

## 项目需求

使用 scrapy 框架编写爬虫程序抓取《安家》影评信息，爬取电视剧的短评(55593条)、评分、有用数量等数据，保存为 json 格式！



结合 Pandas、Numpy和Matplotlib，系统存储和处理爬取的大量数据，使用中文 Jieba 分词工具对爬取的短评信息文本处理，wordcloud 库处理数据关键词绘制词云图展示观众情感倾向和影片评分统计等信息！分别从评论时间、评分、评论内容进行数据可视化分析！

## 前期准备

- scrapy 框架的学习

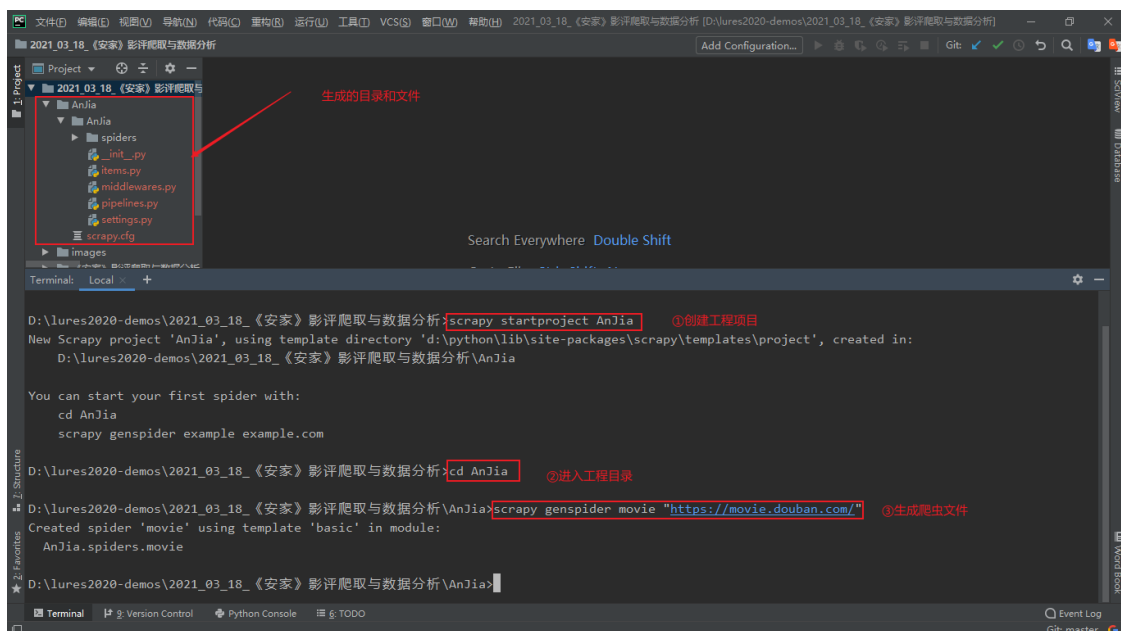
入门视频：<https://www.bilibili.com/video/BV1yf4y1B7S8>

