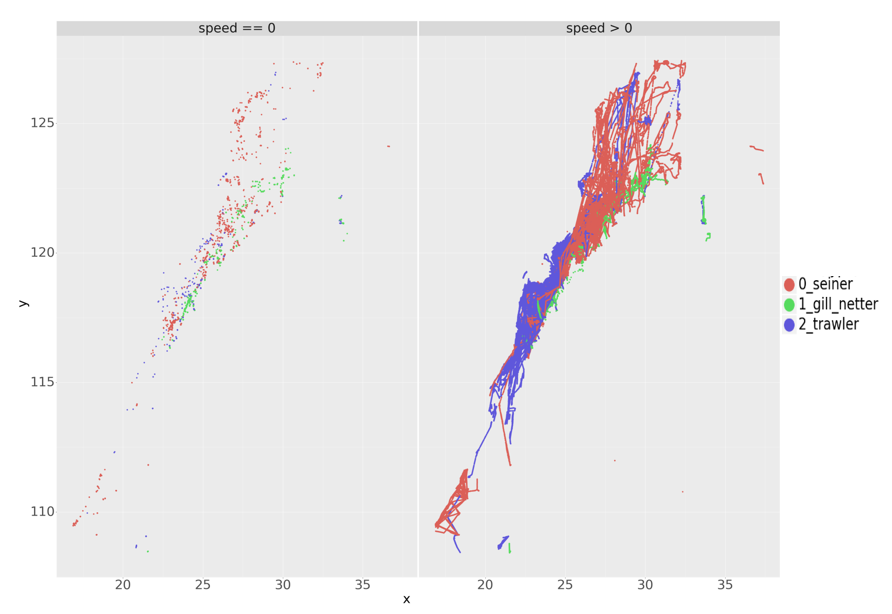
具体方案：

根据查询的相关资料，清楚了解到渔船针对海中生物群和水域差异，拖网、围网、刺网采用不同路线、不同速度、作业时间差异设计出的捕捞方式。根据这些特点，我们为此提取三个方面的特征，最后用lgb的gbdt模型进行学习，未进行模型融合。

**No.1 按照同一个渔船id速度为0和非0两部分进行分析**



思路：

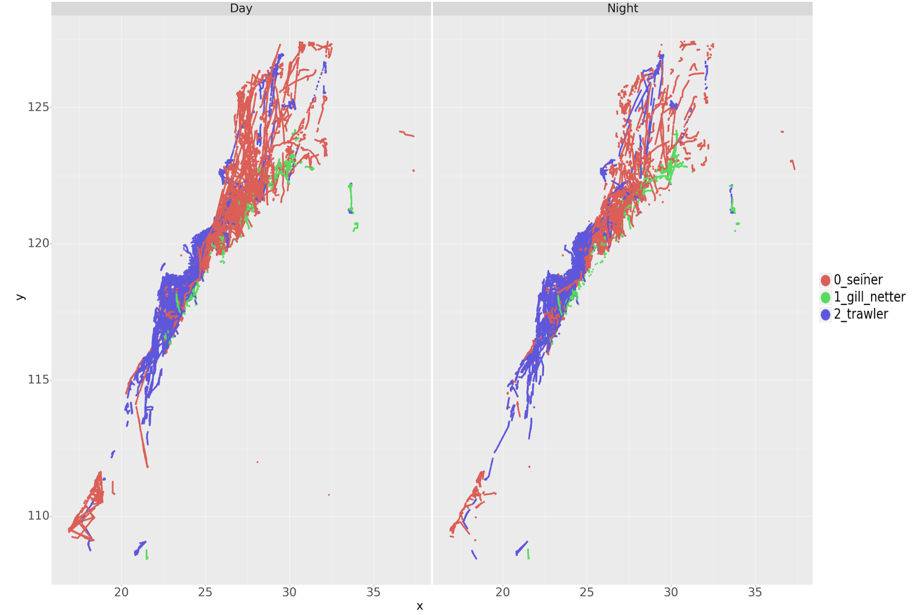
1、针对同一艘渔船，将其数据分为 速度为0和非0两个部分。分别统计该船在速度为0 和 非0情况下，经纬度xy、方向direction这些原始特征的变化情况（均值、方差、极值、峰度、偏度等统计特征）

