项目需求

- 爬取"新疆棉事件"单条微博转发时间、转发用户地区、用户的性别、年龄，输出`csv`文件存储数据(10w条数据)以及保存到`MongoDB`数据库中
- 对数据进行可视化(分布散点图、各时段转发微博数量面积图、用户的性别占比图、用户的年龄段占比饼图(诸如00后，90后，80后.....))
- 研究转发微博数据时间维度、空间维度上的分布情况以及用户特征

开发环境

- python3.7.0下载安装
- Pycharm2019专业版下载安装
- MongoDB数据库管理软件下载安装

需要安装的模块

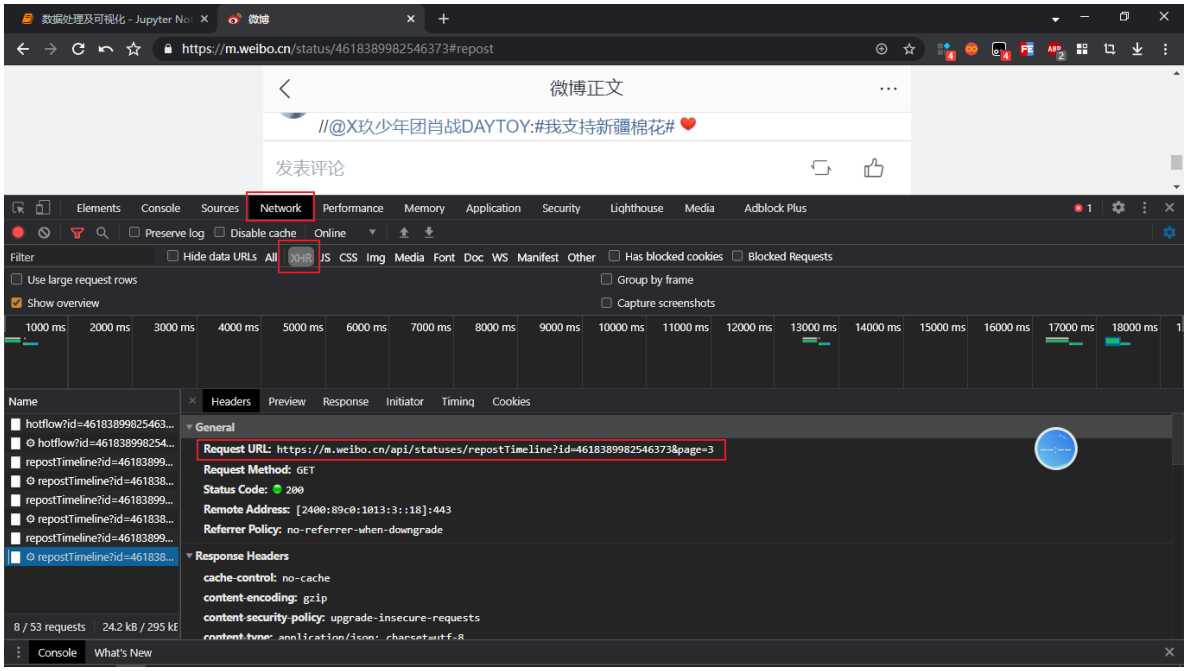
友情提示：安装之前先配置pip为国内镜像源，以上问题请自行解决！

```
pip install requests=2.22.0
pip install pymongo==3.10.1
pip install fake-useragent==0.1.11
pip install pandas==0.25.1
pip install matplotlib==3.1.1
pip install pyecharts==1.8.1
```

