ICS 13.020.01
CCS Z 06

团　　体　　标　　准

T/ACEF □□□－20□□

环境文本数据加工处理技术规范

Technical specification for processing environmental text data

（征求意见稿）

20□□-□□-□□发布　　20□□-□□-□□实施

中 华 环 保 联 合 会 发 布

目　次

前　言

本文件按照GB/T1.1-2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由生态环境部环境规划院提出。

本文件由中华环保联合会归口。

本文件起草单位：生态环境部环境规划院、北京市生态环境监测中心。

本文件主要起草人：李勃 王建童 卢亚灵 赵大地 张鸿宇 蒋洪强 杨懂艳 马俊文。

环境文本数据加工处理技术规范

1 适用范围

本文件规定了环境文本数据加工处理的程序、方法及技术要求。

本文件适用于环境文本数据的加工处理。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 42777 基于文本数据的金融风险防控 知识图谱构建技术框架指南

GB/T 34950 非结构化数据管理系统参考模型

3 术语和定义

下列术语和定义适用于本文件。

3.1

逆向文件频率 inverse document frequency, IDF

一个词语普遍重要性的度量，由总文件数目除以包含该词语的文件数目，将得到的商取对数获得。

3.2

词频 term frequency, TF

某一给定词语在文件中出现的次数。

3.3

关系 relationship

实体与实体之间在特定时间、特定行为下产生的联系。
［来源：GB/T 42777-2023, 3.2］

4 环境文本数据加工处理程序

环境文本数据加工处理涵盖从原始环境相关文本数据的采集，到利用现代自然语言处理技术进行清洗和预处理，再到采用嵌入算法将文本非结构化数据转化为结构化数据的全过程。加工处理程序分为数据采集、数据标准化与清洗、文本嵌入与向量化等步骤，见图 1。

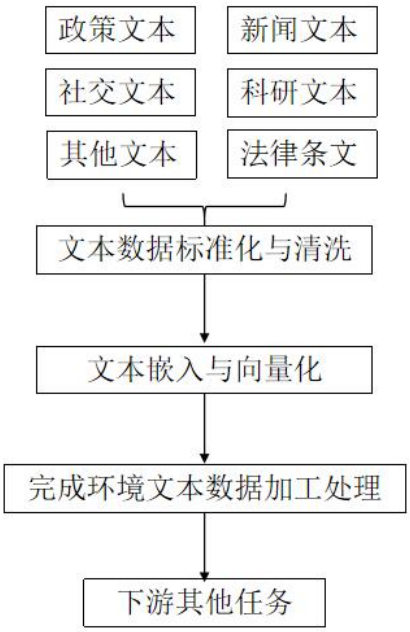

图 1 环境文本数据加工处理程序

5 数据采集

首先确定目标来源，如政府网站、新闻门户、社交媒体平台、学术数据库及法律信息库等；对于版权允许的文档，可进行下载。

6 数据清洗

主要包括去除无关字符、统一编码格式、转换大小写、校正拼写错误、消除重复项等步骤。具体操作如删除标签、数字、特殊符号；将所有文本转换为小写形式；使用自然语言处理工具纠正错别字；过滤停用词（如“的”、“是”等常见但无实际意义的词汇）；以及修正不一致的表达方式。

6.1 清洗前评估

在进行数据清洗之前，应对不同类型的环境文本数据进行评估，确定适当的处理方法。对于 PDF、Excel、TXT、Word、JPG 等多源异构文本数据，评估内容主要包括：

(a) PDF 文件应检查是否含有嵌入式图像或扫描件，采用 OCR 技术提取文字;

(b) Excel 表格应确认是否存在空值、重复记录或格式不一致等问题;

(c) TXT 和 Word 文档应核查是否存在特殊字符、编码错误或段落缺失等情况;

(d) JPG 图像应评估图像清晰度，确保 OCR 识别的准确性。

根据上述评估结果，采取相应的数据清洗措施，例如文本格式转换、去重、缺失值填充、异常值处理等，确保数据质量和一致性。

6.2 清洗策略

将 PDF、Excel、Word、TXT 等不同格式的文本数据统一转换为纯文本格式，如特殊字符、数字、标签等；进行分词处理，将文本分割成词语单元。统一词汇形式，包括转换为小写、去除停用词、词干提取或词形还原等操作。

6.3 清洗后验证

通过随机抽样检查、统计指标分析、重复性与完整性检验、逻辑一致性校验等进行清洗后的验证与质量控制。从清洗后的数据集中随机抽取样本，人工检查数据的质量。计算数据集的基本统计量，如逆向文件频率、词频等，评估数据的一致性和完整性。检查是否有重复文档或词汇，确保数据集没有冗余信息；同时核实数据集是否完整，没有遗漏重要文本。对文档内容进行逻辑一致性校验，确保文档内部信息相互一致，没有明显的逻辑错误或矛盾。

7 数据处理

7.1 实施流程

采用嵌入算法将文本数据转化为结构化数据，进而可以表征不同数据之间的关系，具体实施流程包括数据准备、模型选择、特征提取、语义分析和结果整合。

（a）数据准备：清洗、标注文本数据，确保输入质量，为后续步骤奠定基础；

（b）模型选择：挑选适合场景的嵌入模型，提升特征表达能力；

（c）特征提取：将文本映射为向量形式，捕捉词语间语义关系，便于机器理解；

（d）语义分析：基于向量空间进行相似度计算，识别文本深层含义，辅助决策；

（e）结果整合：汇总分析结果，形成可解释信息，支持进一步应用或模型优化。

7.2 质量控制

通过抽样人工验证的方式来评估模型的表现，确保模型在不同的数据集分割下表现稳定，使用合适的评估指标来衡量模型性能，如准确率、召回率等，同时也可以使用困惑度来评价词嵌入的质量。

附录 A

（资料性）

环境文本加工处理实例

图 A.1 展示了将文本非结构化数据转化为结构化数据的过程，包括数据采集、数据清洗、数据处理等三个步骤。

图 A.1 环境文本数据加工处理实例