实验报告：中英文本信息熵计算与分析

1. 实验目的

本次实验的核心目标是加深对信息论中“熵”这一核心概念的理解，并探究其在自然语言处理中的实际应用，具体目的如下：

- 利用网络爬虫工具，从互联网上分别采集足量的英文和中文文本样本，并清洗乱码，去除噪声。
- 设计并实现算法，计算所采集样本中英文字母和汉字的香农熵（Shannon Entropy）。
- 通过逐步扩大文本规模（例如 2M、3M、4M、5M、6M 字符等不同尺度），重新计算并分析信息熵的变化规律。
- 将实验计算结果与课件中给出的理论值进行对比，并对实验方法和结果进行深入分析，探讨其“伸缩空间”。

实验项目中所涉及到的技术细节和代码已上传至 GitHub 代码仓库，链接地址为：https://github.com/originality666/nlp_homework/tree/main/entropy_task。

2. 实验原理

信息熵是信息论的基石，用于度量一个随机变量的不确定性程度。对于一个由离散字符组成的文本，其信息熵定义为每个字符的信息量的统计平均值。其计算公式如下：

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \cdot \log_b P(x)$$

其中，$\mathcal{X}$是随机变量 X（字符）的取值空间，$P(x)$是字符在文本中出现的概率。熵的单位通常是“比特/字符”（bits/char），表示每个字符平均携带的信息量。熵值越高，代表文本的随机性越强，字符分布越复杂。

3. 实验环境

（1）编程语言： Python 3

（2）核心库：

- requests：用于发送 HTTP 请求，获取网页的原始 HTML 内容。
- BeautifulSoup4：用于解析 HTML 文档，高效地提取纯文本内容。
- matplotlib：用于数据可视化，绘制文本规模与信息熵关系的曲线图。
- argparse：用于解析命令行参数，提高脚本的灵活性和可配置性。
- collection：用于统计字符频数

4. 实验步骤

整个实验流程主要分为两个阶段：数据采集和数据计算分析。

4.1. 样本采集与预处理

（1）数据来源

为保证样本的多样性，本次实验从多个网站采集数据。

- **中文样本：**从 https://www.gutenberg.org/网站中采集的中国古典文学作品节选，如《红楼梦》、《三国志》、《西游记》、《狄公案》等共 16 部作品，附录中给出了详细的作品网址。
- **英文样本：**从 https://www.gutenberg.org/网站中采集的外国古典文学作品节选，如 Pride and Prejudice、MOBY-DICK、Alice's Adventures in Wonderland、Frankenstein 等共 8 部作品，附录中给出了详细的作品网址。

（2）爬虫实现 (scraper.py)

实验使用 Python 编写了 scraper.py 脚本来自动化样本采集过程。该脚本首先从一个预设的 URL 列表（chinese_seed_urls.json 和 english_seed_urls.json）读取待抓取的网址，网址在附录中给出。随后，它模拟浏览器行为发送 HTTP 和 GET 请求，获取网页内容。

（3）数据清洗

从网页直接获取的 HTML 源码包含了大量与文本内容无关的标签（如 <script>, <footer>）和样式信息。为提取干净的文本，clean_text 函数执行了以下清洗操作：

- 利用 BeautifulSoup 库解析 HTML 结构。
- 移除所有<script>, <style>, <footer>等非内容承载标签。
- 提取并拼接剩余标签中的文本内容。
- 使用正则表达式将文本中连续的空白符（包括空格、换行符、制表符）替换为单个空格，确保文本格式的规范化。
- 清洗后的文本被保存为本地.txt 文件，用于后续分析。

4.2. 熵计算与规模扩展 (analyzer.py)

