# 项目中使用的python

我没有单独使用过python完成一个大的项目，但是我最近一个项目用到了python，我准备将这个项目中使用的python描述一下，首先简单介绍下这个项目。

**一、作品简介**

**作品主题&解决问题：**2019年末，新型冠状病毒疫情突然出现，转眼间庚子鼠年来临了。本该是举国上下阖家欢乐的时刻，然而新型冠状病毒导致的肺炎感染牵动着全国人民的心，一场没有硝烟的战斗打响了，疫情之下，人间有爱。抗击疫情的过程中每天都产生庞大的数据，合理利用大数据可视分析技术对疫情防控具有指导意义。

疫情数据呈现出非常明显的地理和时序特征，属于典型的时空数据，利用数据可视化与可视分析方法，展示疫情的分布、多维信息和走势的同时，充分关联多源数据（比如：新闻数据、社交媒体信息等），分析疫情传播模式、比较各地传播差异、检测异常传播事件，分析社交媒体话题与情感的动态演变、对社会舆情进行态势感知，能够为目前和今后的疫情防控工作提供帮助和指导。

**目标用户：**全国各级防疫部门的工作人员可对比分析不同地区疫情传播模式，及时检测异常传播事件；有关部门可使用该系统进行社会舆情监测，探索疫情期间的议论话题走势与民众心理的变化，为正确引领舆论提供辅助性意见；普通民众可及时从时间、空间角度了解疫情发展最新态势；该系统也面向其他可视分析爱好者，便于大家共同学习。

**应用价值：**我们针对2020年1月23日至2020年5月18日共计117天，全国共计31个省级行政区，全球共计188个国家的新冠疫情每日新增病例数据、微博热搜数据、各大主流媒体新闻数据等，从时间、空间两个角度进行新冠疫情的态势分析与疫情期间社会舆情监测，设计并开发了一套可视分析系统：COVID-19：时空态势分析与舆情监测，旨在对疫情传播模式与大众舆情进行可视分析，为从事疫情相关工作的人员目前和今后的疫情防控工作提供帮助和指导。

|  |  |
| --- | --- |
| **作品主题** | 利用数据分析技术研究国内外疫情态势以及疫情期间舆情监测。 |
| **解决问题** | 分析疫情传播模式、检测异常传播事件，分析社交媒体话题与情感的动态等。 |
| **目标用户** | 全国各级防疫部门的工作人员及普通民众。 |
| **应用价值** | 为从事疫情相关工作的人员目前和今后的疫情防控工作提供帮助和指导。 |

**二、项目中使用的数据**

**数据来源：**本系统所使用的原始数据是中国各省市每日疫情态势数据（包括累计确诊病例，累计死亡病例，每日新增病例等），全球各国家每日新增病例数据，每日新闻数据，每日微博热搜数据。所有数据的时间段均是2020年1月23日至2020年5月18日，其中微博热搜数据是使用python爬虫从新浪微博爬取的，国内疫情数据来源于丁香园官网，每日新闻来源于人民日报，全球疫情数据来源于美国霍普金斯大学。

**数据格式：**所有的数据均以csv文件存储。

**数据严谨性：**使用的所有数据均是现实生活中的真实数据，可以保证数据是严谨，准确，合理的。

**数据清洗：**数据清洗采用csv文件存储+python处理的方式，具体步骤如下:

1. 缺省值清洗：首先确定缺省值的范围，将数据备份后删除不需要的字段，最后在不影响数据严谨性的情况下对某些缺省值进行填充。
2. 格式内容清洗：将数据中的时间，日期，数值等字段清洗成我们想要的格式。
3. 逻辑错误清洗：对数据进行去重以及去除逻辑不合理的字段。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **原始数据** | **国内疫情数据** | **全球疫情数据** | **每日新闻数据** | **每日微博热搜** |
| **数据来源** | 丁香园官网 | 霍普金斯大学 | 人民日报 | 新浪微博 |
| **数据格式** | csv文件 | csv文件 | csv文件 | csv文件 |
| **是否进行清洗** | 是 | 是 | 是 | 是 |