## 项目步骤

### 1、创建工程目录

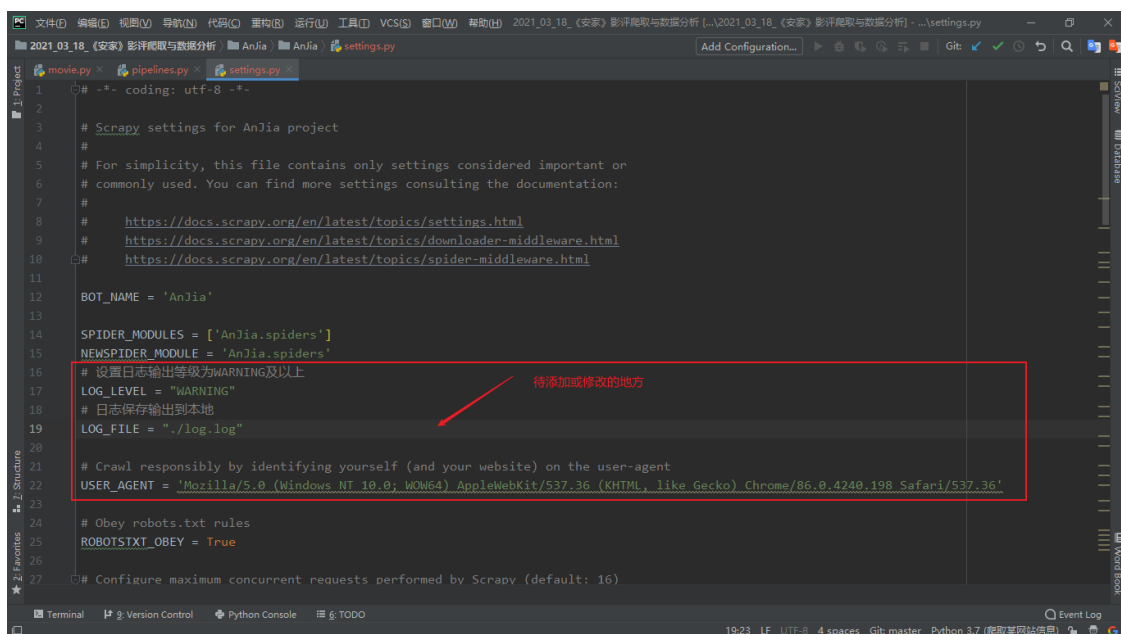
- 使用 scrapy startproject AnJia 创建工程目录 AnJia
- cd AnJia 进入工程目录
- scrapy genspider movie "https://movie.douban.com/" 生成爬虫项目，并命名为 movie.py

以上三步截图效果如下：



得到的相关文件及用途：

**settings.py**：爬虫相关操作的设置文件



以及68行左右取消注释代码：

```
ITEM_PIPELINES = {  
    'AnJia.pipelines.AnJiaPipeline': 300,  
}
```

**pipelines.py**：用于数据保存以及处理的模块，暂时无太大改动

```
1 # -*- coding: utf-8 -*-
2
3 # Define your item pipelines here
4 #
5 # Don't forget to add your pipeline to the ITEM_PIPELINES setting
6 # See: https://docs.scrapy.org/en/latest/topics/item-pipeline.html
7 from logging import getLogger
8 logger = getLogger(__name__)  # 日志输出库
9
10
11 class AnjiaPipeline(object):
12     def process_item(self, item, spider):
13         return item
14
```

movie.py：爬虫爬取的文件，包括一些常见的页面提取方法等

```
1 # -*- coding: utf-8 -*-
2 import scrapy
3 from logging import getLogger
4
5 # 生成迭代对象，进行日志处理
6 logger = getLogger(__name__)
7
8 class MovieSpider(scrapy.Spider):
9     name = 'movie'
10     allowed_domains = ['https://movie.douban.com/']  # 改为短评首页url
11     # 这是爬取的首页链接
12     start_urls = ['https://movie.douban.com/subject/30482003/comments?limit=20&status=P&sort=new_score']
13
14     # 下面是页面解析的函数，爬虫处理都在此函数中
15     def parse(self, response):
16         pass
17
```

