**数学与计算机学院**

**Python程序设计**

**课程设计说明书**

**题 目： 历年春晚数据分析**

**专 业： 数据科学与大数据技术**

**班 级： 1901**

**学 号： 1905140015**

**姓 名： 黎仕勇**

**指导老师： 徐军利**

**2020年　12月　31日**

**目录**

一、 课程设计概述--------------------------------------------------1

1.1 课程设计的目的--------------------------------------------1

1.2 课程设计的任务以及简介------------------------------------1

二、 需求分析-------------------------------------------------------4

2.1 系统功能分析-----------------------------------------------4

2.2 数据采集来源-----------------------------------------------4

2.3 解决问题所采用的算法策略----------------------------------

三、 第三方库、所采用的库方法介绍-----------------------------------8

3.1 requests库及方法-------------------------------------------8

3.1.1 requests.get()---------------------------------------

3.1.2 response的属性和方法--------------------------------

3.2 lxml.etree------------------------------------------------

3.2.1 etree.HTML()------------------------------------------

3.2.2 etree.tostring()--------------------------------------

3.2.3 bytes.decode()----------------------------------------

3.2.4（xpath解析对象）.xpath--------------------------------

3.3 pandas库及方法-------------------------------------------

3.3.1 pd.read\_html()方法-------------------------------------

3.3.2 pd.read\_csv()方法--------------------------------------

3.3.3 pd.to\_csv()方法----------------------------------------

四、 算法设计------------------------------------------------------19

4.1 爬取节目单模块--------------------------------------------19

4.2 爬取导演、主持和时间模块----------------------------------

4.3 节目类型对比模块-------------------------------------------

* 1. “钉子户”筛选模块------------------------------------------
  2. 港台明星上春晚次数对比模块----------------------------------
  3. 时间相同年份选择模块---------------------------------------
  4. 导演次数对比模块-----------------------------------------
  5. 主持次数对比模块-----------------------------------------

五、 课程设计总结--------------------------------------------------33

六、 附录----------------------------------------------------------33

七、 致谢----------------------------------------------------------33

1. **课程设计概述**

**1.1课程设计的目的**

《Python程序设计》课程设计是数据科学与大数据技术专业、信息管理与信息系统专业学生的集中实践性环节之一，是学习完《Python程序设计》课程后进行的一次全面的综合练习。**其目的是使学生熟练掌握**Python**语言的基本知识和技能**，训练学生综合运用已学课程的基本知识，独立进行Python语言应用程序设计,掌握Python语言程序设计、调试的方法和技巧**，从而**提高动手编程解决实际问题的能力。

**1.2课程设计的任务以及简介**

上网搜索历年春晚节目数据，通过Python数据分析，用各种图表回答下列问题：

1. 谁导演的次数最多？
2. 谁主持次数最多？
3. 哪两年的除夕是同一天？
4. 谁上春晚次数最多，堪称‘钉子户’
5. 港台明星上春晚次数对比
6. 歌曲、小品、相声类节目数量对比
7. **课程设计概述**

**2.1系统功能分析**

**2.2数据采集来源**

维基百科全书。网站：<https://bk.tw.lvfukeji.com/zh-cn/>

**2.3解决问题的算法策略**

爬虫的步骤流程图：

基于目标网页特征的爬虫所抓取、存储并索引的对象一般为网站或网页。根据种子样本获取方式可分为：①预先给定的初始抓取种子样本；②预先给定的网页分类目录和与分类目录对应的种子样本。