**三、项目中使用的python**

**算法模型**

**（1）MDS（Mulitiple Dimensional Scaling）降维：**MDS通过利用对点（数据）做平移，旋转，翻转等操作，点的距离是不变的这一特性来对原始数据进行操作来使高维空间（原始空间）中样本之间的距离在低维空间中得以保持，从而达到降维的目的。

**（2）KMeans聚类:** K-Means算法是以距离作为相似度的评价指标，用样本点到类别中心的误差平方和作为聚类好坏的评价指标，通过迭代的方法使总体分类的误差平方和函数达到最小的聚类方法。

**绘制散点图**

我是用jupyter notebook在浏览器上使用MDS和KMeans，先将数据使用python中的pandas读入，然后对数据进行切片，处理成我们需要的格式，运行结果如图3.1所示：

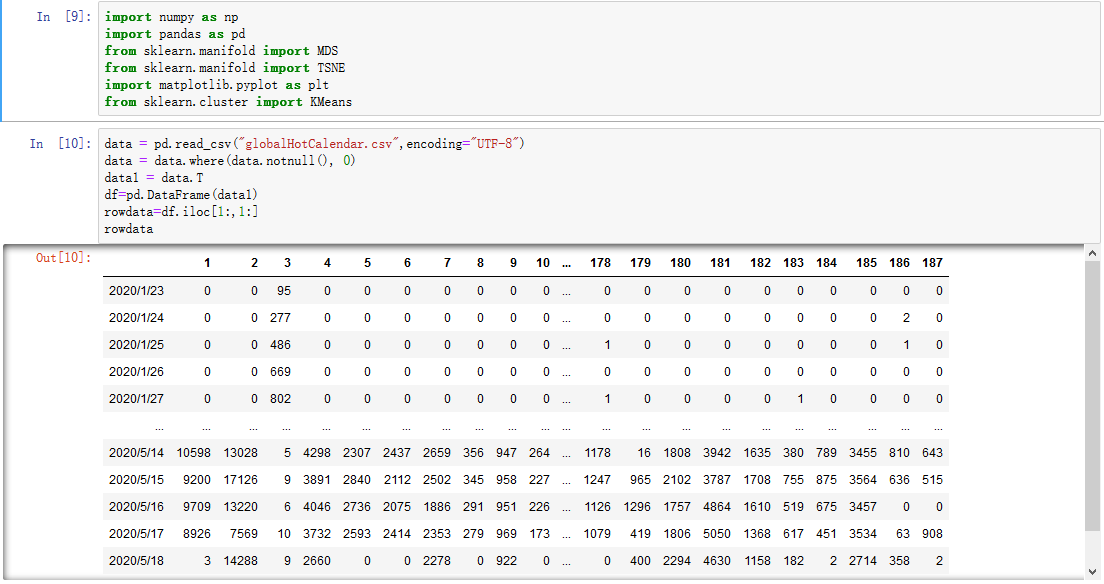


图3.1.

之后采用mds将数据降至一维，之后使用KMeans聚类，降维聚类的结果如图3.2所示，之后不断调整MDS降维的参数，观察KMeans聚类标签，直到获得比较整齐的标签（相同的标签都在一起），说明聚类结果比较好。最后调节类数，发现聚成五类效果最好。

