2023 年全国大学生数据分析大赛

基于文本内容的违规信息识别

近年来，短视频平台越来越受到广大民众的喜爱。短视频的快节奏、视觉冲击力以及互动性，吸引了亿万人的眼球。但是随之而来的问题就是平台上涌现出了大量的低俗、暴力等违法违规内容。这些内容严重影响了社会公共道德和互联网生态，也让短视频平台面临着巨大的挑战。面对日益增长的用户上传量，人工审核已经无法满足需求。因此，短视频平台开始运用人工智能技术来协助审核工作。AI 技术可以实现对视频中的文本、图像等信息的自动分析和识别，例如色情、暴力、恐怖、政治敏感等内容都可以通过 AI 技术进行检测。AI 技术对于重复性高的视频审核起到了很好的辅助作用，能够在短时间内处理大量的违规内容。本题采集了某短视频平台的大量文本数据，请根据提供的数据进行数据的清洗、分析与挖掘，并回答下列问题。

1. 分别获取违规(sensitiveness.csv)和非违规(insensitiveness.csv)数据，并进行数据合并得到整体数据集。(10 分)

2. 绘制统计图分析所有文本内容的词长；分析违规信息数量和非违规信息数量的占比。（20 分）

3. 对数据中的评价内容进行分析，制作词云图，给出违规信息中出现次数最多的 10 个词。（20 分）

4. 构建主题模型分析所有违规信息的主题，并对各个主题进行分析。

（25 分）

5. 建立违规信息的识别模型，并对模型的性能进行评估。基于模型对附件中的测试数据 test.xlsx 进行评测，将评测结果补充到第一列中，并将此文件一起上传到竞赛平台。(25 分)