对爬取数据的清洗：①将带空值的元组去除；②去除异常数据；

1. **第三方库、所采用的库方法介绍**

**3.1 requests库及方法**

**3.1.1requests库的简介：**

**requests是**网络爬虫三大库（requests、lxml、BeautifulSoup）之一。该课程设计涉及到的方法有：

1. Requests.get()方法：向服务器发出http请求；
2. Response的各种属性和方法：

属性：response.status\_code、response.url、response.encoding

**3.2 lxml中的etree**

**3.2.1 etree.HTML()**

构造了一个XPath解析对象并对HTML文本进行自动修正。用于解析html内容。返回值为Xpath解析对象。

3.2.2 **etree.tostring:**

输出为修正后的结果，类型为bytes。

**3.2.3 Bytes.decode()**

该方法以指定的编码格式解码 bytes 对象。默认编码为 'utf-8'。

**3.2.4（xpath解析对象）.xpath**

解析html文本获取数据。

**3.3 pandas库及方法**

Pandas是一个专门用于数据分析的开源Python库。它能够以最简单的方式提供数据处理、数据抽取和数据操作所需的全部工具。

**3.3.1 pd.read\_html()**

读取网页Table表格数据。

**3.3.2 pd.read\_csv()**

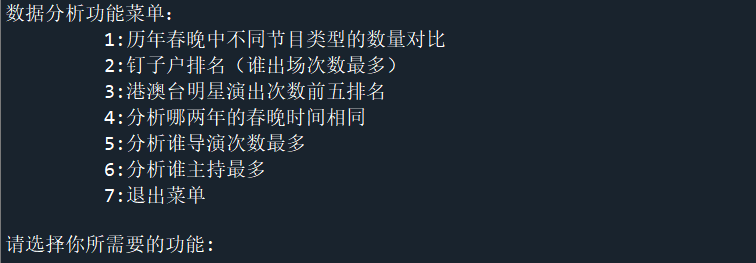
作用：将csv文件读入并转化为数据框形式。

**3.3.3 df.to\_csv()**

to\_csv()是DataFrame类的方法，read\_csv()是pandas的方法。

1. **算法设计**

**最终功能菜单界面：**

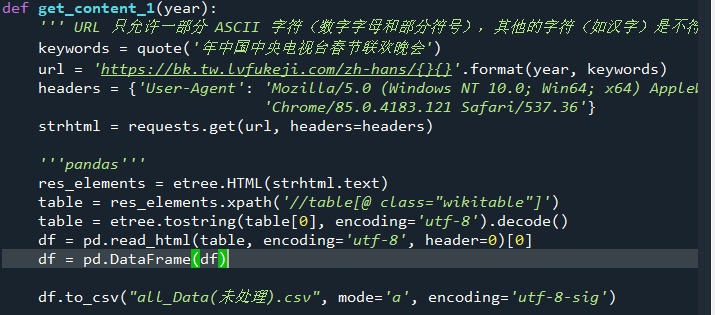


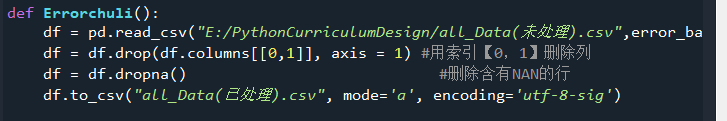
**4.1 爬取节目单模块**

**4.1.1模块功能简介：**

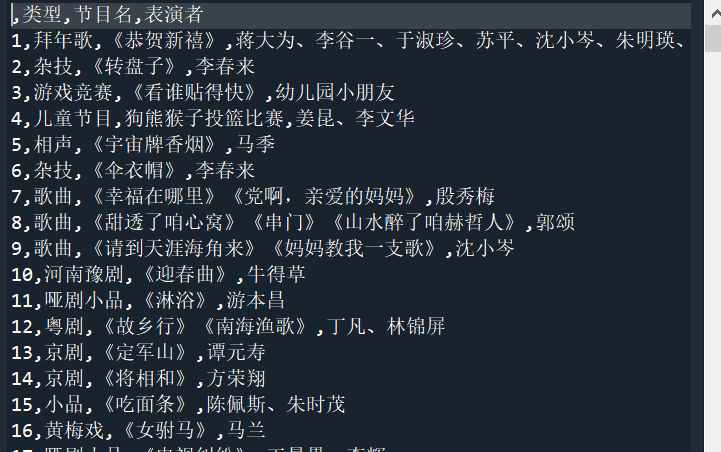
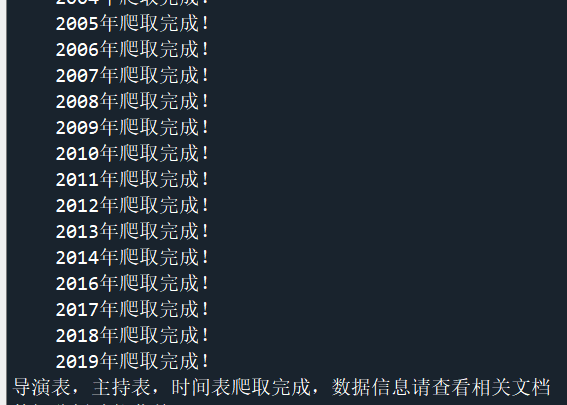
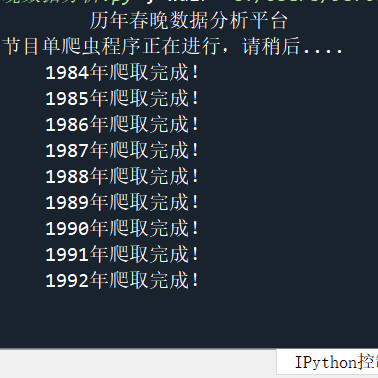
爬取1986年到2019年春晚节目表，该表有三列（类型、节目、表演者）

**4.1.2 模块代码（爬取信息+清洗信息）：**





**4.1.3 模块运行结果**

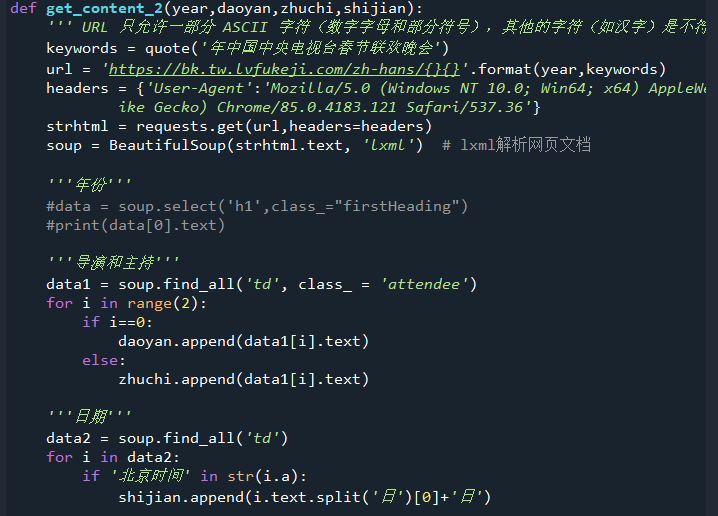


**4.2 爬取导演、主持和时间模块**

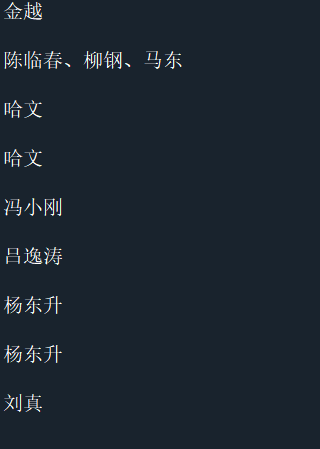
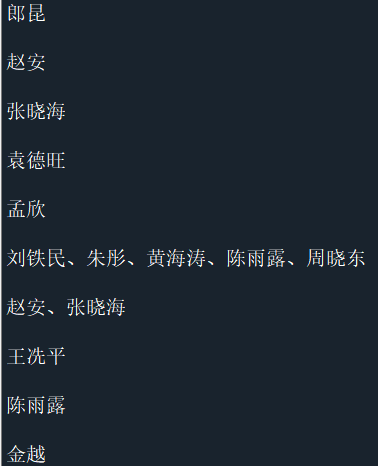
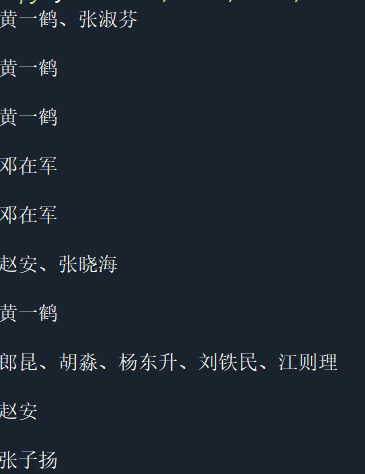
**4.2.1 模块功能介绍**