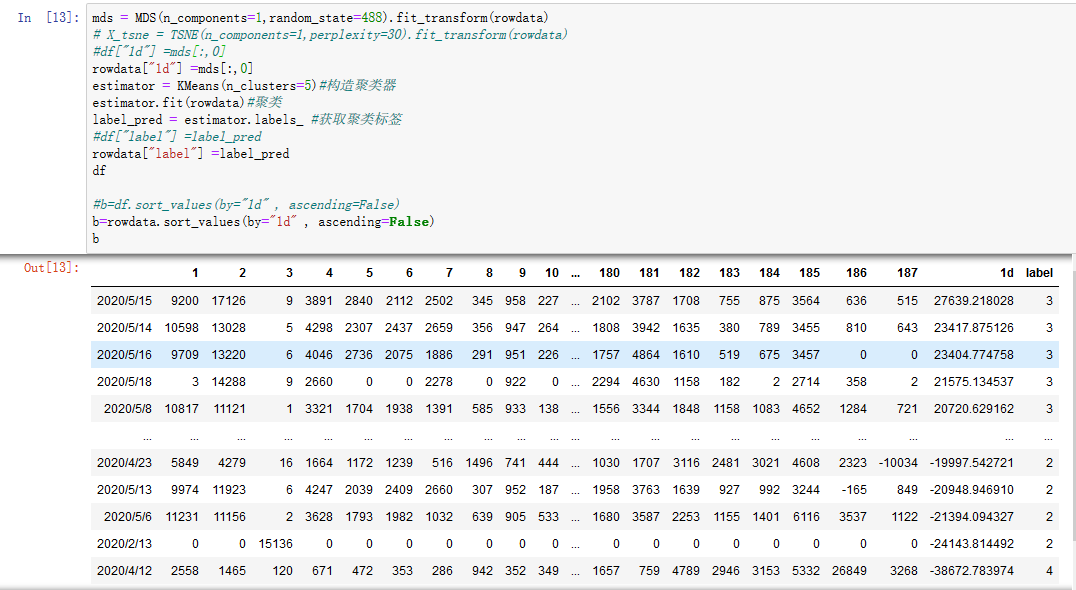


图3.2

将聚类结果的x和y坐标写入文件，方便绘制散点图（如图3.3所示）。

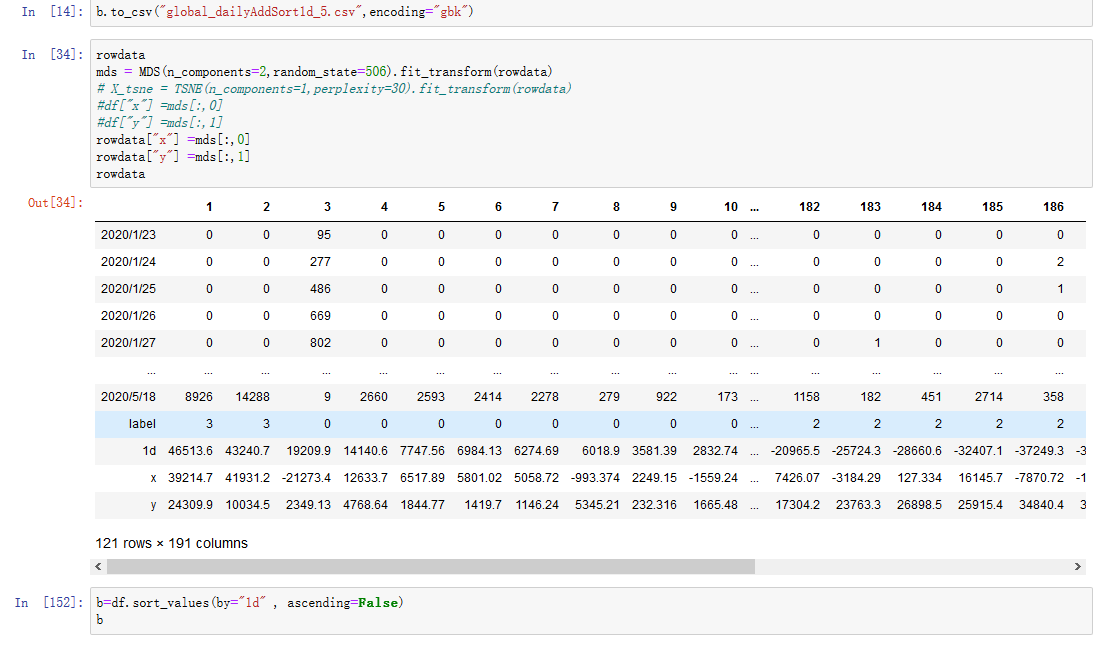


图3.3

散点图聚类结果，将全球188个国家按照1.23-5.18（共计117天）疫情增长情况分成四类，蓝色的类是最轻的，黄色的类是最严重的（如图3.4所示）。

