> 豆瓣电影分析实战之细节篇

- 话接上文，在展示技术实现细节的基础上，本篇文章将着重展示进行信息提取后的呈现细节，毕竟，面向普罗大众的文字和图片要远比代码更直接、易懂。
- 技术细节只需要考虑怎么“实现”即可，但在实际操作过程中，必须考虑为什么做，以及怎么做效果更好。
- 本篇文章将主要围绕着提取有效信息，情感分析，语义理解，文本比较和可视化等四个方面进行阐述，这四个方面并非单纯割裂，而是有机统一，互为表里和因果。

> > 有效信息归并

1. 需要指出的是，本篇分析针对的是一部电影，即《变3》的分析，成果仅做展示用。
2. 在拿到信息后，需要进行提取，因为重点在文本，所以数据分析方面省略处理，但依旧会存留众多杂糅信息，在此背景下，提取信息分为两部分：

- 第一是《变3》的演职人员基本情况；

```
1  1    迈克尔 · 贝 导演
2  2    希亚 · 拉博夫     演员
3  3    罗茜 · 汉丁顿 - 惠特莉   演员
4  4    乔什 · 杜哈明     演员
5  5    泰瑞斯 · 吉布森     演员
6  6    约翰 · 马尔科维奇   演员
7  7    弗兰西斯 · 麦克多蒙德     演员
8  8    郑肯   演员
9  9    凯文 · 杜恩 演员
10 10   朱丽叶 · 怀特     演员
11 11   帕特里克 · 德姆西   演员
12 12   艾伦 · 图代克     演员
```

- 第二是豆瓣评分情况

```
1  主演   希亚 · 拉博夫 / 罗茜 · 汉丁顿 - 惠特莉 / 乔什 · 杜哈明 / 泰瑞斯 · 吉布森 /
2  5 星 12.8%
3  制片国家 / 地区    美国
4  编剧   伊伦 · 克鲁格
5  4 星 36.9%
6  类型   动作 / 科幻
7  豆瓣评分    7
8  3 星 41.1%
9  上映日期    2011-07-21
10 1 星 1.7%
11 2 星 7.5%
12 电影名称    变形金刚 3 Transformers: Dark of the Moon
13 片长   154 分钟
```

- 第三是评论的摘取。

```
1  电影名称    评论人 星级   评论时间     题目   评论内容     有用   没用   回应   地理位置
2  0    变形金刚 3 的影评 (1590)   漫科   1   2011-07-08 02:11:09 麦克贝你丫给我站住
3  1    变形金刚 3 的影评 (1590)   李胶囊 1   2011-07-08 02:11:09 迈克尔贝你这是作
4
```

```
5  2    变形金刚 3 的影评 (1590)    狒狒   1    2011-07-08 02:11:09 记念变形金刚 3
   3    变形金刚 3 的影评 (1590)    KILL88  1    2011-07-08 02:11:09 看完变 3 来详
```

