文本信息熵对比分析报告

一、研究背景

信息熵作为量化信息不确定性的核心指标，在语言信息处理领域具有重要应用价值。本研究通过构建中英文平行语料库，分别从字符与词语维度开展熵值计算，揭示两种语言的信息结构特征差异。

二、数据与方法

1. 语料构成：

- 中文：2019维基百科语料（预处理保留纯汉字）
- 英文：Gutenberg文学语料（标准化处理后使用）

2. 处理流程对比：

中文语料：
汉字提取 → Jieba分词 → 分块处理（优化内存）

英文语料：
文本清洗 → 小写转换 → 空格分词

三、实验结果

语言维度	字符级	词语级
中文	9.85	13.36
英文	4.16	9.73

关键发现：

1. 词汇级熵值普遍高于字符级（+36%中文，+134%英文）
2. 中文字符熵显著高于英文（2.37倍差异）
3. 高频词影响显著（如英文"the"占比达6.2%）

四、语言特性分析

1. 中文特征：

- 字词分离的表意体系
- 汉字字形信息冗余度高
- 词语组合灵活性强

2. 英文特征：

- 字母系统的有限符号集
- 形态变化的规则性约束
- 功能词高频集中现象