**Report of Deep Learning for Natural Langauge Processing**

21374019 孙效宇

21374019@buaa.edu.cn

**摘要（Abstract）**

本报告探讨了中英文文本的平均信息熵，分别以字和词为基本单位计算信息熵。我们使用NLTK工具处理英文语料库Gutenberg，采用jieba库处理中文语料库wiki\_zh。结果显示，中文的字信息熵和词信息熵均高于英文，体现了两种语言在结构和表达上的差异。

**引言（Introduction）**

信息熵是度量文本信息量的重要指标，反映了文本的不确定性程度。本研究旨在比较中英文文本的平均信息熵，揭示两种语言在字母（字符）和词汇层面的复杂性。通过计算中英文文本的字信息熵与词信息熵，我们可以更深入理解语言结构，并为自然语言处理（NLP）任务提供理论支持。

**方法论（Methodology）**

1. 英文文本信息熵计算

英文语料库的选择上，我们选用了NLTK中的Gutenberg语料库，提取了《哈姆雷特》作为实验文本，提取文本中的字母统计频率，按空格分词统计词频，分别利用香农公式（式1）计算了字母信息熵和词信息熵。

2. 中文文本的信息熵计算

中文语料库的选择上，我们采用wiki\_zh语料库，其中包含AA至AM多个子文件夹，每个文件夹中有多个名如wiki\_00的无后缀文本文件。中文文本的字信息熵计算方法与英文类似，即直接统计所有文本字符的频率并计算熵，但中文的词没有空格区分，这里引入jieba分词器对文本进行分词，统计词频并计算熵。这里由于中文语料库较大（共5.65亿字符），为避免内存不足问题，我们将文本分块处理，每次处理100万字符。

英文实验结果：字母信息熵约为4.4813 bits每字，词信息熵：约为 9.0451bits每词

中文实验结果：字信息熵约为9.1328 每字，词信息熵约为11.1843 bits每词

**表 1 中英文字词平均信息熵对比**

|  |  |  |
| --- | --- | --- |
|  | Bits/Letter | Bits/Word |
| Chinese | 9.1328 | 11.1843 |
| English | 4.4813 | 9.0451 |

**结论（Conclusion）**

实验结果表明，中文的字信息熵显著高于英文，这是由于中文字符承载的信息量更大。而在词层面，中文词信息熵也高于英文，反映出中文词语组合的丰富性和复杂性。这些结果对自然语言处理任务如机器翻译、语言建模等具有重要参考价值。未来研究可进一步探讨不同语料库的影响，并尝试更高效的分词算法和计算方法，以提升实验的准确性和可扩展性。