爬取1984年到2019年各个年份春晚的导演、主持人和时间

**4.2.2 模块代码**



**4.2.3 模块运行结果**

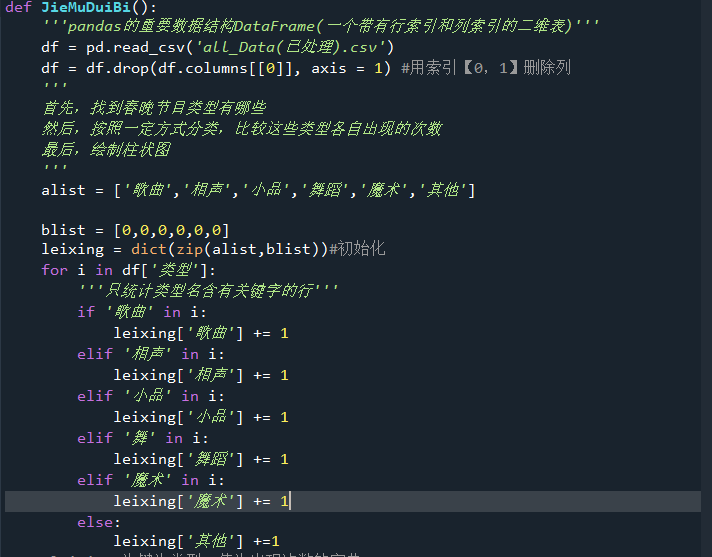


**4.3 节目类型对比模块**

**4.3.1 模块功能介绍：**

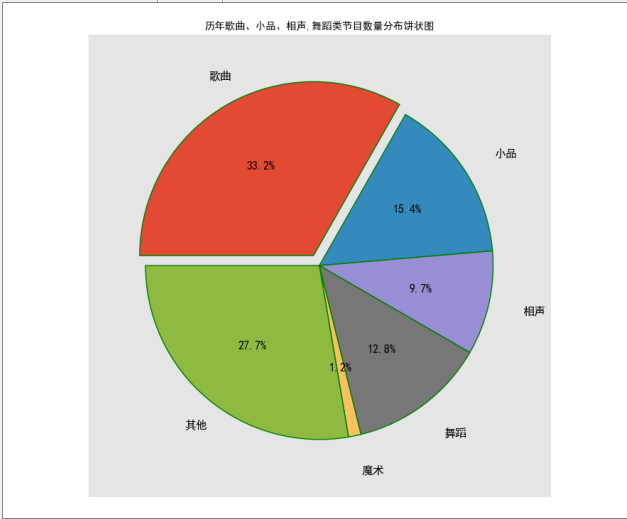
对比热门节目在春晚上的出场次数

**4.3.2 模块代码**





**4.3.3 模块运行结果：**



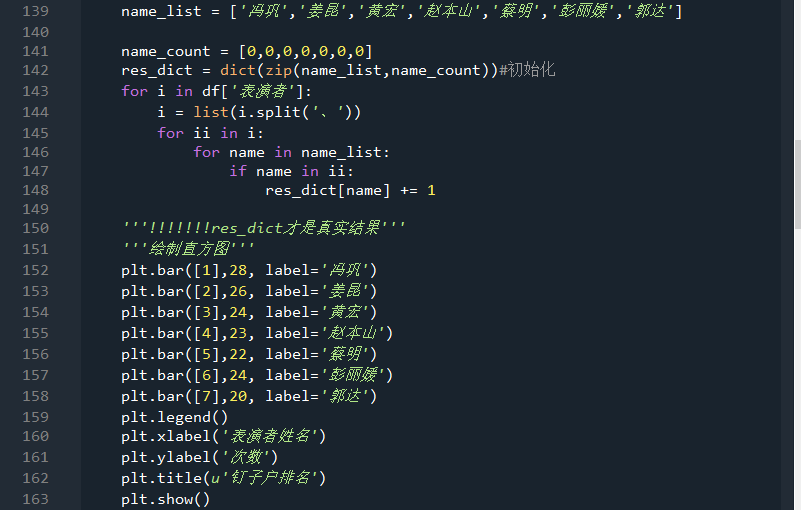
**4.4 “钉子户”筛选模块**

**4.4.1 模块功能介绍**

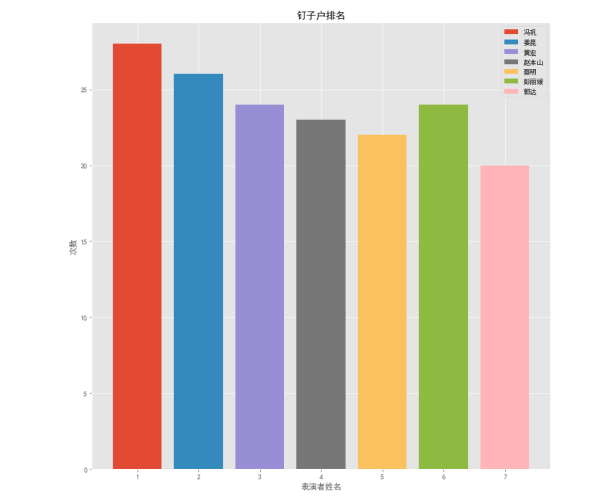
