特别声明：

本软件仅供个人交流学习使用，请勿用于商业使用(BUG 超多的哦)。

源代码及程序下载：

github: https://github.com/SumTwilight/spider_bilibili_comment/tree/GuiAndTread
链接: https://pan.baidu.com/s/1Ht_WRfhOzPRoALSRBKm_NQ 提取码: 2333

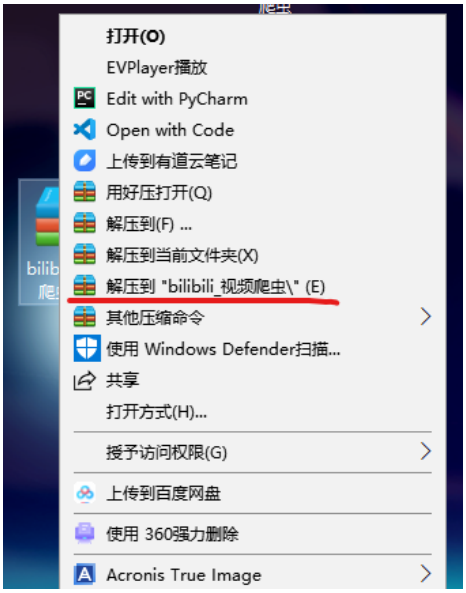
Copyright：SummerTwilight

问题 or BUG 反馈：

ncutzl@outlook.com
(联系了我大概率也懒得改 emmmm)

Bilibili_spider 使用说明

1. **先将压缩包解压在一个文件夹中**



2. 打开软件和你想爬取的 b 站视频网页：

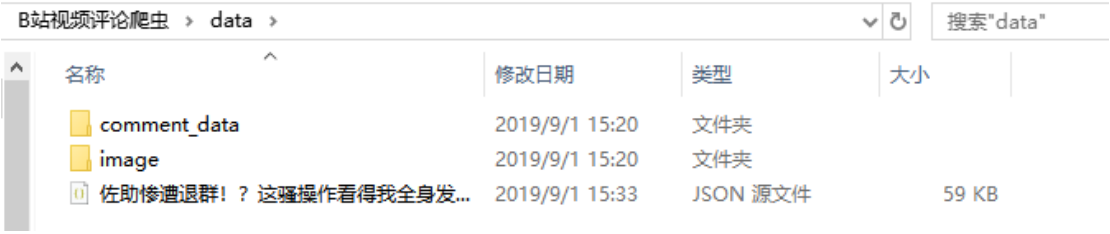
3. 复制网页链接粘贴在这里

4. 点击按钮后等待提示

5. 输入要爬取的评论页数（不要大于总页数，会出错，尽管有错误处理，但是我也不知道会发生什么 emm）后点击按钮

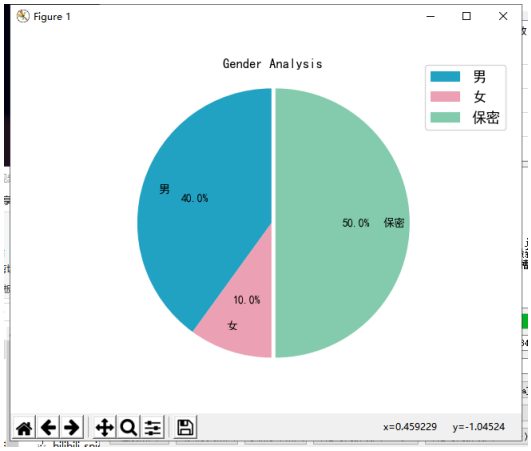
6. 点击“开始爬取”。
7. 爬取完成后：

我一般都不会点开去看那些数据（直接在网页上看不好吗？）



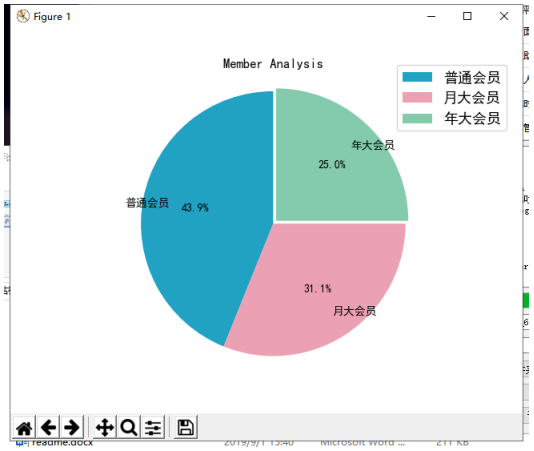
这就是保存的数据。

8. **点击“加载分析数据”后，继续点击下方按钮，就可以进行简单的数据分析**
9. **以下分别是本视频的六个数据分析图像和根据图像可以得出的简单结论**
   - 性别分析



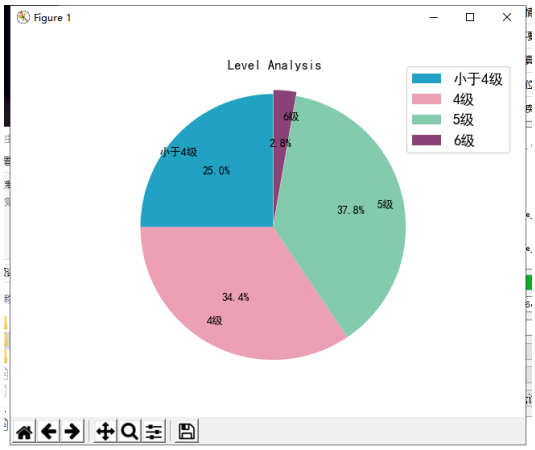
果然对火影感兴趣的大多是男生~（tips：B 站用户在注册如果没有选性别的话会自动设置为保密。）

