**《信息分析》课程作业**

**——线性降维与聚类分析**

**专业：** 信息管理学院 **学号：** 211820073 **姓名：** 胡涂

**专业：** 信息管理学院 **学号：** 211840182 **姓名：** 邓雨茵

**专业：** 信息管理学院 **学号：** 211820274 **姓名：** 郑宇

# 基于K-means的视频垃圾评论识别方法

**摘要：**内容文字用楷体，五号字

**关键词：**楷体；五号字；不多于五个词；各词之间用分号分隔

## 前言

## 数据集及特征构建

### 数据集介绍

数据来自Kaggle平台的Youtube Videos and the Comments(<https://www.kaggle.com/datasets/japkeeratsingh/youtube-videos-and-the-comments/code>)。 共379032条记录，5项特征。

数据字典如下：

|  |  |
| --- | --- |
| 特征名 | 描述 |
| User | 视频作者 |
| Video Title | 视频标题 |
| Video Description | 视频描述 |
| Video ID | 视频ID |
| Comment (Displayed) | 展示评论 |
| Comment (Actual) | 真实评论 |
| Comment Author | 评论作者 |
| Comment Author Channel ID | 评论者频道ID |
| Comment Time | 评论时间 |

在数据集中，并没有提供可供分析的特征以及相应的标签，所以特征需要人工进行构建，并进行无监督学习以达到识别与标注目的。

### 描述性特征构建

### 语义特征构建

除了上述描述性特征构建，我们小组考虑了对评论源文本进行向量化以进行相似度比较。

垃圾评论在语义上可能并没有任何的相似程度，但是有价值的评论在语义上是较为相似的，在这种情况中，我们希望对文本进行语义上的比较，而不是仅仅比较单词或字符的相似性。因此，我们需要将文本转换为能够表示语义信息的向量。

而传统的词袋模型只关注文本中单词的个数，而不考虑单词之间的关系以及语义信息。

因此在语义向量化的任务中，我们采用BERT预训练模型进行向量化操作。

BERT-Whitening是一个使用预训练的语言模型BERT来进行文本表示的技术，以及针对该表示做出的一种嵌入学习策略。BERT是一种双向Transformer模型，它在预训练阶段使用了大量的无标签数据，将该数据输入模型并将模型调优，在此过程中模型可以构建出单词和句子的表示，并不仅仅是简单的词向量，同时包括了语义和上下文信息。然后，BERT-Whitening使用协方差矩阵和白化矩阵来处理BERT表示，使相似的文本在嵌入空间中更加接近，从而提高文本相似度计算的效果。

## K-means及结果选择

## 总结