

自 然 语 言 处 理

技术报告

班级：2025秋季自然语言处理班

姓名：

学号：

报告主题：计算中英文信息源的熵

2025年11月6日

一．报告摘要

本项目实现了一个完整的多语言语料库采集与分析系统，包含三个核心模块：

* 异步网络爬虫 (async\_corpus\_crawler.py)：负责从指定网站爬取中英文语料
* 语料合并工具 (merge\_corpus.py)：将爬取的零散文本文件合并为指定大小的语料库
* 语料分析系统 (analyse\_corpus.py)：对语料进行熵值计算、Zipf定律验证等语言学分析

针对中文熵的计算，选取了以下网站：



图 1 中文网站

针对英文熵的计算，选取了以下网站：



图 2 英文网站

二．网络爬虫模块

本项目采用基于异步并发架构的BFS（广度优先搜索）策略实现网络爬虫模块，与传统的同步爬虫相比具有显著的性能优势。



图 3 爬虫核心代码

爬虫核心使用aiohttp库进行异步HTTP请求，配合asyncio实现高并发处理，能够同时处理多个网页请求而不会阻塞程序执行。通过ClientSession和连接池管理，有效复用网络连接，大幅提升爬取效率。

在网页遍历策略上，爬虫采用经典的广度优先搜索算法。从初始URL列表开始，首先爬取这些种子页面，然后解析每个页面中所有的内部链接，将其加入待访问队列。这种层级式扩展确保优先爬取重要页面，避免陷入过深的链接路径。BFS策略能够快速覆盖网站的广泛内容区域，在相同时间内获取更多样化的文本数据。

为了平衡爬取效率与资源消耗，系统实现了多重控制机制。通过max\_concurrent参数限制并发请求数量，防止对目标服务器造成过大压力；利用delay参数设置请求间隔，体现良好的网络礼仪；max\_pages参数确保爬取过程在可控范围内终止。在链接管理方面，采用visited\_urls集合进行网址去重，结合urls\_to\_visit双端队列实现高效的URL调度，确保每个页面只被爬取一次，避免重复劳动和资源浪费。

针对不同网站结构的自适应处理，爬虫实现了智能的内容过滤机制。通过分析网页的base\_domain，确保只爬取同一域名下的内部链接，避免爬虫失控地追随外部链接。在解析HTML时，优先提取<p>段落标签中的文本内容，同时主动移除script、style、nav等非内容性标签，显著提升语料质量。这种针对性的内容提取策略特别适合新闻网站、百科类站点等以文本内容为主的源站，能够高效获取高质量的语言材料。



图 4 内容过滤

此外，爬虫模块还集成了完善的伦理合规特性，严格遵守robots.txt协议，通过RobotFileParser自动检测并尊重网站的爬取规则。编码自适应系统能够智能识别网页字符编码，支持UTF-8、GBK、GB18030等多种中文编码格式，确保中文文本的正确解析。这些优化措施使得本爬虫不仅在性能上优于传统同步爬虫，在稳定性、合规性和数据质量方面也表现出色，为语料库建设提供了可靠的技术保障。