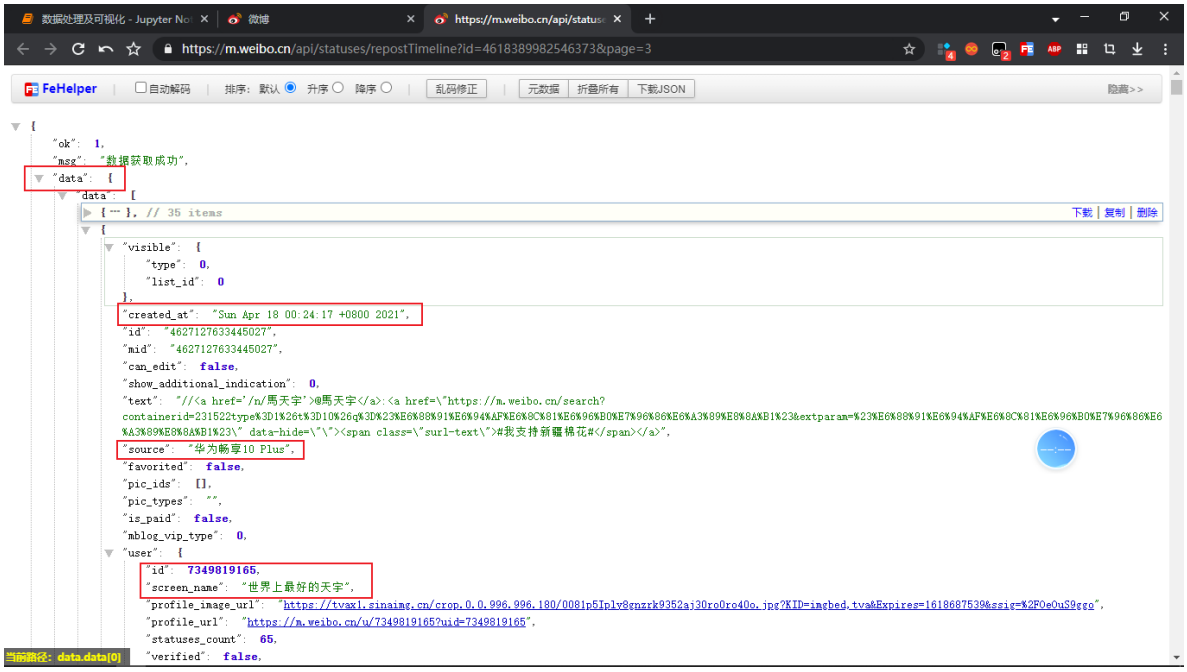
注意：安装`fake-useragent`时需要替换本地的`json`见https://cloud.tencent.com/developer/article/1636419