筛选出1984年到2019年上春晚次数最多的几个演员

**4.4.2 模块代码**





**4.4.3 模块运行结果**



**4.5 港台明星上春晚次数对比模块**

**4.5.1 模块功能介绍**

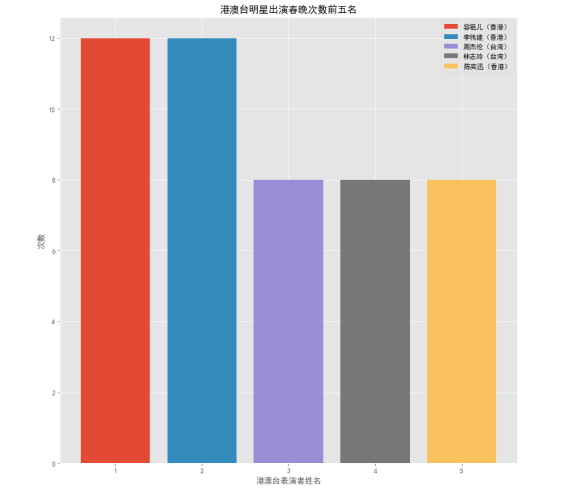
筛选出1984年到2019年上春晚次数最多且属于港、澳、台地区的几个演员

**4.5.2 模块代码**





**4.5.3 模块运行结果：**

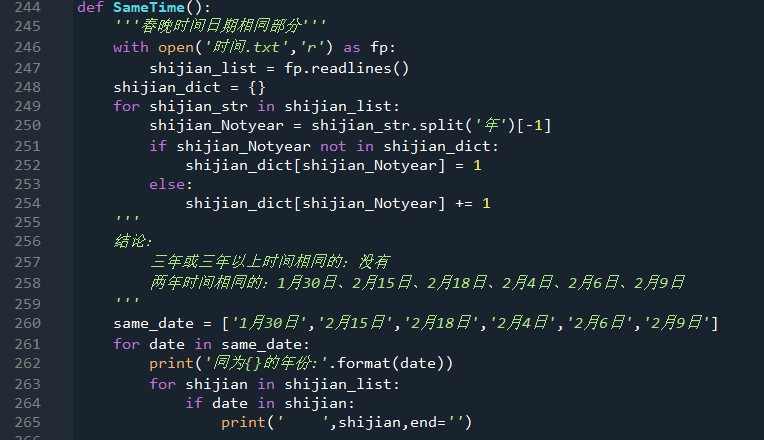


**4.6 时间相同年份选择模块**

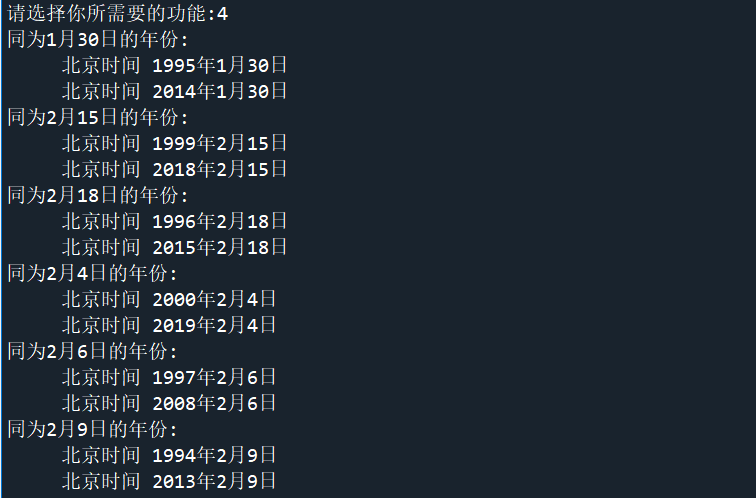
**4.6.1 模块功能介绍**

找出哪两年的除夕日期相同