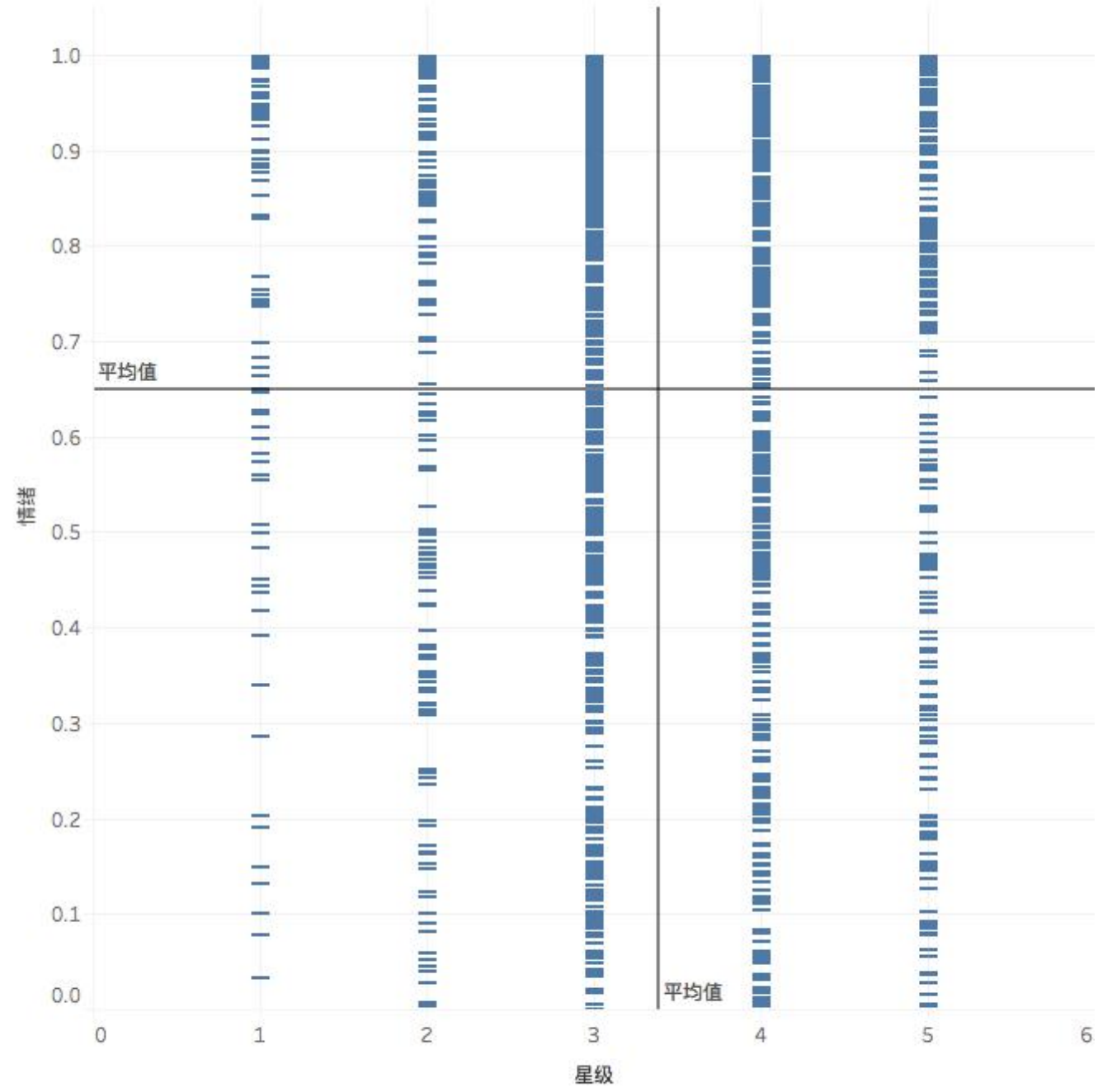
> > > 确定基本信息后，开始文本分析

- 注意到邮件上认为抓取的是变3一部电影的评论,其实并不是一部,而是很多部,但在实践中,我们在分析时,也是要分门别类进行分析,数据分析可以减轻一些人工检索和探索特征的工作,但其也有局限性,筛选有意义的数据,祛除无意义的数据占用了大量时间,因此,为贴合实际,查看汇总信息，并思索下一步。
-
  - 对题目分词并过滤停用词
    针对/霸天虎/战略/我谈/几点/看法 #结果查看
-
  - 对评论分词，并查看结果
    变形金刚/整体实力/2/有所提高/最后/一段/红蜘蛛/枪射/瞎/威震/妞儿/游说/汽车/抱头/击毙/场面/实再/差强人意/机器人/拍得/太少/满意/没有/梅根/惊艳/香肠/嘴/满意/伊利/国内/名声/不怎么样/花/多钱/强奸/人眼/满意/想/大黄

> 1. 情感分析

情感分析实际上是二元分类的推演,通过概率来判断给定样本的概率大小,随后设定阈值,超过此阈值则给出判断,在本案例中,情感分析实际上有某种预示,五星好评和一星差评的情感倾向一定是不同的,这也是进行情感判断的出发点.