微博爬虫

选取话题进行爬取转发用户的信息！通过F12查看真实的网络`url`，获取到数据接口：

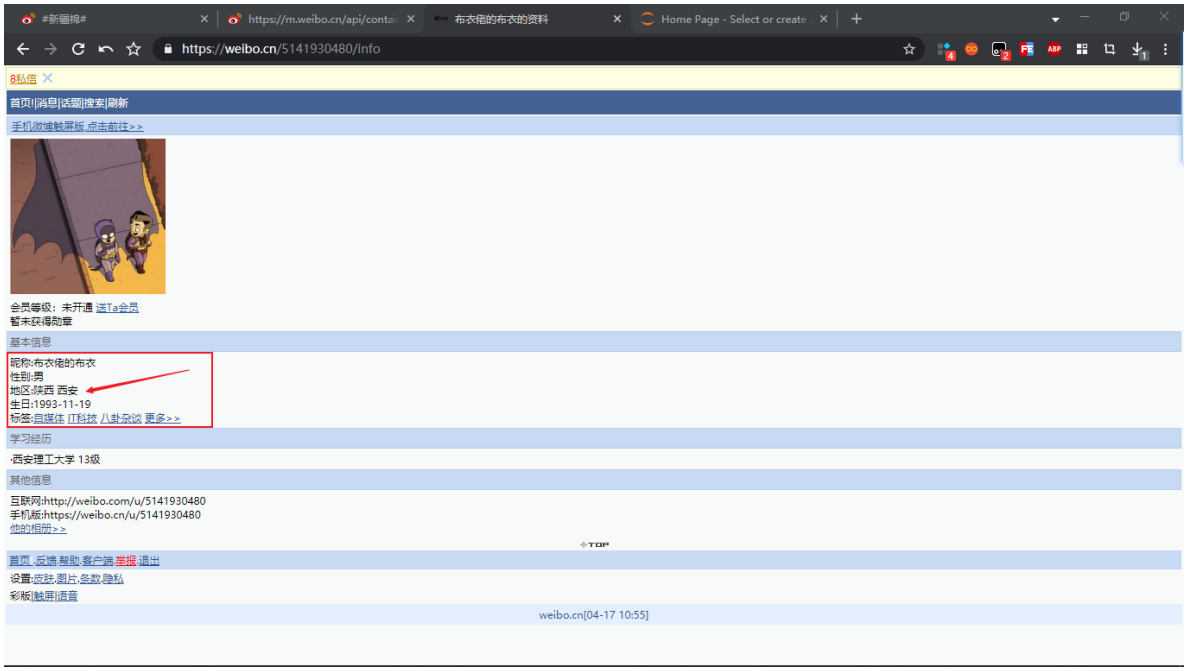


需要爬取10000页，每页10条数据，找到对应的链接构成规则：

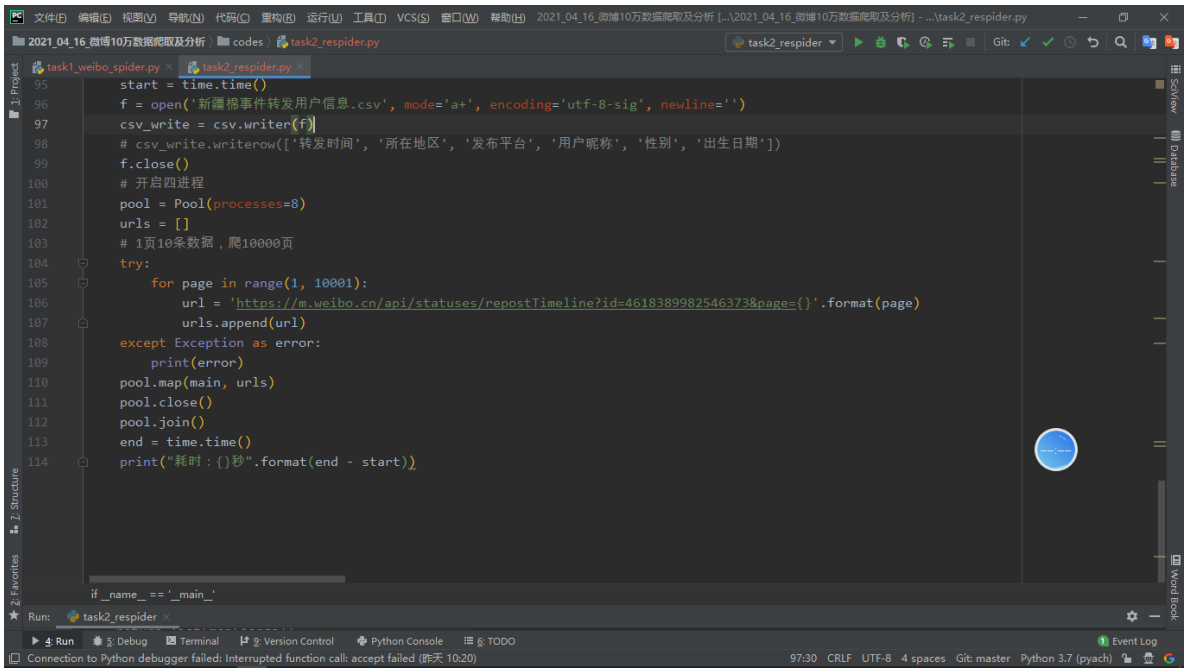


经过分析，发现只有后面的`page`带的数字不同！以及需要爬取的数据！

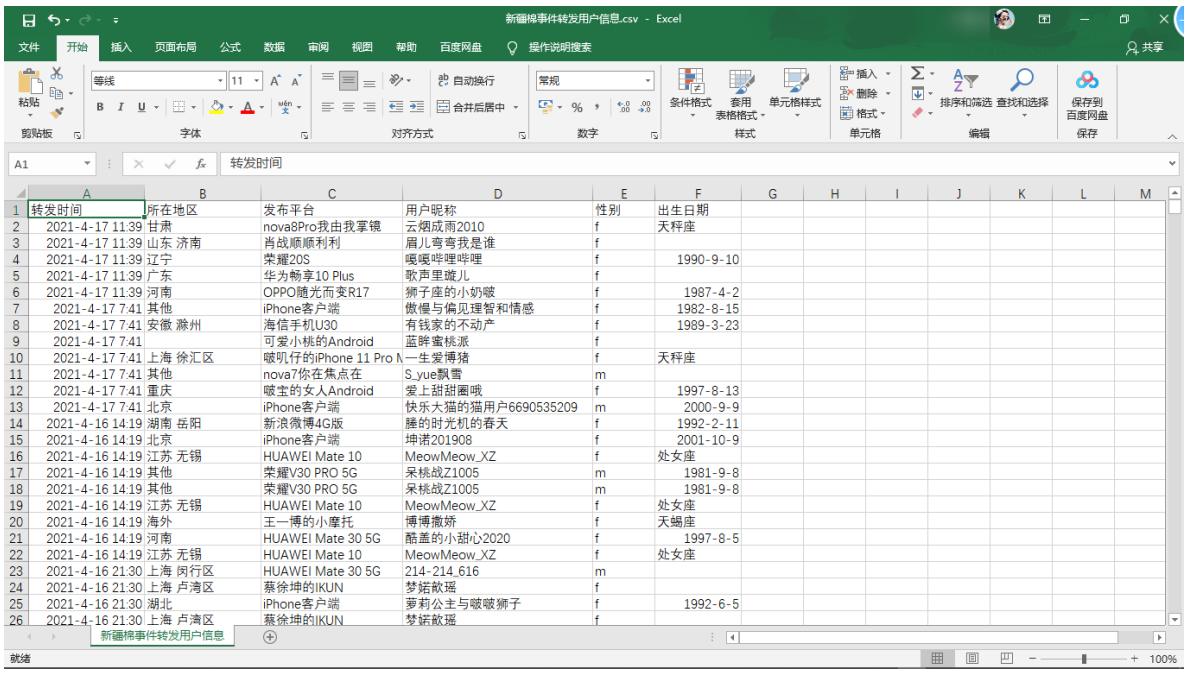
查看用户的生日和地区：



构建爬虫规则爬取响应的信息保存到csv文件和 MongoDB 数据库中：

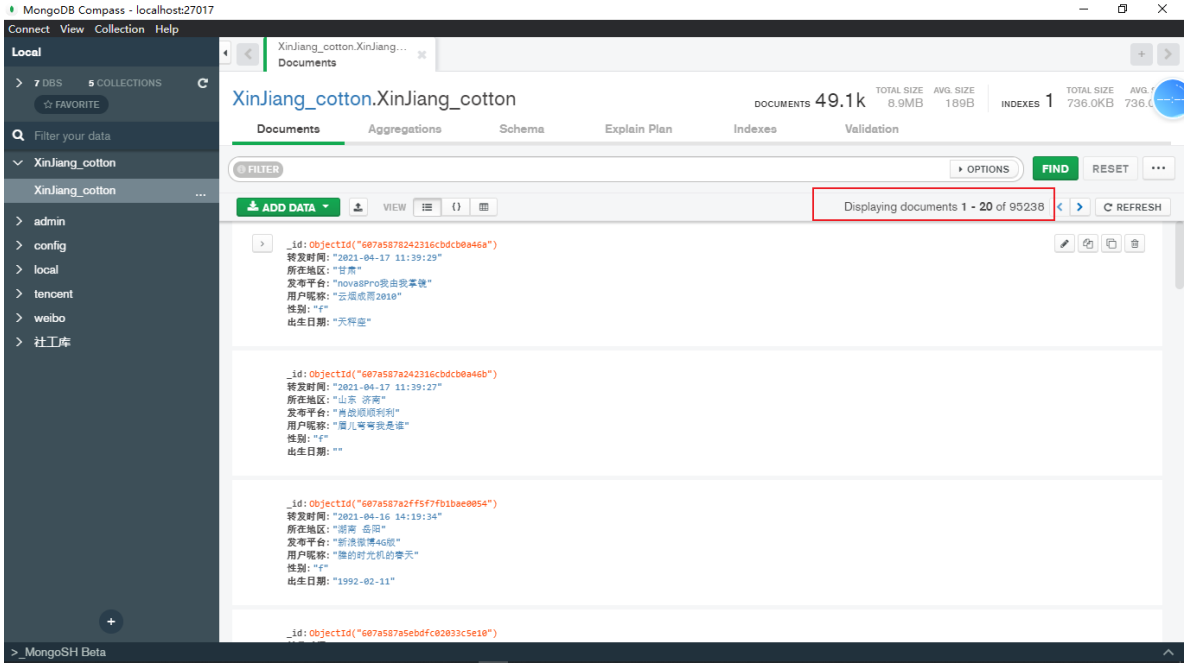


获取到的csv文件截图：



转发时间	所在地区	发布平台	用户昵称	性别	出生日期
2021-4-17 11:39	甘肃	nova8Pro我由我掌镜	云烟成雨2010	f	天秤座
2021-4-17 11:39	山东 济南	肖战顺顺利利	眉儿弯弯我是谁	f	
2021-4-17 11:39	辽宁	荣耀20S	嘿嘿哔哩哔哩	f	1990-9-10