（1）算法设计

熵的计算和分析由 analyzer.py 脚本完成。其核心算法步骤如下：

- **语料加载：**使用 load_corpus 函数读取指定语言类型的所有.txt 样本文件，并将其合并成一个单一的、庞大的基础文本语料。
- **字符频率统计：**设计 char_entropy 函数，并利用 collections.Counter 对整个文本进行遍历，高效地统计出每个唯一字符的出现次数。
- **概率与熵计算：**根据字符频率计算每个字符的出现概率，然后代入香农熵公式，计算出整个文本的平均信息熵。

（2）文本规模扩展

为了分析熵随文本规模的变化，expand_text 函数被用于扩展文本规模。具体

方法是：将加载的基础文本语料进行重复拼接，直到达到目标长度（如 200 万、500 万字符），但实际如果目标语料规模小于实际语料规模时，并不需要复制拼接，直接截取到目标要求的精确字符数即可，本项目的中文和英文语料规模均在 700 万字符左右，满足统计规模。通过逐步扩大文本规模，可以不断逼近真实的字符分布。

5. 实验结果与分析

5.1. 计算结果

我们分别对中英文本在 200 万至 600 万字符的不同规模下进行了熵值计算，结果如下。

表 1 中文语料熵计算结果

文本规模 (字符)	独立字符数	香农熵 (bits/char)
2,000,000	5,614	8.9793
3,000,000	6,062	9.0842
4,000,000	6,186	9.0994
5,000,000	6,520	9.1209
6,000,000	7,122	9.2224

表 2 英文语料熵计算结果

文本规模 (字符)	独立字符数	香农熵 (bits/char)
2,000,000	112	4.4571
3,000,000	114	4.4485
4,000,000	121	4.4574
5,000,000	131	4.4580
6,000,000	137	4.4581

5.2. 与理论值对比分析

我们将实验结果与课件中给出的理论值进行比较。

表 3 实验值与理论值对比

语言	实验最大熵值	理论值
英语	4.46	4.03
汉字	9.22	9.71

- **英语分析：**实验测得的熵值（约 4.46）略高于理论值（4.03）。主要原因是，我们的计算单元是**字符**，包含了大小写字母、数字、标点符号和各种特殊符号，而理论值通常是基于 26 个英文字母的统计模型。字符集的增大直接导致了不确定性的增加，因此熵值更高。
- **中文分析：**实验测得的熵值（约 9.22）低于理论值（9.71）。这可能是因为个人采集的样本主要来自中文古典文学作品，没有充分覆盖的汉字范围（尤其是罕用字和生僻字）。理论值是基于一个更庞大、更多样化的语料库（如全量的文学、历史、科技文献）计算得出的，因此熵值更高。

