

# 奥运视频热议度分析

## 一、数据介绍:

hot\_or\_not: 评论情况  
type\_sv: 话题类型  
nar\_angle: 叙事角度  
zf\_sv: 态度倾向  
length\_log: 视频长度 (对数)  
title\_sv: 视频标题  
media\_type: 发布者类型  
media\_fans\_log: 发布者粉丝数 (对数)  
type\_time\_eff: 时效性  
month1: 季节性  
love\_sv\_log: 视频点赞数 (对数)  
com\_sv\_log: 视频评论数 (对数)

变量	变量名	类型	取值范围	说明
因变量	评论情况	定性变量	热议、非热议	抖音网站收集短视频评论数取对数,再以阈值=2做分组后得出
	视频长度	连续变量	0.6~3.0	抖音网站收集,统一整合成秒为单位,由于呈尖峰分布,故对其作对数变换
	话题类型	定性变量	传统话题 非传统话题	抖音网站收集短视频主题文本,依据短视频主题文本进行分类
	叙事角度	定性变量	被动回应 品牌宣传、明星团队 要素解读、突发事件 赛事周边	抖音网站收集短视频主题文本,依据短视频主题文本进行分类
	态度倾向	定性变量	正、负、中	抖音网站收集短视频主题文本,依据短视频内容进行态度判断与分类
	发布者类型	定性变量	主办方及授权传播主体、非授权商业媒体、 体育专业媒体、其他机构或个人自媒体	从抖音官网采集的原始视频数据,按照当下新媒体传播主体的主要类型对其进行分类
	发布者粉丝数	连续变量	0.3~8.1	抖音网站收集发布者总粉丝数,其呈尖峰分布,因此对其作对数变换
	时效性	定性变量	时间敏感、短期、长期	发布时间从抖音网站上获得,事件发生时间由视频内容或相关新闻报道中检索获得,二者相减后以2天,60天为分界点进行分组
	季节性	定性变量	1: 7或8月发布、 0:其他月份发布	从抖音官网采集短视频发布事件中的月份信息进行分组

## 二、分析任务

1. 绘制评论量和点赞数的直方图
2. 将评论量对数进行分组 ( $\geq 2$  为“热议”,  $< 2$  为“非热议”并将原数据中其他符号类型变量转化为因子变量
3. 描述性分析,重点分析视频长度、话题类型、叙事角度、态度倾向这四个自变量对因变量的影响,要求作图+解读
4. 将样本按照 8: 2 比例随机生成训练集和测试集 (种子设置为 2023),针对视频是否为热议视频这个分类任务,在训练集上实现决策树、AdaBoost 以及随机森林模型 (注意: 需在训练集上利用五折交叉验证进行参数调优),最后在测试集上测试三种模型的效果并进行对比,需作出 ROC 曲线图并计算 AUC 并写出模型比较的结果。