2021-4-17 11:39	广东	华为畅享10 Plus	歌声里璇儿	f	
2021-4-17 11:39	河南	OPPO随光而变R17	狮子座的小奶皱	f	1987-4-2
2021-4-17 7:41	其他	iPhone客户端	傲慢与偏见理智和情感	f	1982-8-15
2021-4-17 7:41	安徽 滁州	海信手机U30	有钱家的不动产	f	1989-3-23
2021-4-17 7:41		可爱小桃的Android	蓝眸蜜桃派	f	
2021-4-17 7:41	上海 徐汇区	皱叽仔的iPhone 11 Pro M	一生爱博猪	f	天秤座
2021-4-17 7:41	其他	nova7你在焦点在	S_yue飘雪	m	
2021-4-17 7:41	重庆	皱宝的女人Android	爱上甜甜圈哦	f	1997-8-13
2021-4-17 7:41	北京	iPhone客户端	快乐大猫的猫用户6690535209	m	2000-9-9
2021-4-16 14:19	湖南 岳阳	新浪微博4G版	滕的时光机的春天	f	1992-2-11
2021-4-16 14:19	北京	iPhone客户端	坤诺201908	f	2001-10-9
2021-4-16 14:19	江苏 无锡	HUAWEI Mate 10	MeowMeow_XZ	f	处女座
2021-4-16 14:19	其他	荣耀V30 PRO 5G	呆桃战Z1005	m	1981-9-8
2021-4-16 14:19	其他	荣耀V30 PRO 5G	呆桃战Z1005	m	1981-9-8
2021-4-16 14:19	江苏 无锡	HUAWEI Mate 10	MeowMeow_XZ	f	处女座
2021-4-16 14:19	海外	王一博的小摩托	博博撒娇	f	天蝎座
2021-4-16 14:19	河南	HUAWEI Mate 30 5G	酷盖的小甜心2020	f	1997-8-5
2021-4-16 14:19	江苏 无锡	HUAWEI Mate 10	MeowMeow_XZ	f	处女座
2021-4-16 21:30	上海 闵行区	HUAWEI Mate 30 5G	214-214.616	m	
2021-4-16 21:30	上海 卢湾区	蔡徐坤的IKUN	梦诺散瑶	f	
2021-4-16 21:30	湖北	iPhone客户端	萝莉公主与皱皱狮子	f	1992-6-5
2021-4-16 21:30	上海 卢湾区	蔡徐坤的IKUN	梦诺散瑶	f	