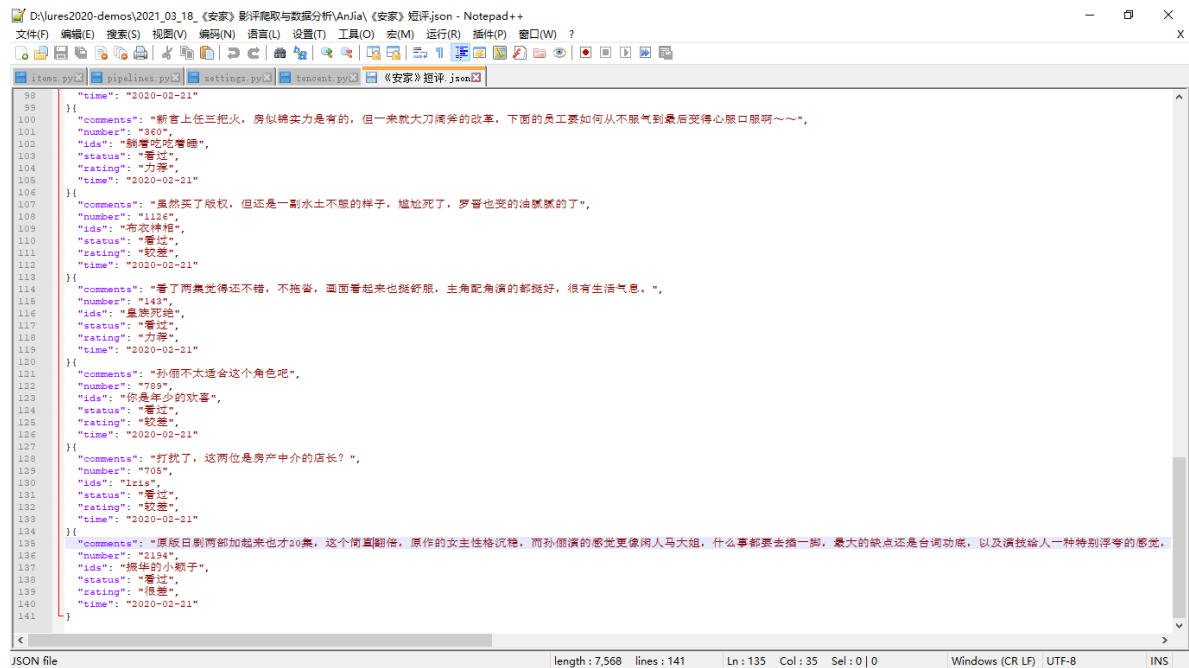
items.py：用于创建生成器以及生成响应的标签，用于爬虫标签信息的临时存取

```
1 # -*- coding: utf-8 -*-
2
3 # Define here the models for your scraped items
4 #
5 # See documentation in:
6 # https://docs.scrapy.org/en/latest/topics/items.html
7
8 import scrapy
9
10
11 class AnjiaItem(scrapy.Item):
12     # define the fields for your item here like:
13     # name = scrapy.Field()
14     pass
15
```