5.3. 熵随文本规模的变化分析

为了更直观地展示熵的变化趋势，我们绘制了关系曲线图。

我们分别对中英文本在 200 万至 600 万字符的不同规模下进行了熵值计算，结果如下。

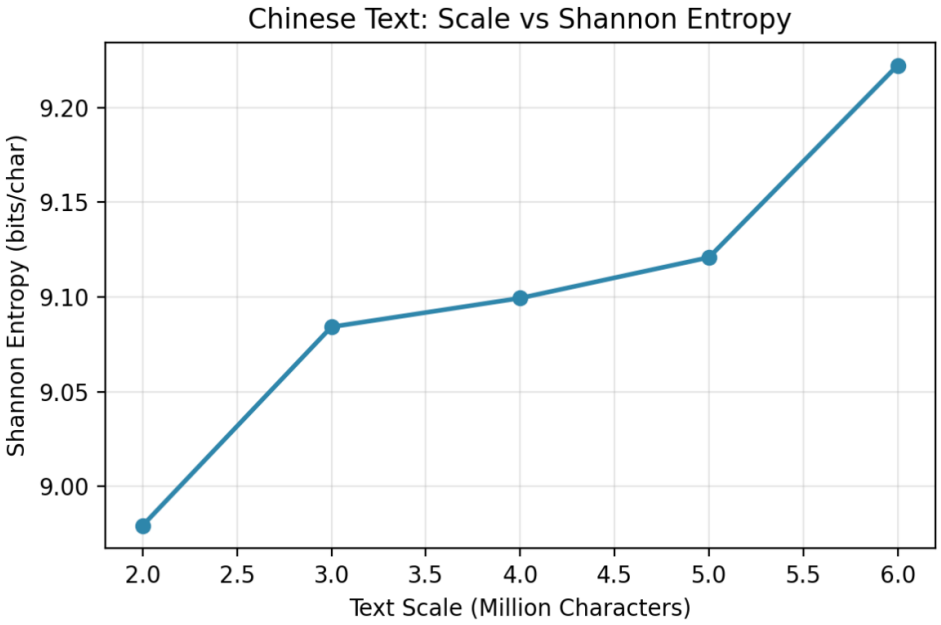

图 1 从 200 万至 600 万字符的不同规模下中文语料熵值统计结果

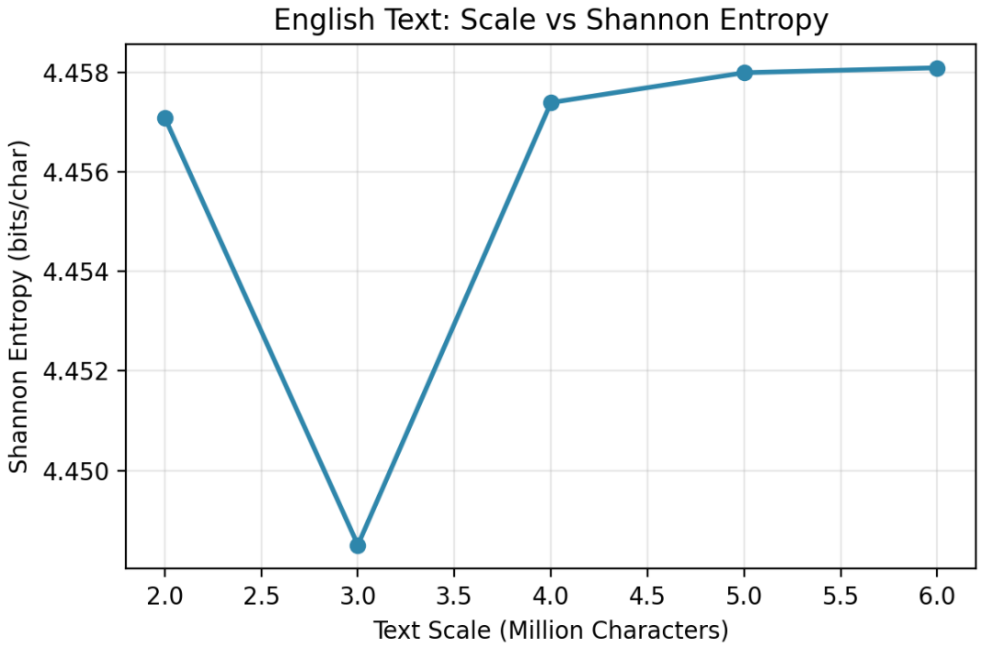

图 2 从 200 万至 600 万字符的不同规模下英文语料熵值统计结果

- **英文文本规模-熵关系曲线分析：**从图 1 可以看出，英文文本的熵值在文本规模从 200 万增加到 600 万的过程中，基本保持稳定，仅有微小的波动。这是因为英文的字符集非常有限（主要是几十个字母和常用符号）。当文本量达到一定规模后，几乎所有常用字符都已出现，并且它们的频率分布也趋于稳定，因此总熵值变化不大。
- **中文文本规模-熵关系曲线分析：**与英文不同，中文文本的熵值随着文本规模的扩大呈现出明显的、持续的增长趋势。这完美地揭示了汉字系统的复杂性。汉字字符集极其庞大，常用字数千，总数上万。在不断扩大的文本中，总能遇到新的、频率较低的汉字，这使得独立字符数持续增加，字符概率分布不断调整，从而导致整体熵值稳步上升。曲线并未出现平缓的迹象，说明即使在 600 万字符的规模下，语料库仍远未饱和。