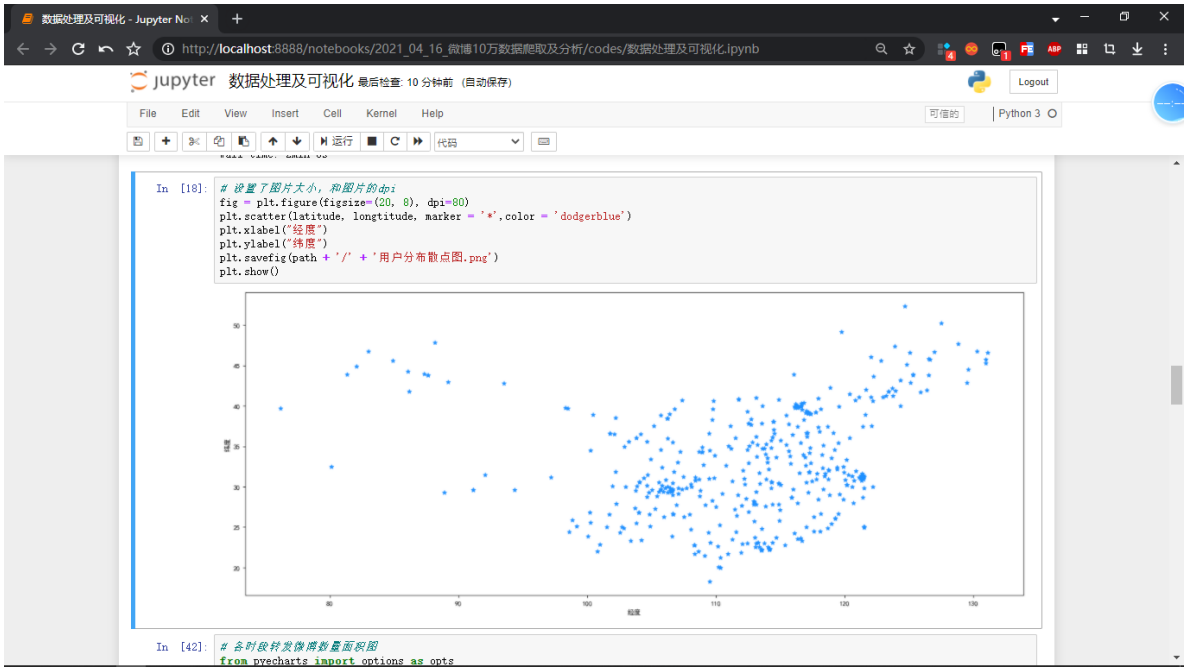
MongoDB数据库信息截图：

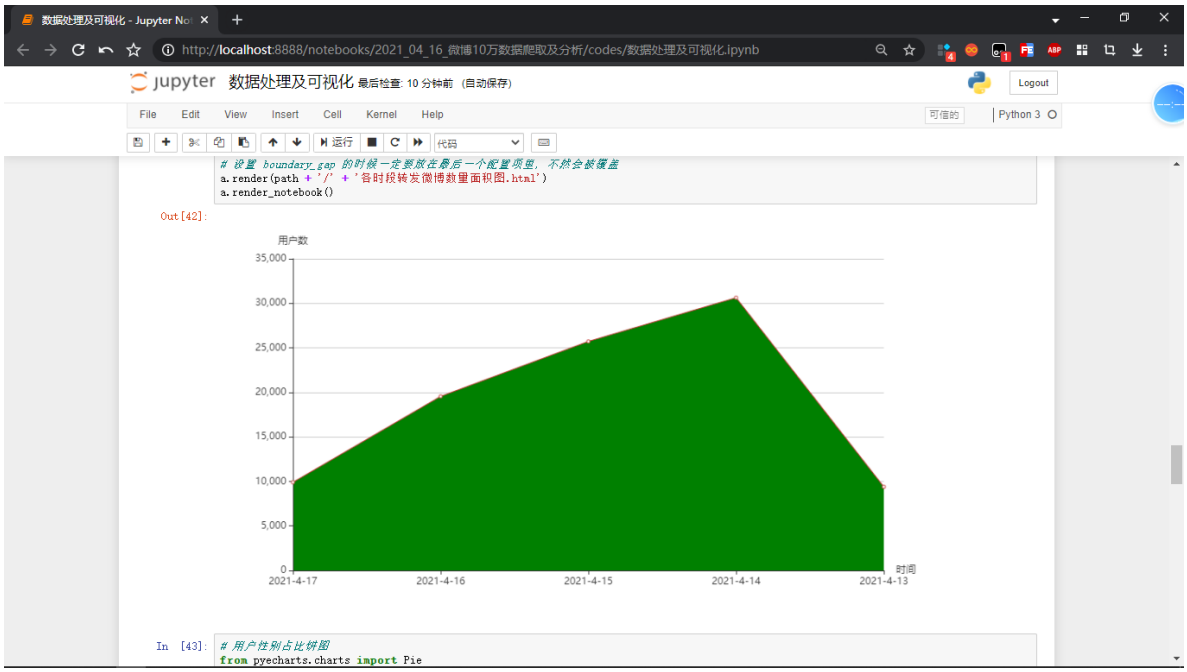