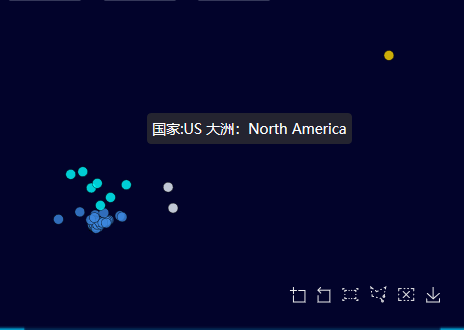


图3.4

使用python中的jieba分词将微博热搜分成关键词加词频的模式，如图3.5所示。

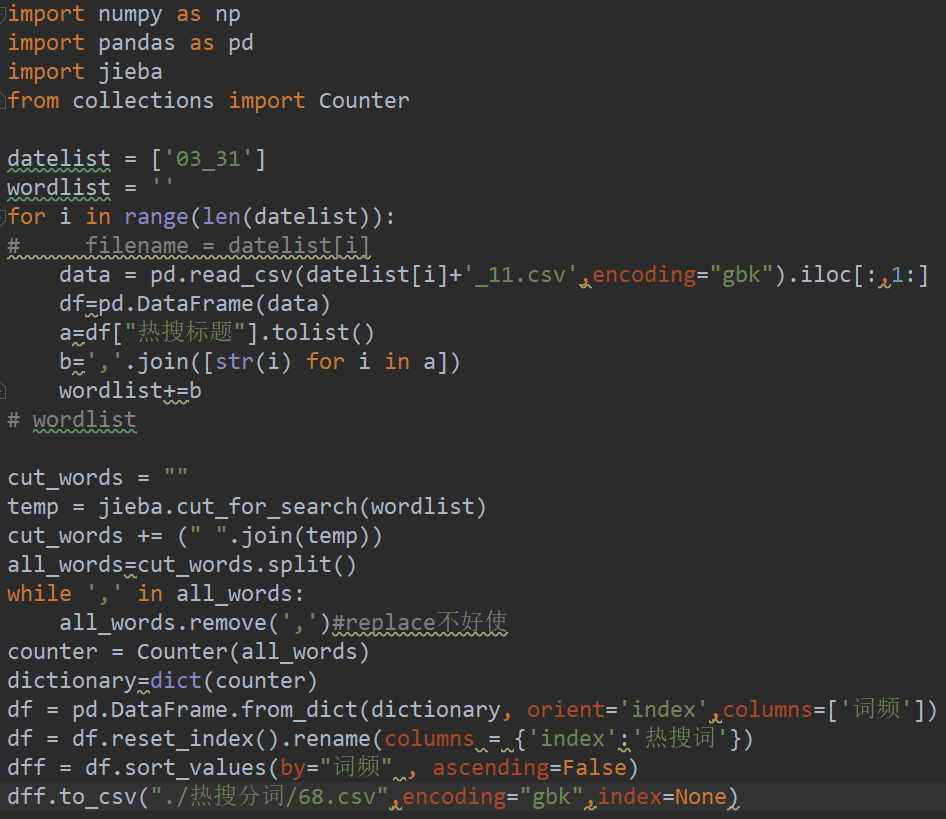


图3.5

**知乎数据爬取**

为了与微博数据做对比，我们还使用python爬取了知乎数据，代码如图3.6所示

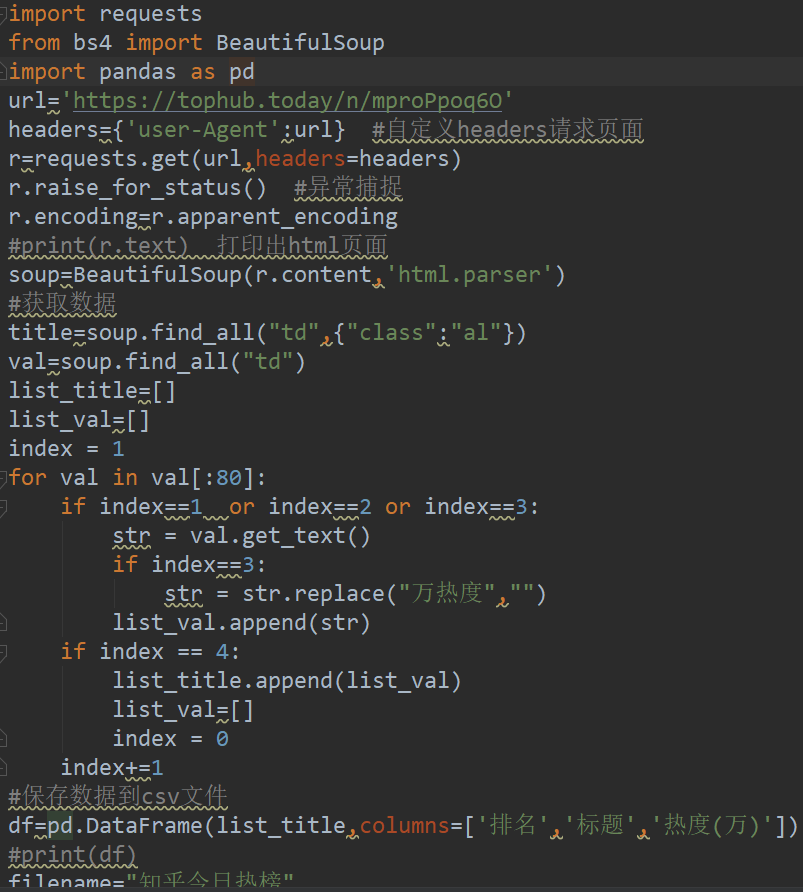