2、根据1构建的特征，原始特征被构造出一系列统计特征，一种含义的特征会被分成速度为0和非0情况。根据这个特点，对这些特征进行一个比值处理。

**No.2 渔船在白天和黑夜会按照同一个渔船id白天和黑夜两部分进行分析**

早6点整至晚8点整设置为白天(**图标识Day**)

晚8点整至早6点整设置为黑夜(**图标识Night**)



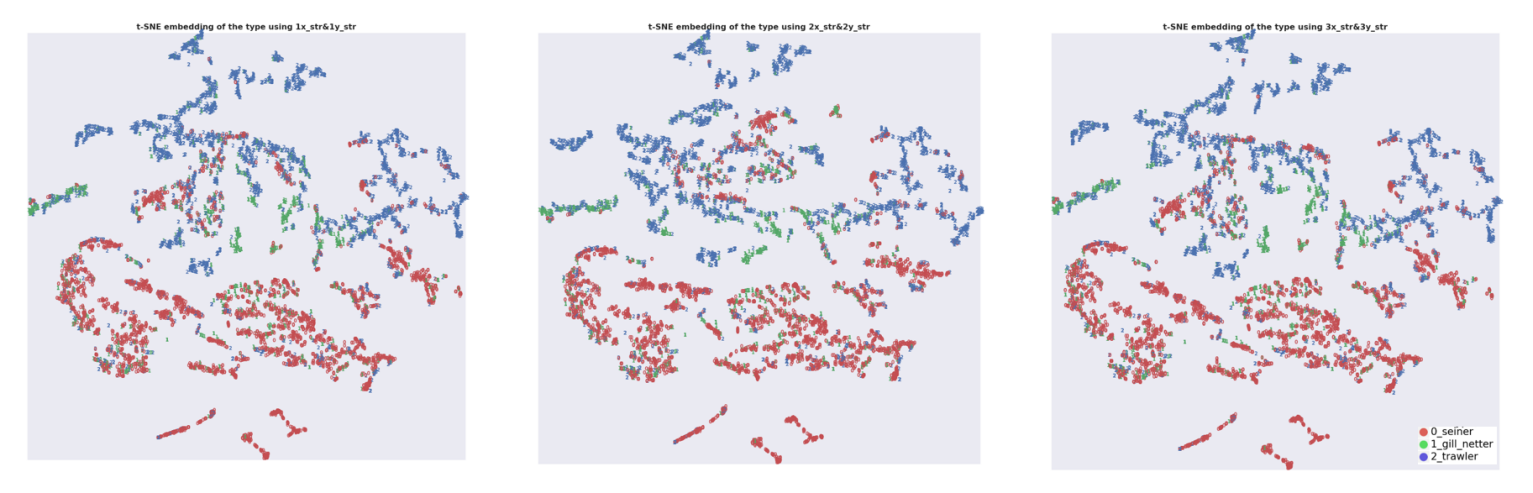
思路：

1. 与No:1类似，数据按照时间划分成白天和黑夜两部分，分别统计该船在不同时间内的作业情况。
2. 根据1构造的两组时间特征，提取关键的速度speed、经纬度xy进行白天与黑夜特征的对比。

