，粉丝们都很在乎“风云榜”这件事，并且要帮助蔡徐坤拿第一名。里面有早安打卡的，有超级话题的，还有很多“开心”、“比心”、“温暖”等字眼。说明绝大部分真粉还是很温暖的。

# 5.思考

此次课程设计主要是对之前学习的使用可视化图表工具分析数据的实践，真实的数据分析远不止这些，还需要扎实的教学、统计学基础，良好的数据敏感度，开拓但严谨的思维等，是一个终生学习的过程。

# 6.总结

在本次Python课程设计中，我查阅了包括Github的使用在内的许多资料，在徐勇老师和同组成员的帮助下，顺利完成了这个蔡徐坤粉丝数及转发数据真假状况的数据分析。在这个过程中，我学到了很多，如怎样去全面的看待并分析一个问题等，同时也在总体上巩固了自己的Python知识，发现自己之前从来没有像现在一样将零散的知识点结合在一起，完成了这个蔡徐坤粉丝数及转发数据真假状况的数据分析后，让我在整体上重新认识了Python。

在这次的数据分析中，因为有平时学习的基础，所以完成本次Python课设的效率也较高，在设计过程中虽然遇到了一些问题，但经过多次的思考及反复的实践，检查出问题可能出现的原因，从而也暴露出了我在这方面知识的欠缺和经验的不足，这使我意识到，自己学习还只停留在表面，所以，我将努力在今后的时间里改正，并将所学的知识吃透、弄懂。

我在本次课程设计中学到了不少新知识或新技能，首先就是我学会了Github的使用，从原来的不知如何下手到现在的能够熟练使用Github上传各种课设相关的资料供组员参考；其次我巩固了这学期的Python课程中学习到的知识，还了解到了一些课上没讲过的包，比如scrapy、pyecharts、requests、seaborn等；然后我从本次课设中感受到了团队的重要性，懂得了如何协同组员一起完成一项工程，提升了人际交往能力；最后就是PPT制作能力了，因为我是本次PPT制作的主要负责人，所以通过此次课设，我的PPT制作能力也上升了一个档次。

我要感谢徐勇给予我们的指导、提供给我们的支持和帮助，导师渊博的知识，严谨的治学态度，一丝不苟的工作作风，平易近人的性格都是我们学习的楷模。徐老师平时讲学很仔细，这学期对我们指导性建设要求提出了很多意见，促进了当代教育的发展。在课设期间，老师给了我们很大的支持和鼓励，才使得课设得以顺利的完成，在此谨向导师表示忠心的感谢和崇高的敬意，这是我们能顺利完成本次课设的主要原因，在此期间，我们不仅学到了许多新的知识，而且也开阔了视野，提高了自己的能力。

本次的Python课程设计还有许多的不足之处，但今后我会更加努力地巩固所学的知识，希望自己能够学以致用，不辜负老师的教导。在这次的设计中，我学会了很多新的知识，那是我以前未曾感觉到的，同时也感到自身知识的缺乏，在日后的学习和工作中能做的更好。最后再次感谢同学的帮助和老师的悉心教导！

**参考文献**

[1] 丁辉.Python基础与大数据应用.北京：人民邮电出版社，2020.1

[2] 董付国.Python可以这样学[M].北京:清华大学出版社，2017

[3] 刘凌霞，郝宁波，吴海涛.21天学通Python[M].北京:电子工业出版社，2016

[4] 基于Python的数据可视化[J].罗博炜.信息记录材料. 2019(12)

[5] 刘宇宙.Python 3.5从零开始学[M].北京:清华大学出版社，2017

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 蔡徐坤粉丝数及转发数据真假状况分析 | | |
| 课程设计学生答辩或质疑记录：   1. 你们小组的分工是怎么样的？   答：章华鼎主要负责代码编写、数据清洗、ppt制作；徐佳清主要负责Github管理及代码、报告编写；张湖川主要负责代码编写及代码调试，但每个人都参与了各个项目环节。   1. 在本次课设中是否使用了数据库？   答：我们没有使用数据库，我们直接将Excel文件导入，对数据进行清洗之后开始对数据进行分析。   1. 在课程设计过程中遇到了哪些困难？   答：首先是对github的使用不是很熟悉，其次是pyecharts的版本不统一，最后是对python的一些包不太了解，需要查阅相关资料才清楚它们的用途及使用方法。 | | | |
| **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、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 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 15日 | | | |