## 2、开始爬取首页

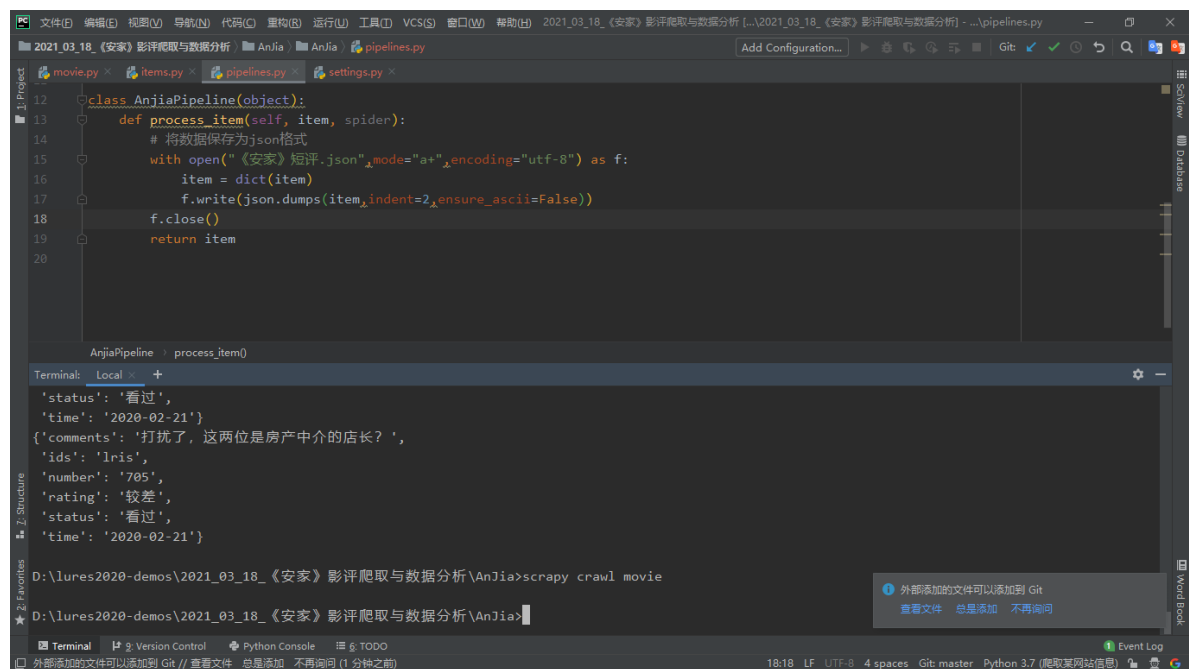
使用 scrapy crawl movie 开始运行爬虫程序

首先配置和数据截图：



```
99  "time": "2020-02-21"
99  }
100  "comments": "新官上任三把火，房似锦实力是有的，但一来就大刀阔斧的改革，下面的员工要如何从不服气到最后觉得心服口服啊～～",
101  "number": "380",
102  "ids": "躺着吃花着睡",
103  "status": "看过",
104  "rating": "力荐",
105  "time": "2020-02-21"
106  }
107  "comments": "虽然买了版权，但还是一副水土不服的样子，尴尬死了，罗晋也变的油腻腻的了，
108  "number": "1126",
109  "ids": "布衣神相",
110  "status": "看过",
111  "rating": "较差",
112  "time": "2020-02-21"
113  }
114  "comments": "看了两集觉得还不错，不拖沓，画面看起来也挺舒服，主角配角演的都挺好，很有生活气息。",
115  "number": "143",
116  "ids": "皇族死绝",
117  "status": "看过",
118  "rating": "力荐",
119  "time": "2020-02-21"
120  }
121  "comments": "孙俪不太适合这个角色吧",
122  "number": "709",
123  "ids": "你是年少的欢喜",
124  "status": "看过",
125  "rating": "较差",
126  "time": "2020-02-21"
127  }
128  "comments": "打扰了，这两位是房产中介的店长？",
129  "number": "705",
130  "ids": "Iris",
131  "status": "看过",
132  "rating": "较差",
133  "time": "2020-02-21"
134  }
135  "comments": "原版日剧两部加起来才20集，这个简直翻倍，原作的女主性格沉稳，而孙俪演的感觉更像闲人马大姐，什么事都要去插一脚，最大的缺点还是台词功底，以及演技给人一种特别浮夸的感觉。",
136  "number": "2194",
137  "ids": "德华的小儿子",
138  "status": "看过",
139  "rating": "很差",
140  "time": "2020-02-21"
141  }
```

