毕设梗概

毕设的主题限定于对大规模日志序列数据的挖掘与分析，目前重点在MOOC数据集上，如果有余力我想再做一两个数据集。

理由：

1 有兴趣，从数据中构建特征以一种新的方式审视问题我觉得是一件很刺激的事
虽然相关材料看着都挺难，但是看得下去。

2 目前接触的几个数据集都是日志类型，对相关的工具熟悉一点

3 应用广泛

MOOC数据集：对问题的重新评估

数据集是表示‘点击’‘观看’等操作的字符串组成的序列，
与数值不同，数值型数据的大小就可以反映其特征。但字符串是‘符号’，直接让机器了解符号的含义是不现实的，我查阅了自然语言处理相关的材料，觉得符号序列其实和文本识别以及翻译等任务的解决方案是很接近的。所以我提出的方案是使用词嵌入的方法来构建mooc数据集中行为符号的词义空间。

存在的难点：

1 对于人来说，‘点击视频’和‘播放视频’这两个操作间直观的相似度是高于‘提出问题’和‘播放视频’之间的。我想确保词嵌入构成的词义空间里，这种相似性仍然存在，也就是训练词嵌入时要让网络学习到符号间的相似性。

2 设备性能瓶颈，兼顾性能的编程方式。经常跑预处理筛选数据的过程中内存溢出，确实数据量大了，问题就都出来了，小问题也被放大了。

解决方案：

1 目前通过优化遍历方式以及设置精度与数据格式暂时解决了

2 剔除了部分主观上觉得优先级不高的特征
如： 剔除操作在开课时间内的分布位置，只保留密集操作间的组合与顺序。
因为若开课时间过长会导致特征过于稀疏。

3 近期有考虑配置一个AWS服务器，kaggle和google的云似乎是因为内存策略的原因预处理时比pc更容易爆内存。

其它

手上还有一个通信基站上下行流量的日志数据集，出于对信号做降维以及平滑噪声的目的，先选用rfft，根据各频率的权重，想人为设定阈值过滤低权重的频率分量以过滤噪声，但效果不好。又试着用tensorflow写了个自编码器来对数据重新编码并解码，结果发现还原出的信号与原信号的误差还可以接受，但是对于大的峰值（高于均值3倍）还是拟合不到，还在看自编码器的材料，还没想到优化的方法。