6. 关于“伸缩空间”的讨论

本次实验虽然达到了预期目标，但在方法和深度上仍有很大的“伸缩空间”，即可以从以下几个方面进行改进和扩展：

- **样本的广度与均衡性：**结果的可靠性高度依赖于样本质量。当前样本主要来中文和英文古典文学作品，未来可以扩展采集范围，覆盖文学、科技、法律、社交媒体等更多领域，构建一个更具代表性的均衡语料库。
- **分析单元的深化：**本次实验以“字”为基本单位。分析可以向更宏观或更微观的层面扩展：
  a) **词熵：**以“词”为单位计算熵，可以更好地反映语言的语法和语义结构。例如，课件中提到英语单词的熵约为 10。
  b) **N-gram 模型：**计算二元组（bigram）或三元组（trigram）的条件熵，可以用来度量语言的上下文相关性和可预测性。
- **数据清洗的精细化：**可以设计更复杂的清洗逻辑，例如，识别并处理文本中的非目标语言（如中英混杂文本）、过滤网络噪声（如“小编说”、“点赞”）等，

7. 结论

本实验成功地通过编程实践，完成了对中英文本信息熵的计算与分析。实验结果清晰地表明：

- 汉字的字符熵（约 9.22 bits/char）远高于英文的字符熵（约 4.45 bits/char），这定量地证实了汉字系统在字符层面的复杂性远超英文。
- 随着文本规模的增加，英文熵趋于稳定，而中文熵持续增长，反映了两种语言字符集大小和使用频率分布的巨大差异。
- 实验结果与理论值的对比分析，不仅验证了熵理论，也揭示了样本选择和计算口径对结果的重要影响。

通过本次实验，我不仅掌握了信息熵的计算方法，更深入地理解了其作为衡量语言复杂性和不确定性的理论工具的强大作用。

附录

本项目所提到的英文语料爬取网址如下：

- https://www.gutenberg.org/files/2701/2701-h/2701-h.htm
- https://www.gutenberg.org/files/1342/1342-h/1342-h.htm
- https://www.gutenberg.org/files/11/11-h/11-h.htm
- https://www.gutenberg.org/files/84/84-h/84-h.htm
- https://www.gutenberg.org/files/98/98-h/98-h.htm
- https://www.gutenberg.org/cache/epub/2759/pg2759.txt
- https://www.gutenberg.org/cache/epub/1184/pg1184.txt
- https://www.gutenberg.org/cache/epub/1257/pg1257.txt

本项目所提到的中文语料爬取网址如下：

- https://www.gutenberg.org/cache/epub/24264/pg24264.txt",
- https://www.gutenberg.org/cache/epub/24185/pg24185.txt
- https://www.gutenberg.org/cache/epub/25146/pg25146.txt
- https://www.gutenberg.org/cache/epub/25606/pg25606.txt
- https://www.gutenberg.org/cache/epub/27582/pg27582.txt
- https://www.gutenberg.org/cache/epub/7367/pg7367.txt
- https://www.gutenberg.org/cache/epub/27686/pg27686.txt
- https://www.gutenberg.org/cache/epub/25142/pg25142.txt
- https://www.gutenberg.org/cache/epub/23910/pg23910.txt
- https://www.gutenberg.org/cache/epub/23962/pg23962.txt
- https://www.gutenberg.org/cache/epub/52200/pg52200.txt
- https://www.gutenberg.org/cache/epub/27582/pg27582.txt
- https://www.gutenberg.org/cache/epub/25393/pg25393.txt
- https://www.gutenberg.org/cache/epub/25327/pg25327.txt
- https://www.gutenberg.org/cache/epub/27166/pg27166.txt
- https://www.gutenberg.org/cache/epub/25134/pg25134.txt