终端首页截图：



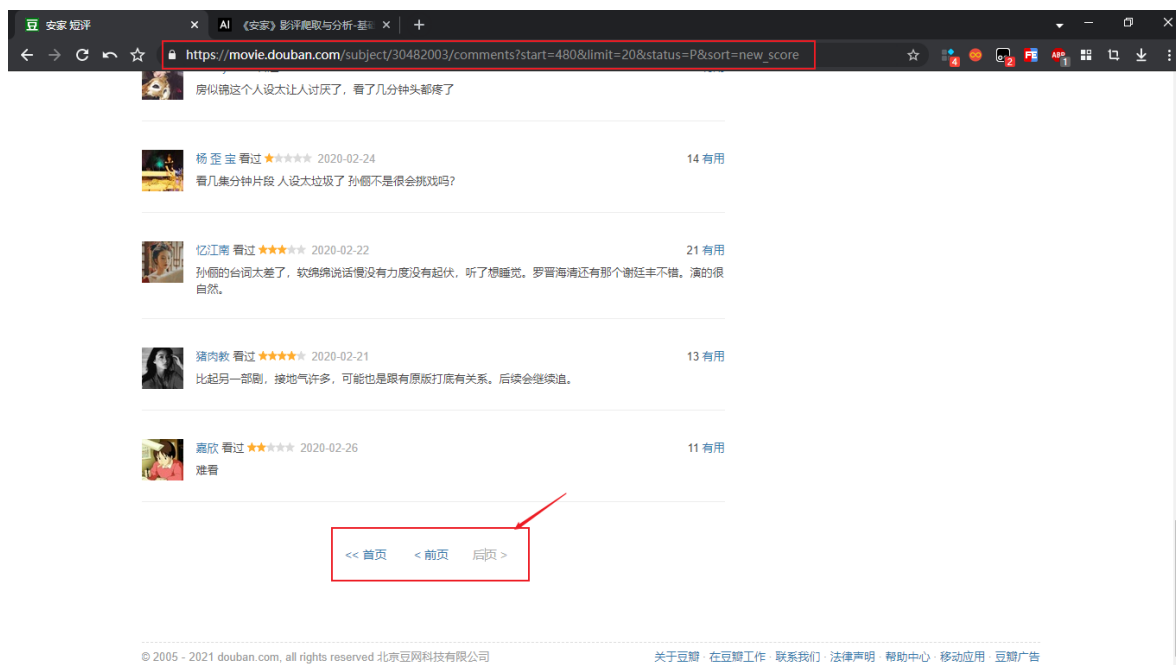
```
12 class AnJiaPipeline(object):
13     def process_item(self, item, spider):
14         # 将数据保存为json格式
15         with open("《安家》短评.json", mode="a+", encoding="utf-8") as f:
16             item = dict(item)
17             f.write(json.dumps(item, indent=2, ensure_ascii=False))
18         f.close()
19         return item
20
```

```
Terminal: Local x +
{'status': '看过',
 'time': '2020-02-21'}
{'comments': '打扰了，这两位是房产中介的店长？',
 'ids': 'Iris',
 'number': '705',
 'rating': '较差',
 'status': '看过',
 'time': '2020-02-21'}
```

```
D:\lures2020-demos\2021_03_18_《安家》影评爬取与数据分析\AnJia>scrapy crawl movie
```