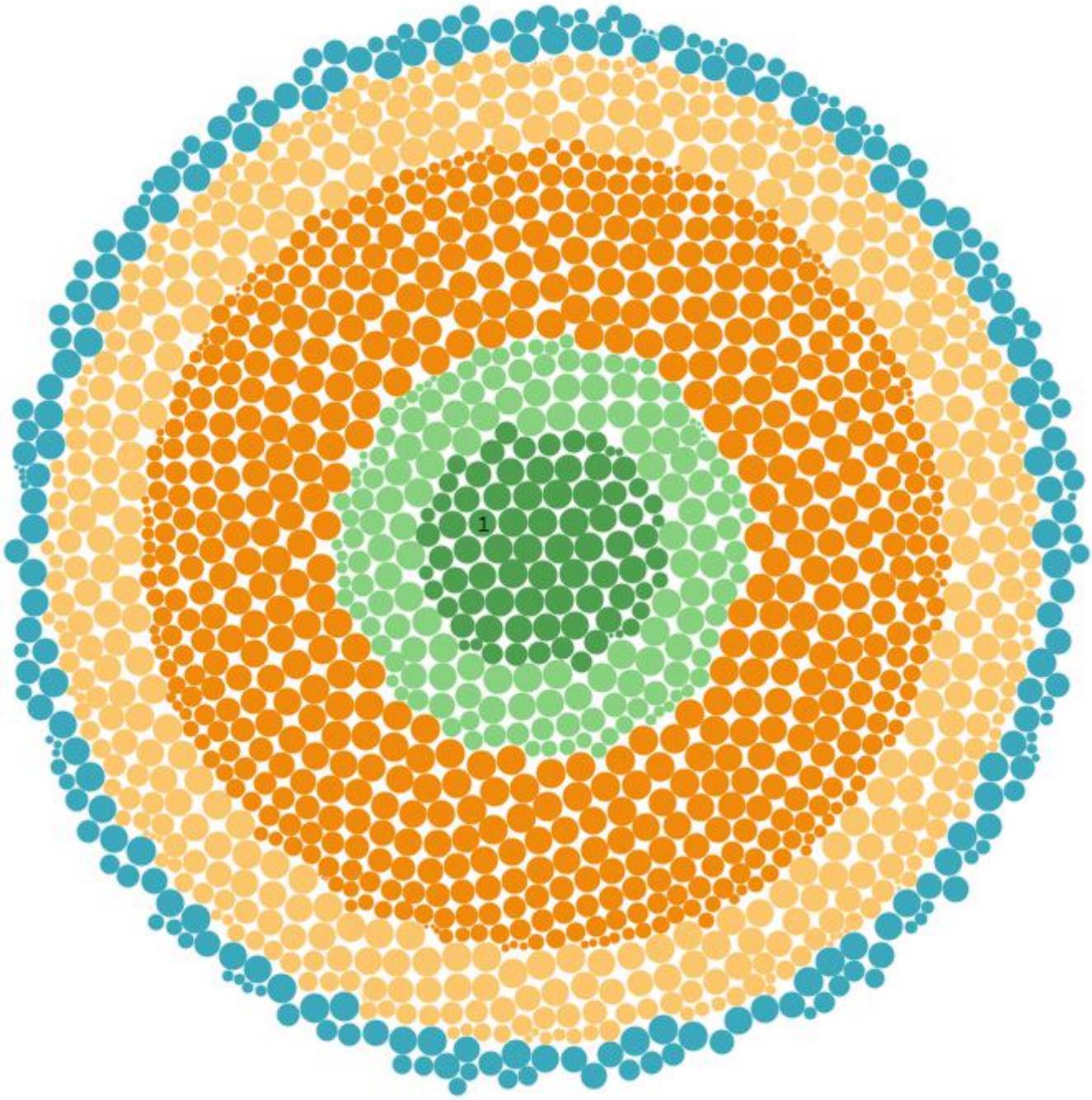
- ['迈克尔贝', '你', '这', '是', '作', '死', '啊'] #案例
- 0.20410726096518372 #情绪得分，情绪得分在（0，1），越高越积极
- 查看分词后评论题目数量
  len(sent)#检查 1570
- 构造新列，将各题目情绪得分作为参数传入，考察星级和情绪得分具有何种关系。



