> 一开始以为20k全要标，就一直往后赶，跳着去标了，感觉这样好像标了很多的感觉😶所以当时只标了因果句，没细致划分.....
>
> 因为跳了太多内容，很零散，我用Screening.py把标了的数据提了出来，方便之后处理。数据放在new_smooth.xlsx。
>
> 提取了之后，我发现实际只标了1000多条，有的不好意思。🫣

- 在这个数据集中，我主要观察的关键词是“导致”，“由于”；还有一些“随着”，“因此”，“由于”，“愿意”等关键词。
- 我目前只手动标注了因果句，但没有标注因、果、关键词（sorry🥹）。
- 对于比较确定没有因果句的，标注是“null”。
- 对于标注的因果句，我是用英文分号划分“;“
  不同的因果句，有一部分似乎用的是中文分号（原文中直接保留的以及我不小心切换了输入法），到时候需要稍微注意一下。
- 我观察到的一个问题是数据集里面很多**因果句根本不是金融事件**，而且比例不小。如下例：

```
1．由于有限的互联网接入，我将无法及时回复您的消息
2．但由于存在能力不匹配、分布不平衡等问题，凸显了危险废物处置能力的不足
```

- 那些几千字以上的长文本，常常会有上述的情况；而短文本很多没有因果关系，但是如果有这样的关系，大概率是我们想要的金融事件，可能是短文本的口水话较少、内容概括性高的原因。
- **一段话如果包含多个因果关系**或者**多段话件之间存在因果关系**，我目前是未分开的。如下面两例：

```
1．由于过去的基础太过薄弱，而道路建设又是一个地区经济社会发展和居民生活水平提高不可忽视的重要方面。
因此，各地都在如何加快公路建设、特别是高速公路建设方面花了很大的力气、下了很大的功夫、投入了很大的人力物力和财力。这也使高速公路建设债务大幅增加，需要通过收费来偿还债务。

2．由于目前网约公交车平台的活跃客户量、用户黏性以及影响力都不高，因此偶尔会出现预约乘客过少，形成只能被迫取消的情况，而预约成功率降低会进一步降低乘客对网约公交车的依赖，形成一种非良性的循环。
```

- 有一些句子虽然包含了一些因果关键词，但我主观感觉不是真的因果关系，我目前没有特别标注出来，暂时忽略了。例子如下：

```
1．通常来说其股票指数在接下来一个月中会平均跑赢全球指数达3.5%。这有可能是由于球迷激动庆祝的情绪传播到了金融市场，推动强劲买盘。
2．整体市场上是不是由于政策限制了供应商，因此业务受到影响，不能使用某个供应商提供的设备？
```