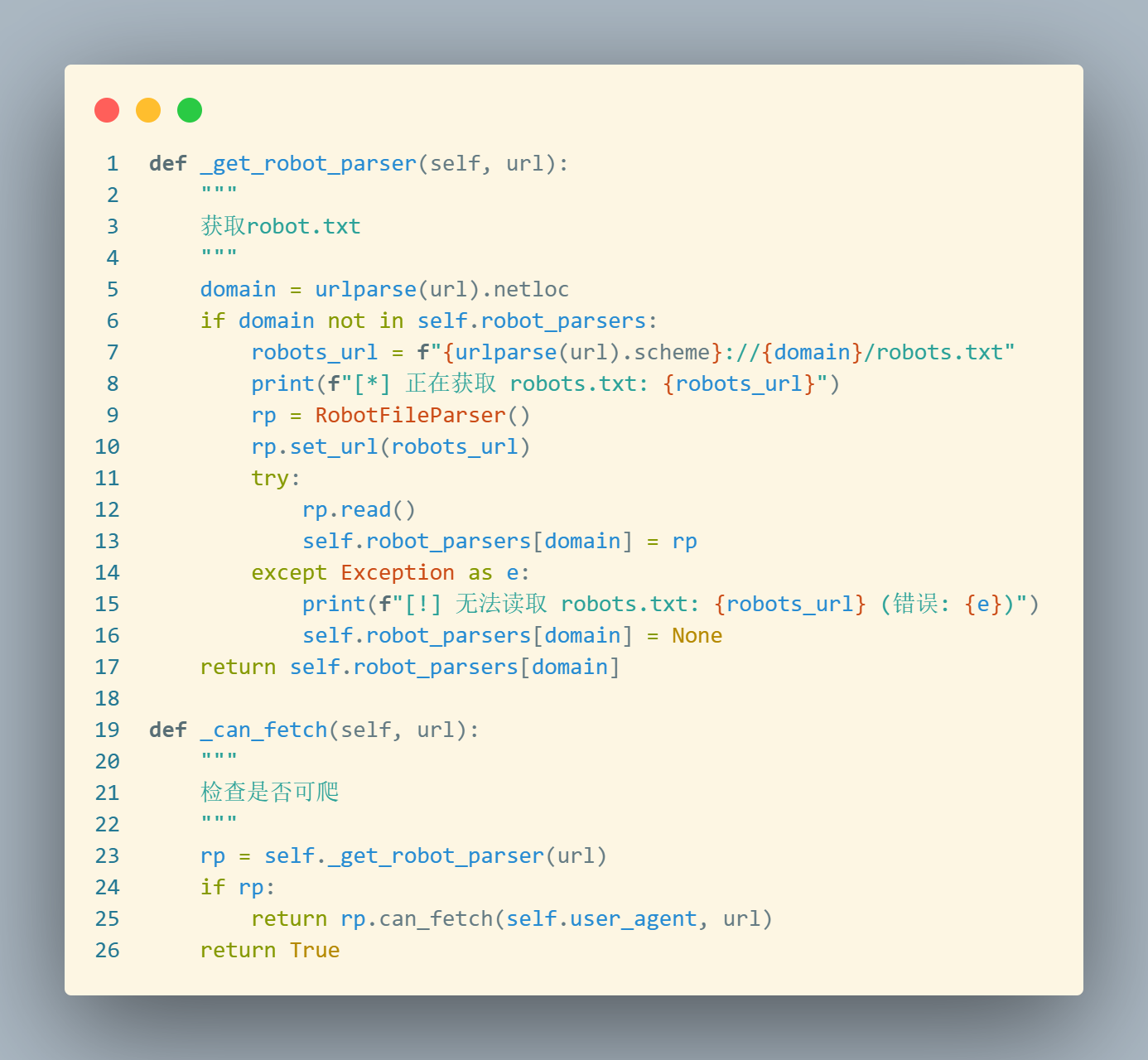


图 5 协议遵守



图 6 多样编码

三．数据清洗

将所有爬虫数据保存至txt文件中，随后进行数据清洗。

针对中文文本，利用正则表达式，结合汉字的GTF-8编码范围，去除除了汉字之外的所有字符，并写回文本文件：



图 7 中文语料清洗