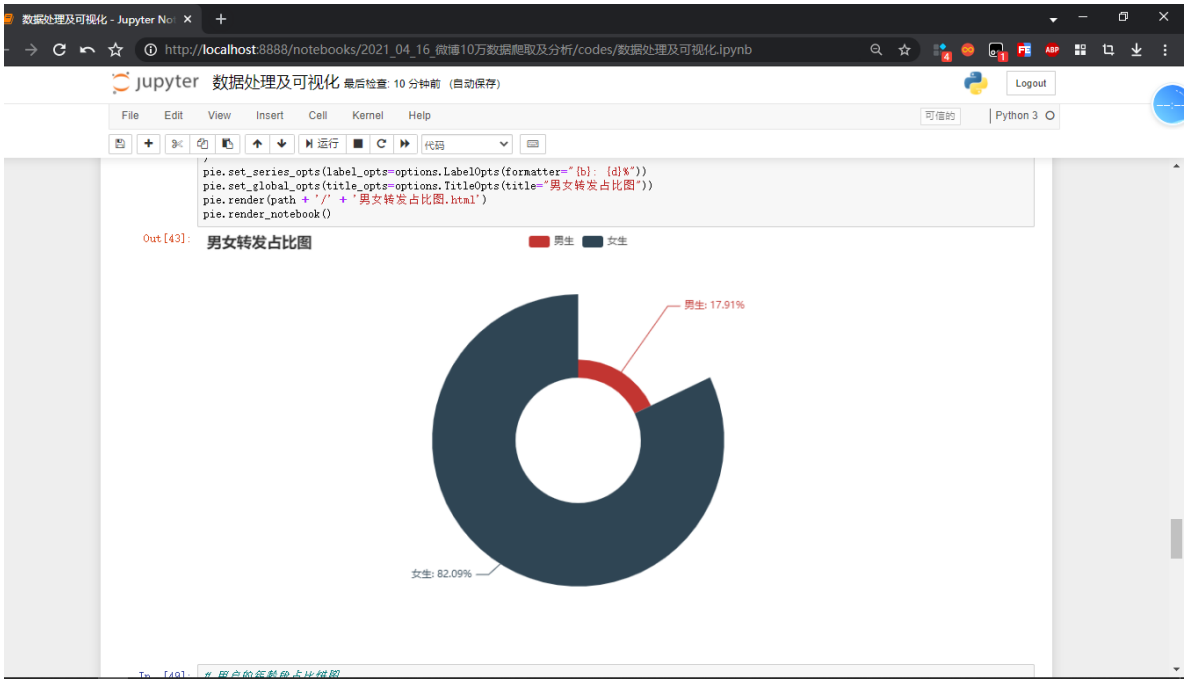
爬虫采用了多进程，但是执行时间依旧漫长，预计几小时！如果时间来的及可以爬！没有耐心则使用文件夹自带的文件！