图3.6

**每日微博热搜数据分类**

将每日的微博热搜分成四类，分别是疫情利向，疫情弊向，娱乐和其他。我们采用基于情感字典的情感分析方法，将微博热搜文本使用jieba分词划分出情感词、否定词以及程度副词，而后使用之前设置好的情感字典根据情感权值将热搜语句分为正向和负向，从而进行分类。



图3.7

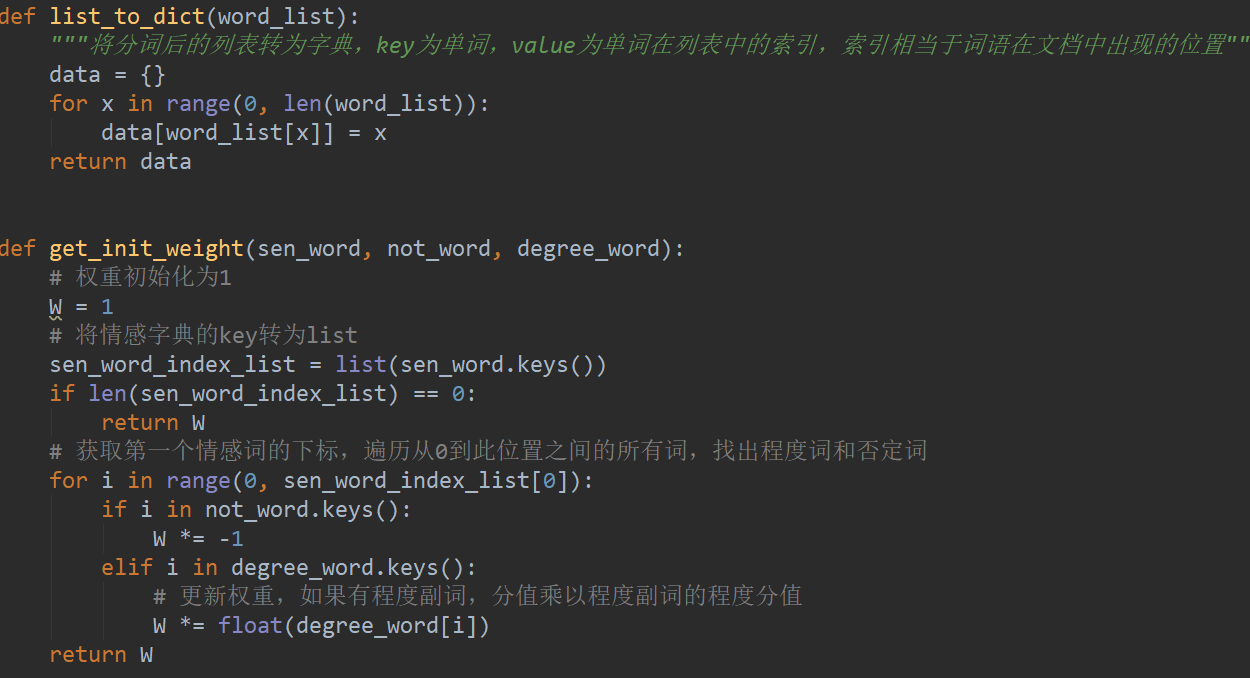


图3.8



图3.9