星级 以及 情绪。

从其中可得出两点结论：1.评分人整体上对电影持认同态度,情绪平均值高于0.5；2.星级和情绪持正相关,即评分越高,印象也较为正面.

-
    - 在此基础上，直观感受下这种情绪变化



说明，该图由内向外依次是1-5星级和情绪关系，可直观感受其三星处为高值，和情绪集中趋势吻合，即情绪高于均值，但也仅仅高于了

评论文本语义理解

1. 查看评论列情况

```
1  0    这篇影评可能有剧透                                  很多人怕，说你不要写变3影...
2  1    这篇影评可能有剧透                                  本戏的主要剧情进展表  开...
3  2    真的猛士，敢于直面惨淡的人生，敢于正视推迟的档期。这是怎样的哀痛者和幸福者？然而…
4  3    买拷贝说，世界上有300个专家，上亿的粉丝，我拍这个电影是给粉丝看的，不是给那300个
5  4    1.  没有理由不看 3D IMAX 版。 2.  前半部分基本上可以看做是 Lenov...
6  Name: 评论内容, dtype: object
```

- 回应邮件主题，查看各主要演员和什么话题相关

-
  - 很遗憾，没有完全做出来，只做出来导演一个人的，原因有点复杂，主要因为中国观众特别喜欢给外国明星起花名[1],比如迈克尔贝叫“买拷贝”，严格意义而言，此类情况可以消除，但是比较复杂。我用卖拷贝索引出的主要话题如下：
    变形金刚、爆米花、圈钱、汽车人

2. 当然，此类索引运用的是分聚类和相似度，将评论文本序列化，并构造索引系统，利用得出的高频主题或感兴趣的主题进行索引，示例如下（示例，选用第一篇文档）：

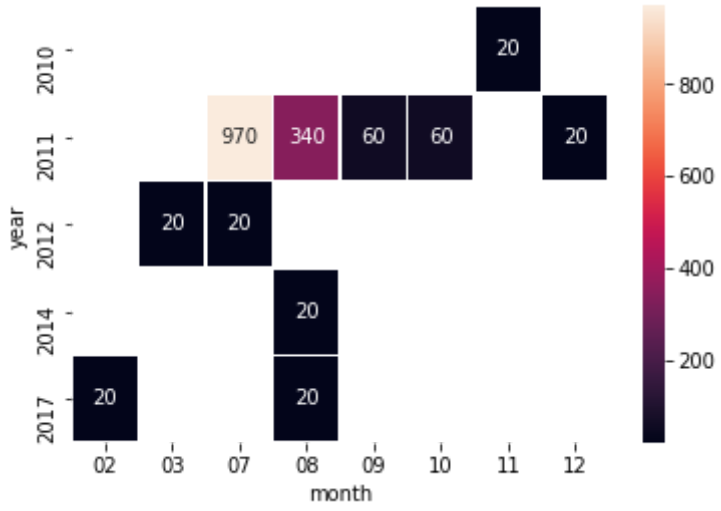
- 返回如下：

```
0
 这篇影评可能有剧透                                    很多人怕，说你不要写变3影评了，我怕看完
1079
 这篇影评可能有剧透                                          看过变形金刚1，曾经认为，又一部
10
 这篇影评可能有剧透                                    这片坑爹大了，当然还是有的是人咆哮几句‘
484
 之前看了好多吐槽的影评，各种抨击、纠结、响应，然后相拥而泣。                    我想，既然花这么
1550
 这篇影评可能有剧透                                          好吧，我大脑进水了，七夕去看电
1065
 这篇影评可能有剧透                                          首先，作为一名根正苗红的801，i
501
 这篇影评可能有剧透                                       大家都说变3只有广告没有故事，我是这
1405
 这篇影评可能有剧透                                          长期的历史实践告诉我们，大部分
1193
 这篇影评可能有剧透                                          虽然一开头，男主就穿的是美邦，
1529
 这篇影评可能有剧透                                    没看过一和二，只看了这一部。TRANSFORM
```

