2-3日 new idea

现阶段已经了解了BERT transformer 的工作原理。 并且希望借助已经有的预训练模型，应用到我们的下游任务。

关于虚假诉讼

诉讼可以分为三个阶段： 诉讼原因，诉讼过程，诉讼判决。

虚假诉讼会导致第三方的利益受损。 这个受损是结果上的。受损的表现方式在民间借贷这个案由上表现为资金的流动。

> 资金的流动： 既包括事实上的流动和潜在的流动（有执行财产和无执行财产都算） ， 也包括流动和不流动。 只要涉及资金，就构成对第三方损害的潜在性。

而从 当事人实施一个虚假诉讼的角度来说，可能需要串通原被告， 伪造证据等方式。 要用人工智能去识别文本中潜在的作假证据或者当事人的串通是困难的。

相对容易且合理的方案是去对资金的流动进行非显示的分析，去判断其对潜在的（未知）的第三方的侵害的可能性。

同时可以配合对过程的分析，去评价这是一个个案还是一个普遍的案件。

实验设计

动机

目前我们用transformer_sentence 的预训练模型对案件的`content`字段进行embedding，但是结果并不呈现出明显的分块，也就是不能很好的聚类。

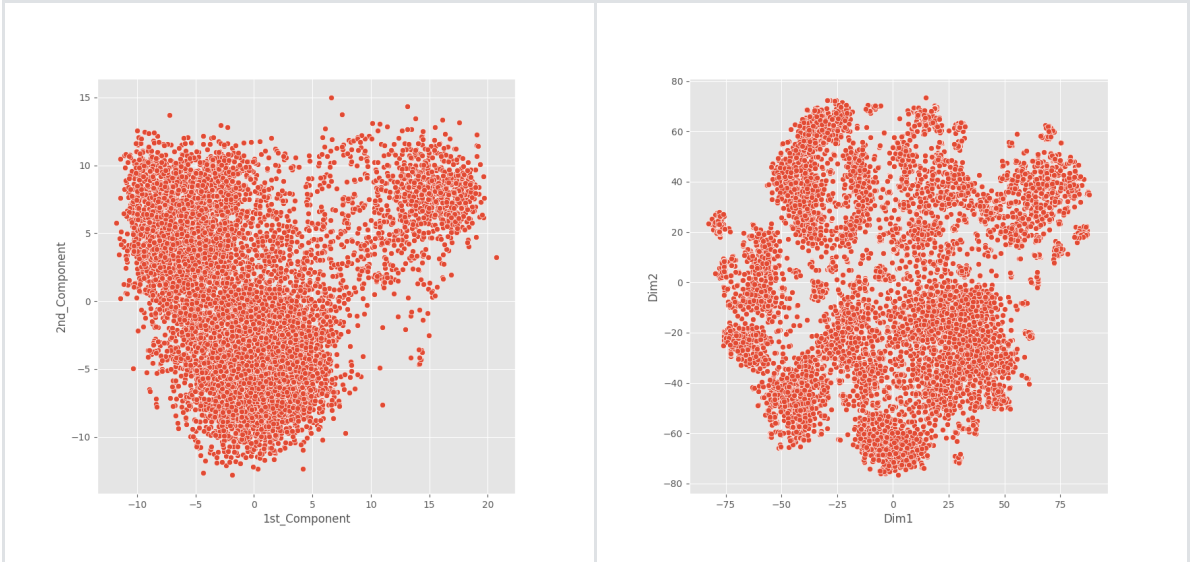
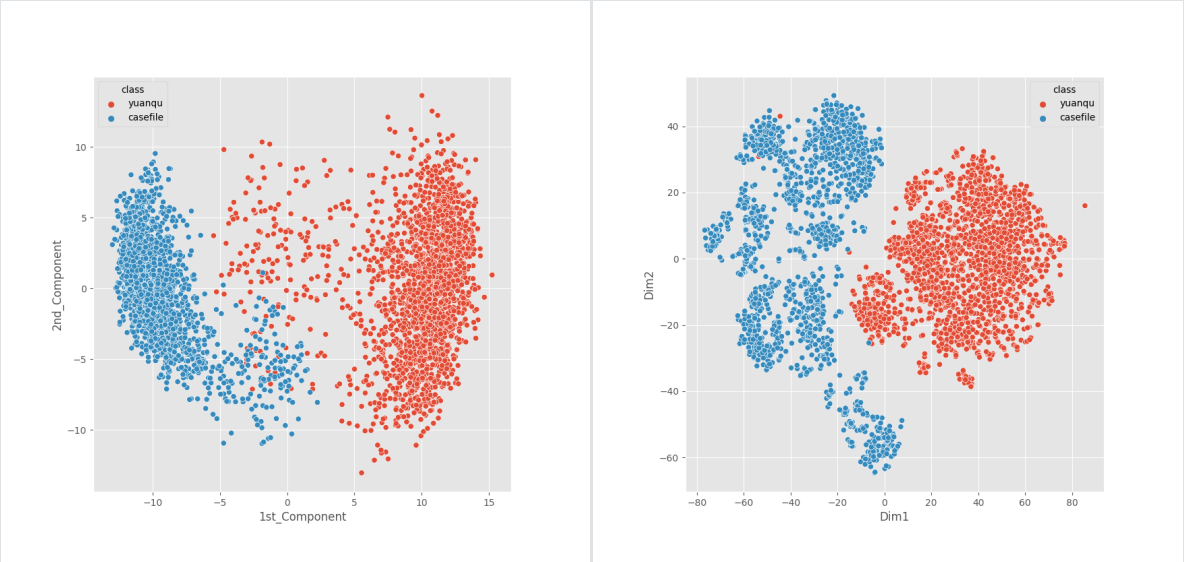


图1 对 随机抽取的10000个案件进行嵌入并降维的结果

可能存在的原因有：1. 可能是预训练模型泛化能力太弱。 2. 可能是2维空间太少，降维的过程中丧失了过多的信息

然后我用元曲数据集进行了嵌入，观察其嵌入结果是否有明显边界。 如果有明显边界，则初步排除预训练模型泛化能力问题。 而是输入的数据集太大，噪音太多，导致案件的点过于密集，连城一片。



可以看到，两个类别样本有明显的界限，接下来是一个svm分类的结果， 同样说明两者之间分类非常容易

	precision	recall	f1-score	support
casefile	1.00	1.00	1.00	1900
yuanqu	1.00	1.00	1.00	1900
accuracy			1.00	3800
macro avg	1.00	1.00	1.00	3800
weighted avg	1.00	1.00	1.00	3800

	precision	recall	f1-score	support
casefile	0.96	0.96	0.96	600
yuanqu	0.96	0.96	0.96	600
accuracy			0.96	1200
macro avg	0.96	0.96	0.96	1200
weighted avg	0.96	0.96	0.96	1200

用sklearn naive 的smc模型，在训练集只有5%比例的情况下，依然能完全正确。 在映射到一维空间的情况下，按照0.3的测试集比例，依然有0.96的正确率。

以上实验结果表明，下一步实验应该清洗content字段，提取有用信息进行嵌入。

实验思路

首先用无监督的方法提取关键词，然后从完整的文本里面切割出判决部分，然后进行嵌入。

实验预期

相比于对整个content进行嵌入的情况，本次的嵌入结果应该具有更加明显的内部界限。