**4.6.2 模块代码**



**4.6.3 模块运行结果**

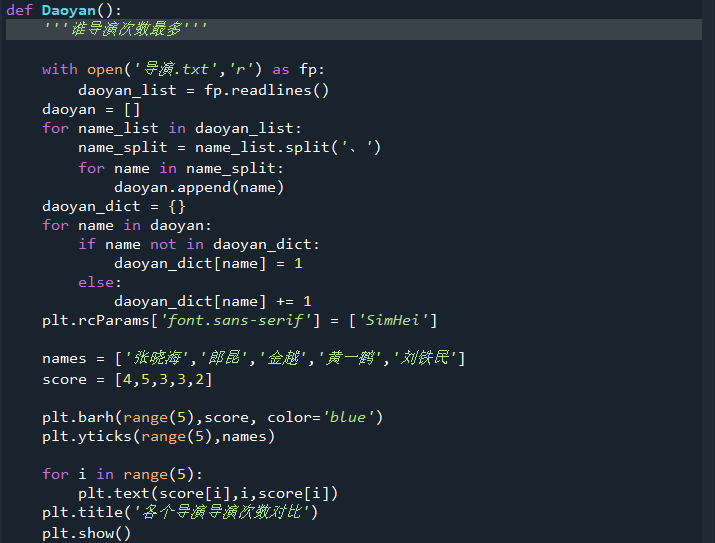


**4.7 导演次数对比模块**

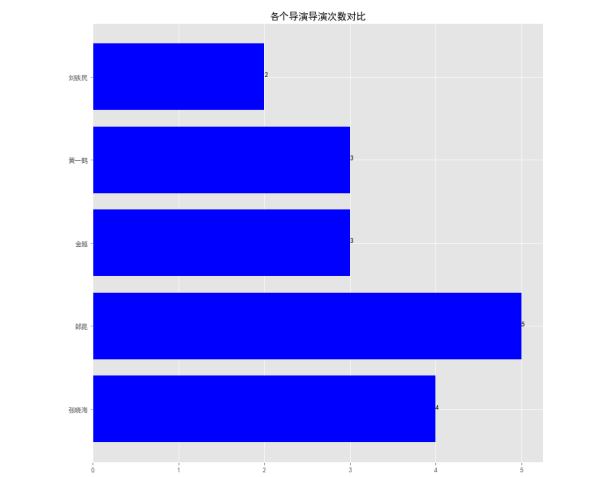
**4.7.1 模块功能介绍：**

对比导演春晚次数最多得几个导演，并筛选出来。

**4.7.2 模块代码：**



**4.7.3 模块运行结果：**

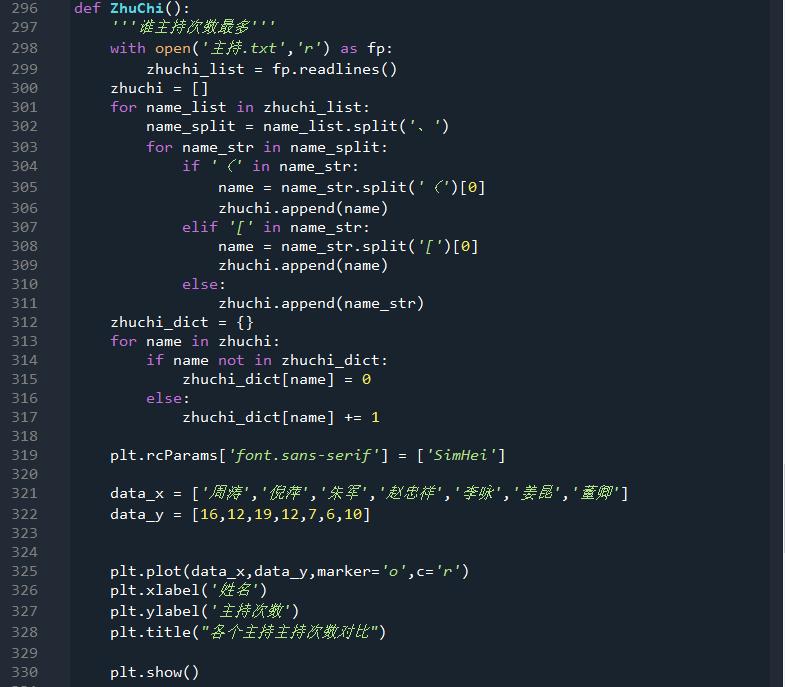


**4.8 主持次数对比模块**

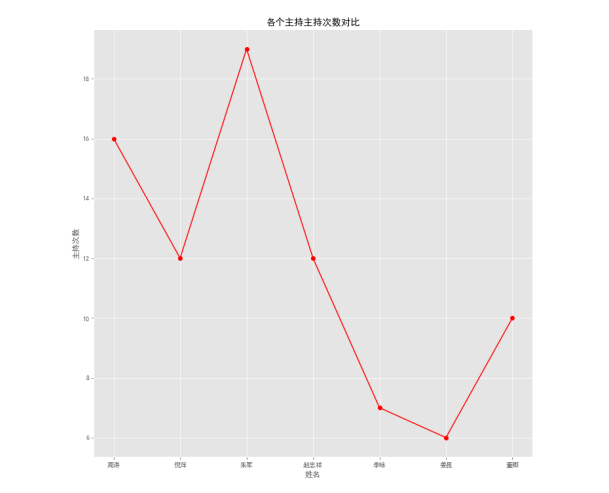
**4.8.1 模块功能介绍：**

对比主持春晚次数最多得几个主持人，并筛选出来。

**4.8.2 模块代码：**



**4.8.3 模块运行结果：**



1. **课程设计总结**

经过这次课程设计，使我学到了很多东西，包括书本上得理论知识以及书本上学不到得知识。具体得收获与体会如下：

一、经过这次课程设计，培养了我耐心、仔细、谨慎得工作态度。这次课程设计得资料对多的是学习python的一些扩展库。在学习库的过程中，会有很对的对象，要记住并且理解对象的属性和方法。

二、经过这次课程设计，使我加深了对所学理论知识的理解与巩固，并能将课本上的纯理论应用到实践当中，进一步加深了对知识的认识。同时，也有助于对其他知识的理解。此刻，我不但对requests库和pandas库有了更加深入的理解，并且能熟练的实现一些简单的网络爬虫技术，也让我有了新的python学习方向，找到了自己感兴趣的地方。