**No.3 借鉴自然语言处理（NLP）角度去处理船的轨迹特征**

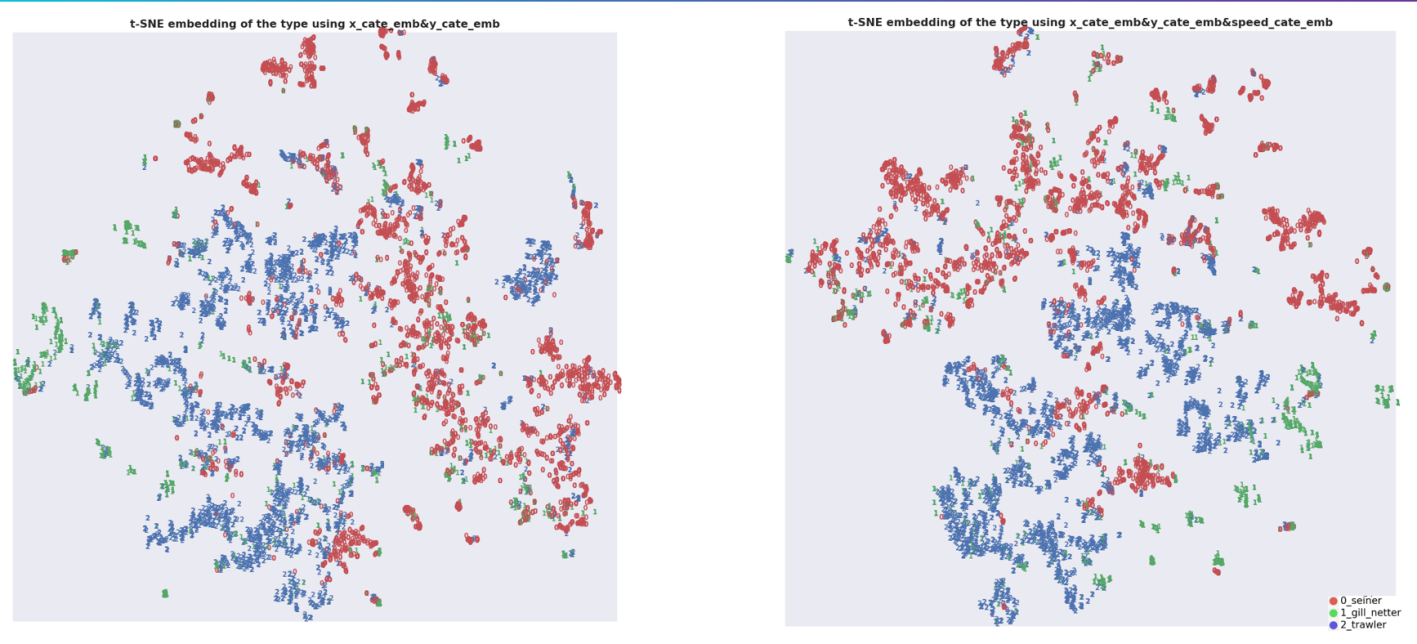
速度speed、经纬度xy按照作业时间排序，可以反映出每艘船的行为规律。而每种作业方式都有其内在的一些规律, 借鉴自然语言处理(NLP)的相关算法进行特征提取。利用nlp的算法对速度、经纬这些序列的学习，尝试挖掘出每艘船的行为特点。

思路一：TF-IDF + NMF(如图，从左到有分别是ngram=1, ngram=2,ngram=3，经过t-SNE降维的可视化结果)



1. 使用不同的ngram去处理每个渔船的速度、经纬度数据，提取出每艘船的TF-IDF特征（**ngram=1, 2, 3**）。
2. 并利用非负矩阵分解(NMF)算法，对处理后的速度、经纬度进行降维生成一个主题分布向量。（此题目分成了8类）。
3. **对每个渔船的主题分布向量进行T-SNE降维，进行可视化。**

思路二：word2vec（左图：利用经纬度； 右图：利用经纬度和速度）



1、每个渔船id的速度、经纬度看做是一个序列信息

2、利用速度、经纬度的分位数统计量，将浮点特征分桶转成一个类型特征。（降低过拟合带来的影响）。

3、使用深度学习的word2vec的CBOW算法无监督训练，获取经纬度(x-y)和速度(speed)的类型向量。

**4、将每个渔船每一步的类型向量取平均得到渔船的特征向量，并采用T-SNE降维，进行可视化。**

思路三 预训练模型（由于时间和数据等因素未在线上实验）

利用NLP的相关技术word2vec有一定的效果，由此可扩展到无监督预训练模型（如BERT/XLNET/ELECTRA等），利用大量无标记数据（即只需要利用所有未标注的渔船航行信息），可极大提升下游任务（分类/匹配/排序等）的效果；