## 3、爬取全部

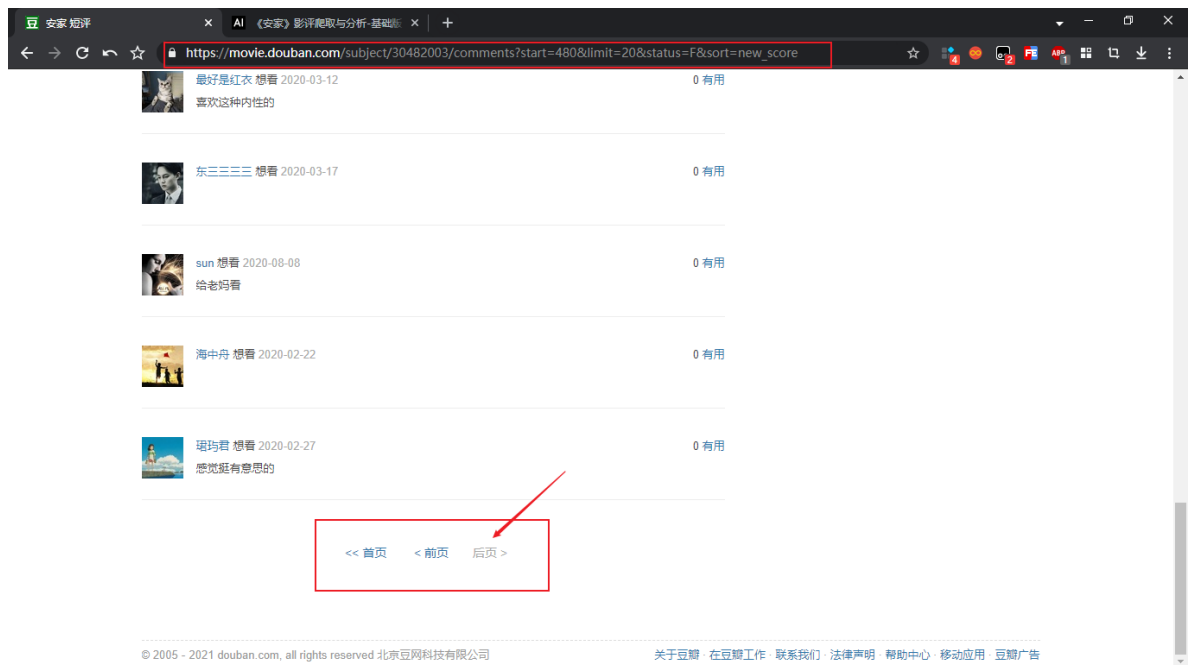
"看过"标签发现最多到480，剩下的没了



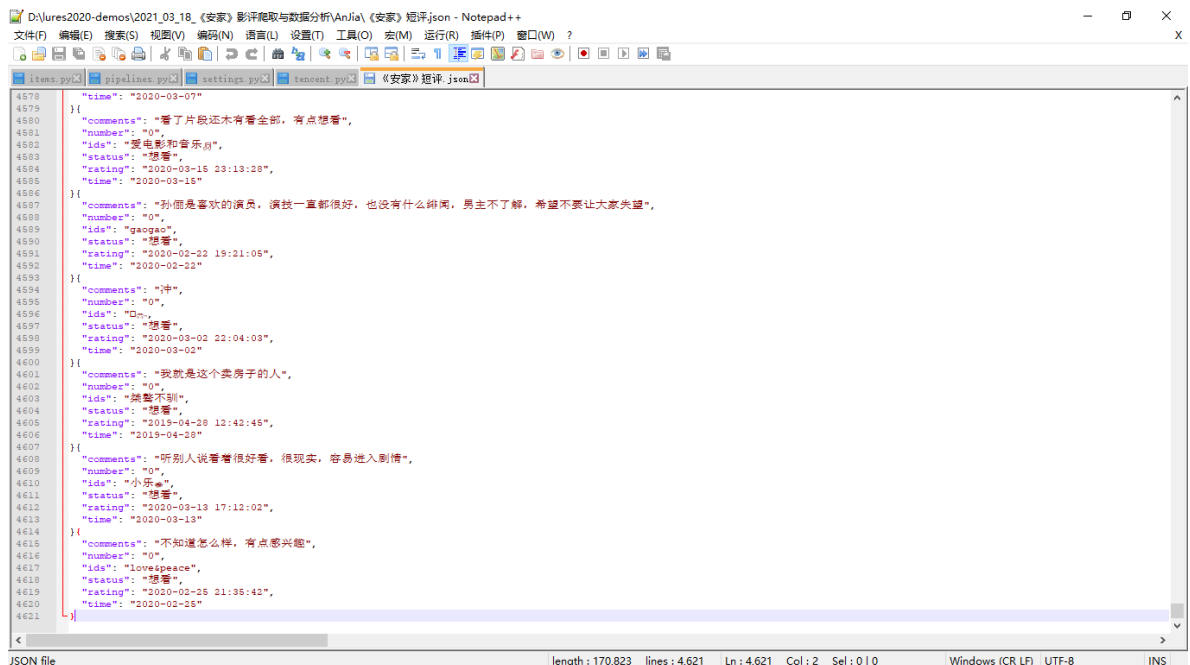
"在看"标签发现最多也到480，剩下的没了，估计是豆瓣公司的设置



"想看"标签发现最多也到480，剩下的没了，估计是豆瓣公司的设置



登录豆瓣前貌似只能爬220条，登录后是500条，全部短评如下：



#### 4、使用jieba分词处理短评

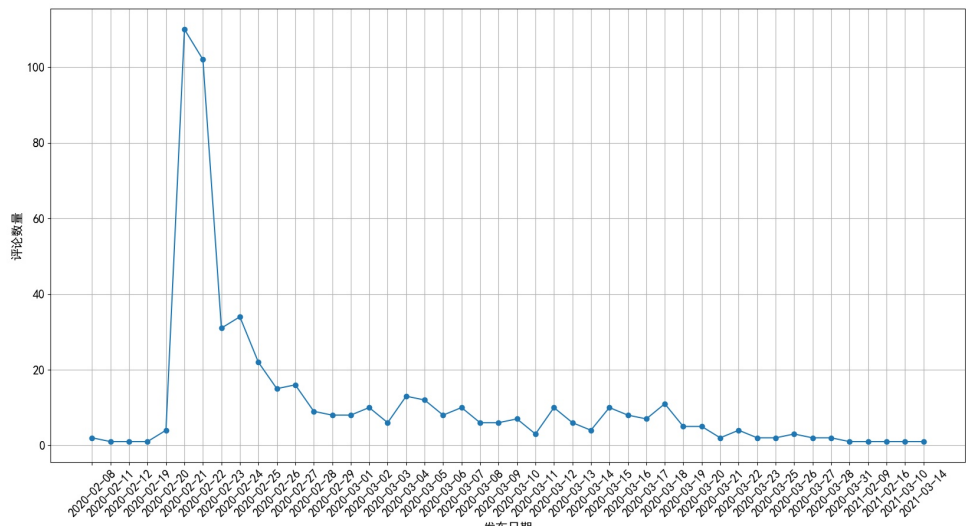
在对应目录下创建 task1\_短评中文分词及绘制词云图.py 用于完成短评的中文分词以及绘制词云图工作！

终端显示：



```
2021_03_18_《安家》影评爬取与数据分析 - ...task2_绘制评论数量趋势图.py
task2_绘制评论数量趋势图.py task1_短评中文分词及绘制词云图.py
63 # 显示图形
64 plt.show()
65
66
67 if __name__ == "__main__":
68     path = "pictures"
69     # 不存在此目录则创建一个
70     if not os.path.exists(path):
71         os.mkdir(path)
72     draw_line_chart(path)
73
74
75 if __name__ == "__main__":
76
Run: task2_绘制评论数量趋势图.py
D:\python\python.exe D:/lunes2020-demos/2021_03_18_《安家》影评爬取与数据分析/AnJia/task2_绘制评论数量趋势图.py
进程已结束，退出代码 0
外部添加的文件可以添加到 Git // 查看文件 总是添加 不再询问 (7 分钟之前)
```