- 分别返回第0，1079，10，484，1550，1065，501等10篇影评，发现其主要聚焦于特效，3D，广告等话题，隐约感觉其为批评，返回其情绪分，为0.2876345890679868，果然不高，其聚合效果尚可。
  正规说法：回溯第一篇影评的直观感受是"讽刺',其情绪得分较低,而此时返回的相近文本具有相同的感受,即"不满

时间序列主题可视化

1. 在前文基础上，展开对主题的可视化操作，回溯数据发现，评论延续到2017年，较长时间可为可视化提供较好的呈现。
2. 标注热度（月度）

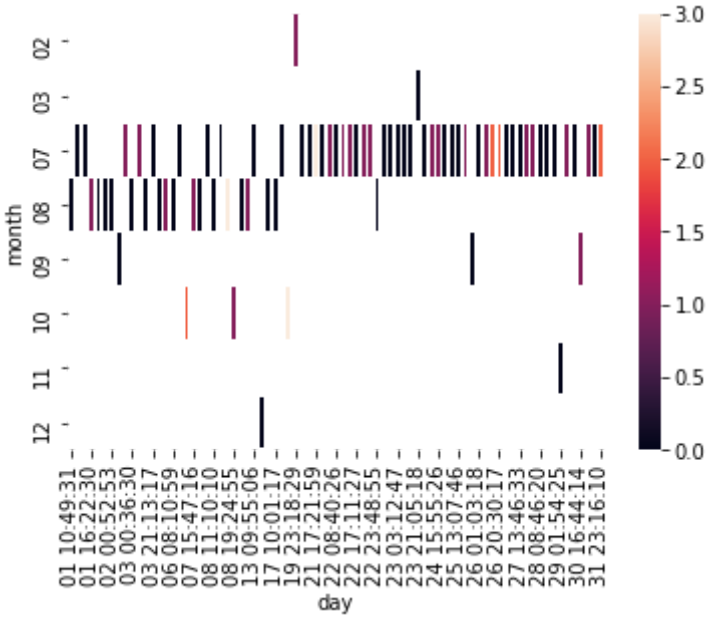


- 可见其最高热度出现在2011年7月份和8月份，此时正式其在北美上映后被引入国内的时期[2],可见电影的热度主要再其上映前后。

3. 看条神评：

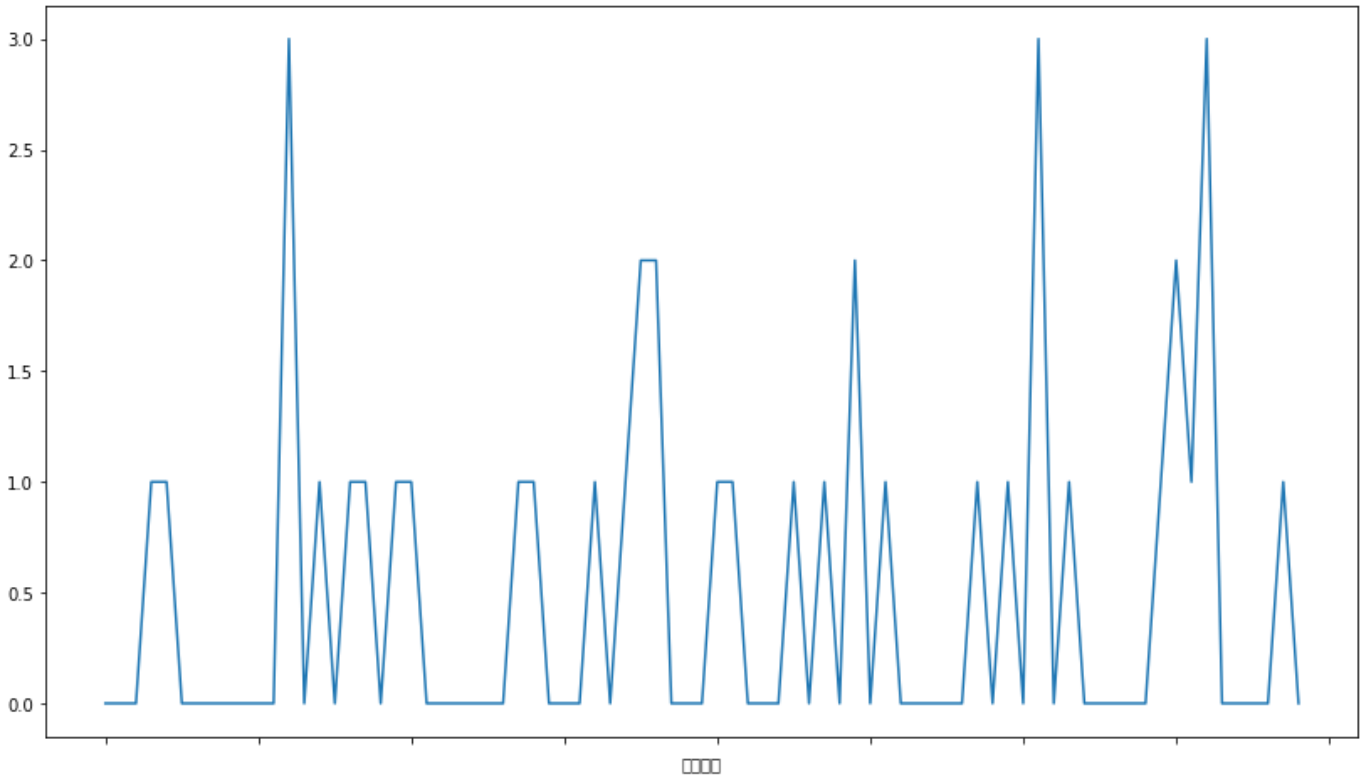
```
1  真的猛士，敢于直面惨淡的人生，敢于正视推迟的档期。这是怎样的哀痛者和幸福者？然而上映又
```

- 至此时，我特别想寻找下“评论”一词的热度，将其可视化，当然，这是我感兴趣的，你不一定感兴趣。



指出一点是，按天这一更细的时间粒度来划分，然后我发现还是7月和8月，看来暑假学生们不上学，上班的太热，还是窝在家里看电影吧，这个结论对吗？我不知道。

last :画幅走势图吧



一个结论，不平稳，对变3这部而言，热度是不均衡的

总结

1. 情绪整体偏正面，但总而言之并不是受到广泛赞誉，甚至可以说饱受争议；
2. 对导演而言，吃瓜群众主要谈论的是电影用来圈钱一类的话题，其他演员，花名我实在搞不懂；
3. 话题主要是特效等技术层面，但在此话题集中的背后，显示的是剧情的薄弱；
4. 随后进行了一些可视化。

---

1. 大家来说说中国粉丝给外国明星起的昵称 ↩
2. 变形金刚3_百度百科 ↩