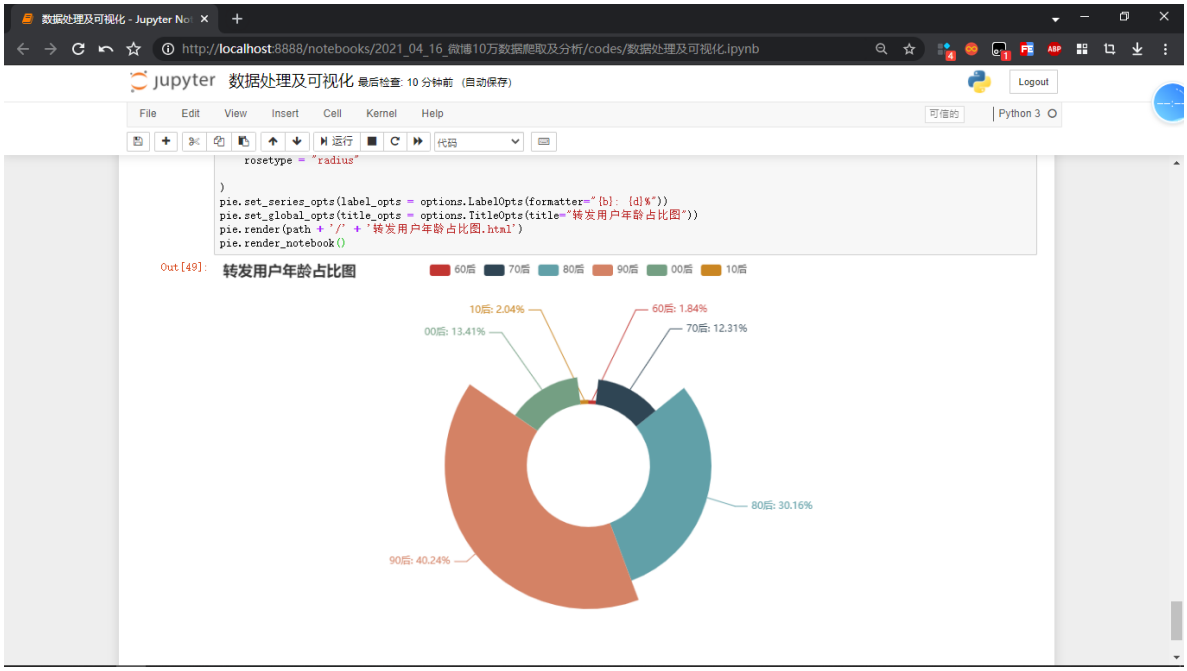
数据处理及可视化

得到的四幅图依次如下：


数据处理及可视化 - Jupyter No
http://localhost:8888/notebooks/2021_04_16_微博10万数据爬取及分析/codes/数据处理及可视化.ipynb
jupyter 数据处理及可视化 最后检查: 10 分钟前 (自动保存)
Logout
File Edit View Insert Cell Kernel Help
可信的
Python 3
运行
代码
In [18]:
# 设置了图片大小，和图片的dpi
fig = plt.figure(figsize=(20, 8), dpi=80)
plt.scatter(latitude, longtitude, marker = '*',color = 'dodgerblue')
plt.xlabel("经度")
plt.ylabel("纬度")
plt.savefig(path + '/' + '用户分布散点图.png')
plt.show()
In [42]:
# 各时段转发微博数量面积图
from pyecharts import options as opts


数据处理及可视化 - Jupyter No
http://localhost:8888/notebooks/2021_04_16_微博10万数据爬取及分析/codes/数据处理及可视化.ipynb
jupyter 数据处理及可视化 最后检查: 10 分钟前 (自动保存)
Logout
File Edit View Insert Cell Kernel Help
可信的
Python 3
运行
代码
# 设置 boundary_gap 的时候一定要放在最后一个配置项里，不然会被覆盖
a.render(path + '/' + '各时段转发微博数量面积图.html')
a.render_notebook()
Out[42]:
用户数
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
2021-4-17
2021-4-16
2021-4-15
2021-4-14
2021-4-13
时间
In [43]:
# 用户性别占比饼图
from pyecharts.charts import Pie





最后的文件目录结构：

```
D:.
|  fake_useragent_0.1.11.json
|  开发文档.md
|  新浪微博的数据爬取及传播的时空特性分析.doc
|
├─.idea
|  |  .gitignore
|  |  2021_04_16_微博10万数据爬取及分析.iml
|  |  misc.xml
|  |  modules.xml
|  |  vcs.xml
|  |  workspace.xml
|  |
|  └─inspectionProfiles
```

```
|          profiles_settings.xml
|
├─codes
|  |  task1_weibo_spider.py
|  |  task2_respider.py
|  |  数据处理及可视化.ipynb
|  |  新疆棉事件话题信息.csv
|  |  新疆棉事件转发用户信息.csv
|  |
|  ├─.ipynb_checkpoints
|  |      数据处理及可视化-checkpoint.ipynb
|  |
|  └─pictures
|          各时段转发微博数量面积图.html
|          用户分布散点图.png
|          男女转发占比图.html
|          转发用户年龄占比图.html
|
└─images
        csv文件截图.png
        年龄.png
        散点图.png
        爬虫0.png
        爬虫1.png
        爬虫2.png
        爬虫3.png
        爬虫截图.png
        男女.png
        面积图.png
```

其中`task1_weibo_spider.py`是第一次处理爬虫，无用!