效果图显示：

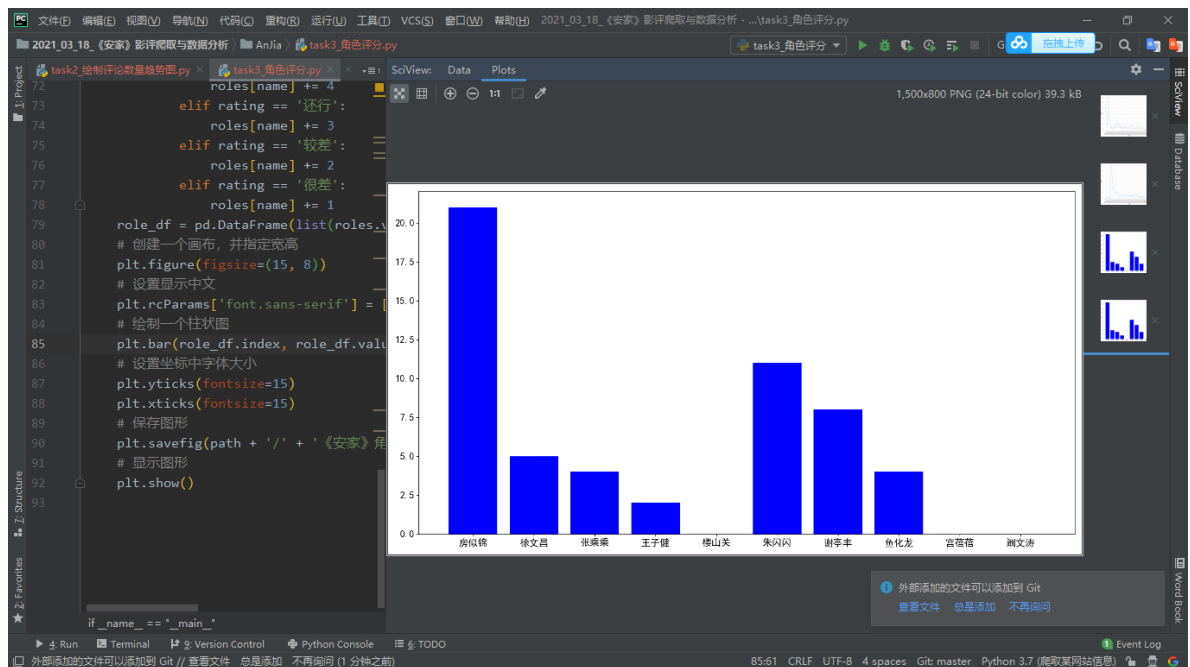


## 6、角色评分

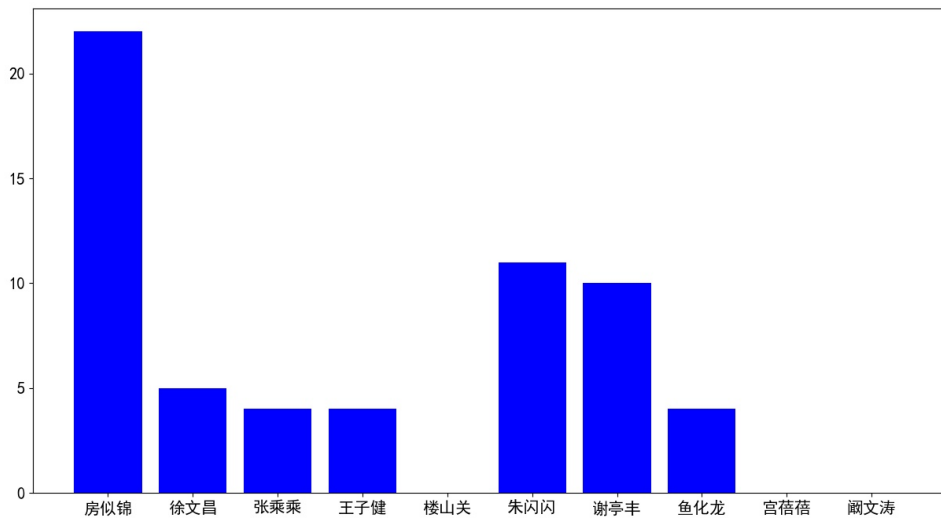
除了剧情狗血，对角色设定的反感也是观众给1星的很大一部分原因。这里的算法是根据评分和内容中出现的角色来进行打分（不是很严谨，但也能说明问题）。举个例子，观众给了1星，然后这个评论内容中出现了几次“房似锦”，大概率说明这个观众对“房似锦”这个角色是比较反感的。其次，1星给1分，2星给2分，依次类推，谁的分高，说明谁更受观众喜爱。

终端显示效果：





实际效果图：



## 7、评分好坏

豆瓣的评分是5星制，5星是力荐，4星是推荐，3星是还行，2星是较差，1星是很差。

部分爬取下来的一些数据，由于用户评价的时候没有给分，因此我们给他归类到“放弃”这部分。

