概率论与数理统计

扩展任务 8 试验报告

GHe

Build: 2025-04-03

Typst Version: 0.13.1

目录

1 **程序设计和原理阐述** **1**

1.1 词频-逆文档频率（TF-IDF） 1

2 **程序实现和结果分析** **1**

2.1 源码代码及其实现原理 1

1 程序设计和原理阐述

1.1 词频-逆文档频率（TF-IDF）

TF-IDF 是一种统计方法，旨在评估某个词对于一个文档的重要程度，其核心思想是：一个词在文档中出现的频率越高（TF），同时在语料库中的常见程度越低（IDF），则认为这个词越能代表文档，其重要性越高.

Definition 1.1.1 (词频(TF)): 词频衡量了某个词在文档中出现的频率，词频越大，词对当前文档的重要性可能越大.

$$\mathrm{TF}(t,d) = \frac{\text{词 t 在文档 d 中出现的次数}}{\text{文档 d 的总词数}}$$

Definition 1.1.2 (逆文档频率(IDF)): 逆文档频率衡量了某个词在整个语料库中的普遍性或稀缺性，逆文档频率越高，词在整个语料库中越稀有，则认为这个词越能代表文档，其重要性越高.

$$\mathrm{IDF}(t,D) = \log\left(\frac{\text{语料库的文档总数 N}}{\text{包含词 t 的文档数} + 1}\right) + 1$$

2 程序实现和结果分析

2.1 源码代码及其实现原理