   - 用户会员分析



B 站用户的消费能力还是不错的，大会员占了有四分之一，现充用户占了过半~

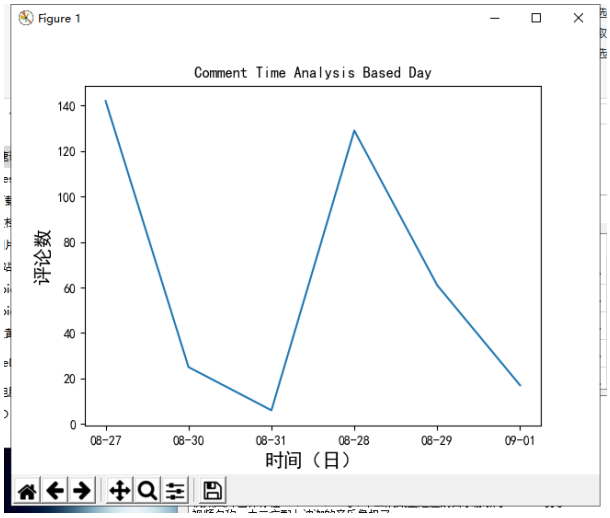
- 用户等级分析



B 站这个视频下大多都是老用户，并且以 4，5 级大佬居多~，6 级用户果然很少。

Ps：其实根据上面三个数据就可以用概率论相关知识，分析一些相关性，例如用户中男生充会员和女生充会员之间的关系，用户等级和会员之间的关系等等等啦，但是——我懒。

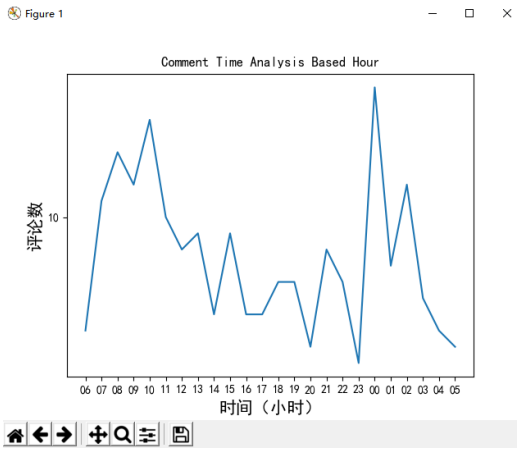
- 评论数与评论时间（日）



横轴是日期，竖轴是当天的评论数

根据本图可以看出，大多数的评论集中在 27，28 日？？
图有点 BUG，这是因为我这点日期直接拷到列表里面，它是根据内存的顺序而不是根据日期的顺序，导入的。然后我懒得改了。(还需要写一个排序的函数。

- 评论数与评论时间（小时）



横轴是当天时间（从早上 6 点到凌晨 5 点），竖轴是当天评论数。
从本图中可以看出大家大多是在早上 9 点 or10 点（刚起床）和晚上 00 点（上床睡觉时）评论（可以侧面反应出大家在这两个时间点看视频的居多）

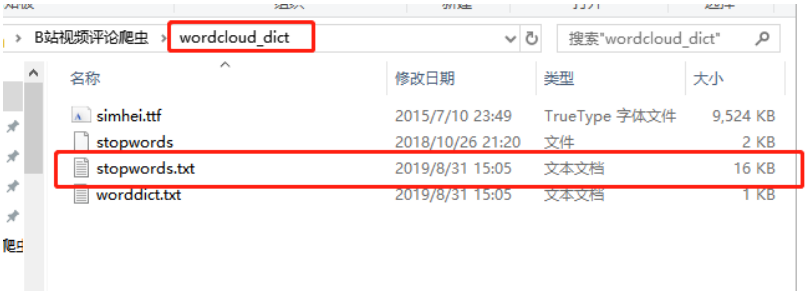
- 评论词云分析

最后这个是词云分析，就是将评论中出现的词抓取出来，按照出现次数排序，次数越多，在图中字就越大。
从此图中我们可以知道本视频的主题是佐助，博人，我爱罗，博人转等等。

关于词云分析的进阶：

1. 消除无效词语

这些我圈出来的词语对于我们来说都是无用的，但是因为在评论中出现的次数很多，所以被抓取了出来，我们可以通过在 stopword.txt 文件中添加上述词语，就可以让生成的图中不再有它们。



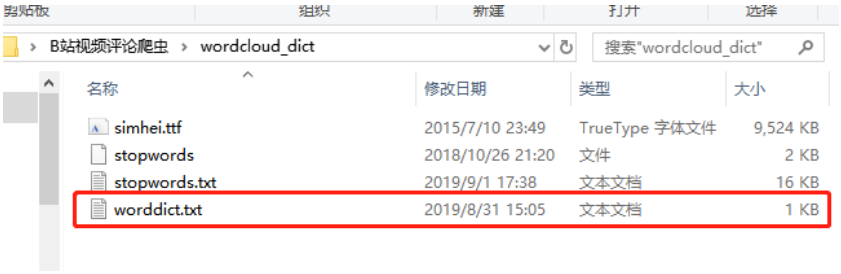
Eg：在 stopword.txt 中添加 amp 后保存重新点击“评论词云图”

啦啦啦啦~，amp 这个单词就没有了~!

2. 自行设定关键词

词云分析中的词语都是常用词语，我们爬取的视频中也许有很多在它原始文件中没有存储过的词语，例如 bilibili，哈利波特，宇智波佐助，漩涡鸣人。。。。。

像这样的关键词，它自身是无法识别出，我们可以在



这个文件中添加与爬取视频相关的特定词，以便程序进行更为精确的词云分析。

格式：词语 （一个空格） 1

其他：