针对英文文本，利用正则表达式，去除除了英文字母以及空格之外的所有字符，并写回文本文件：



图 8 英文语料清洗

四．字母/汉字出现特征统计及熵的计算

针对中文语料，我们依次分析1MB、2MB、……、7MB的中文熵，篇幅有限，仅展示1MB和7MB的分析结果：



图 9 1MB中文语料分析图

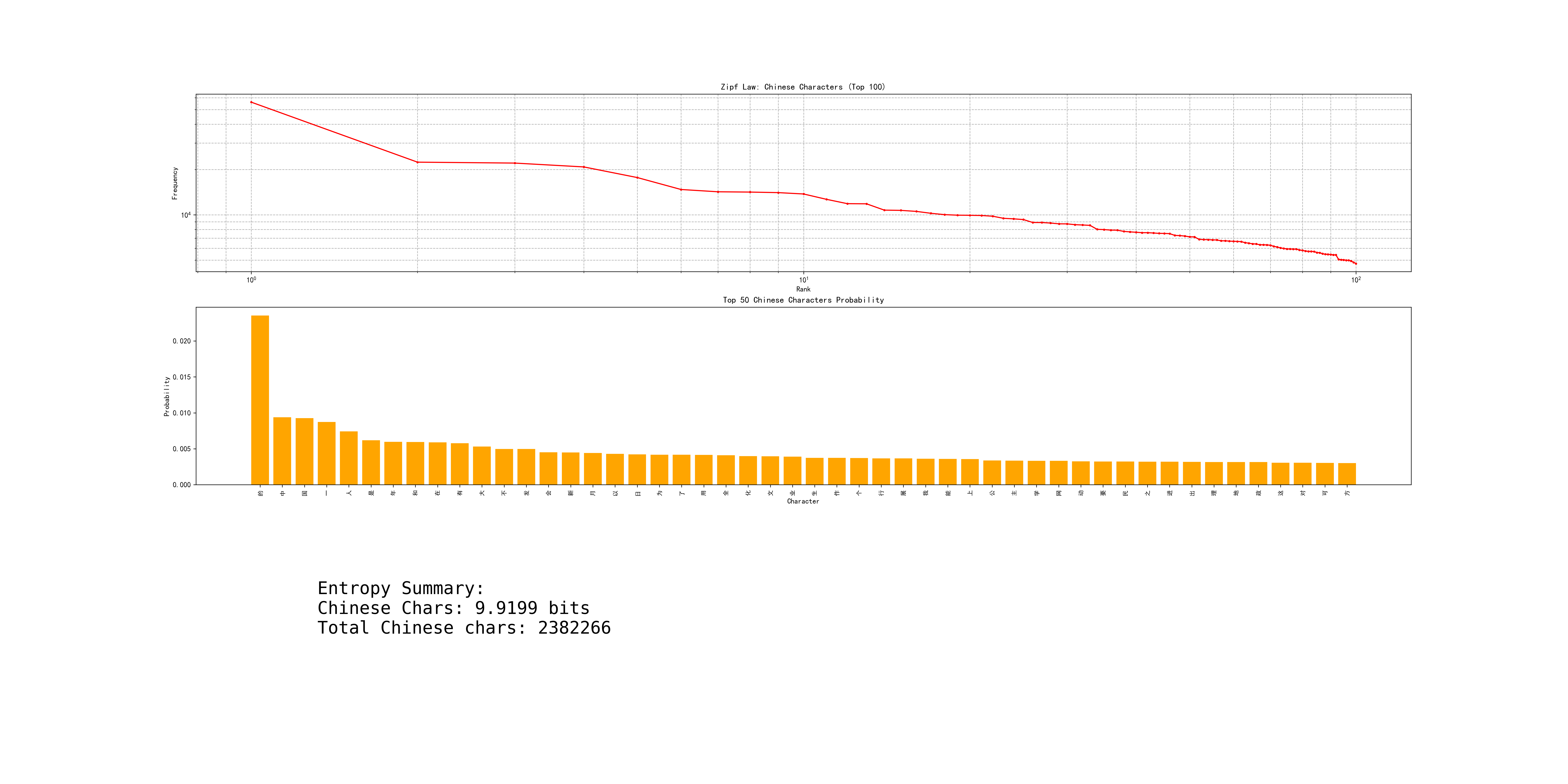


图 10 7MB中文语料分析图

针对英文，我们依次分析了2MB、4MB、……、14MB的英文熵，篇幅有限，仅展示2MB和14MB的分析结果：

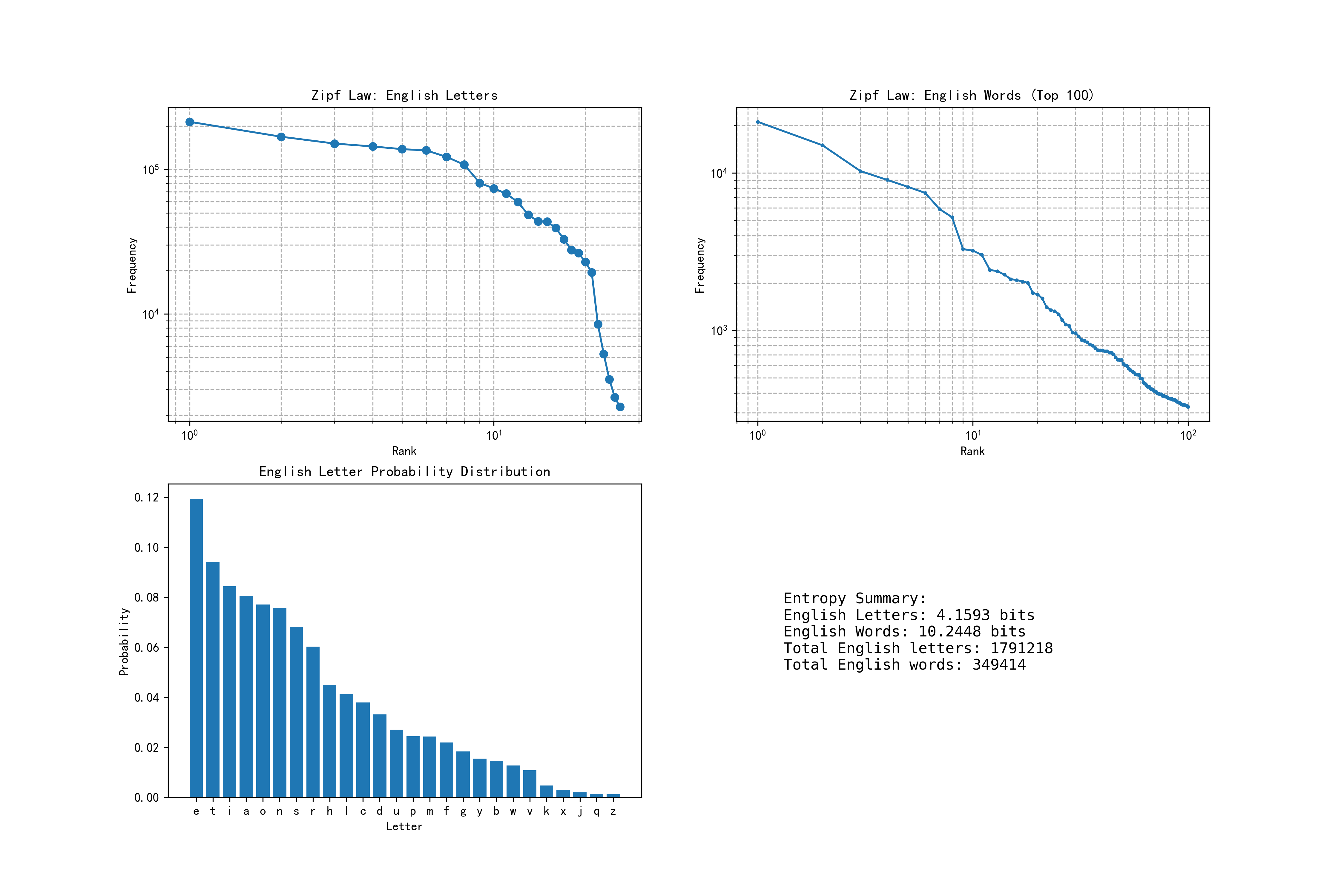


图 11 2MB英文语料分析图

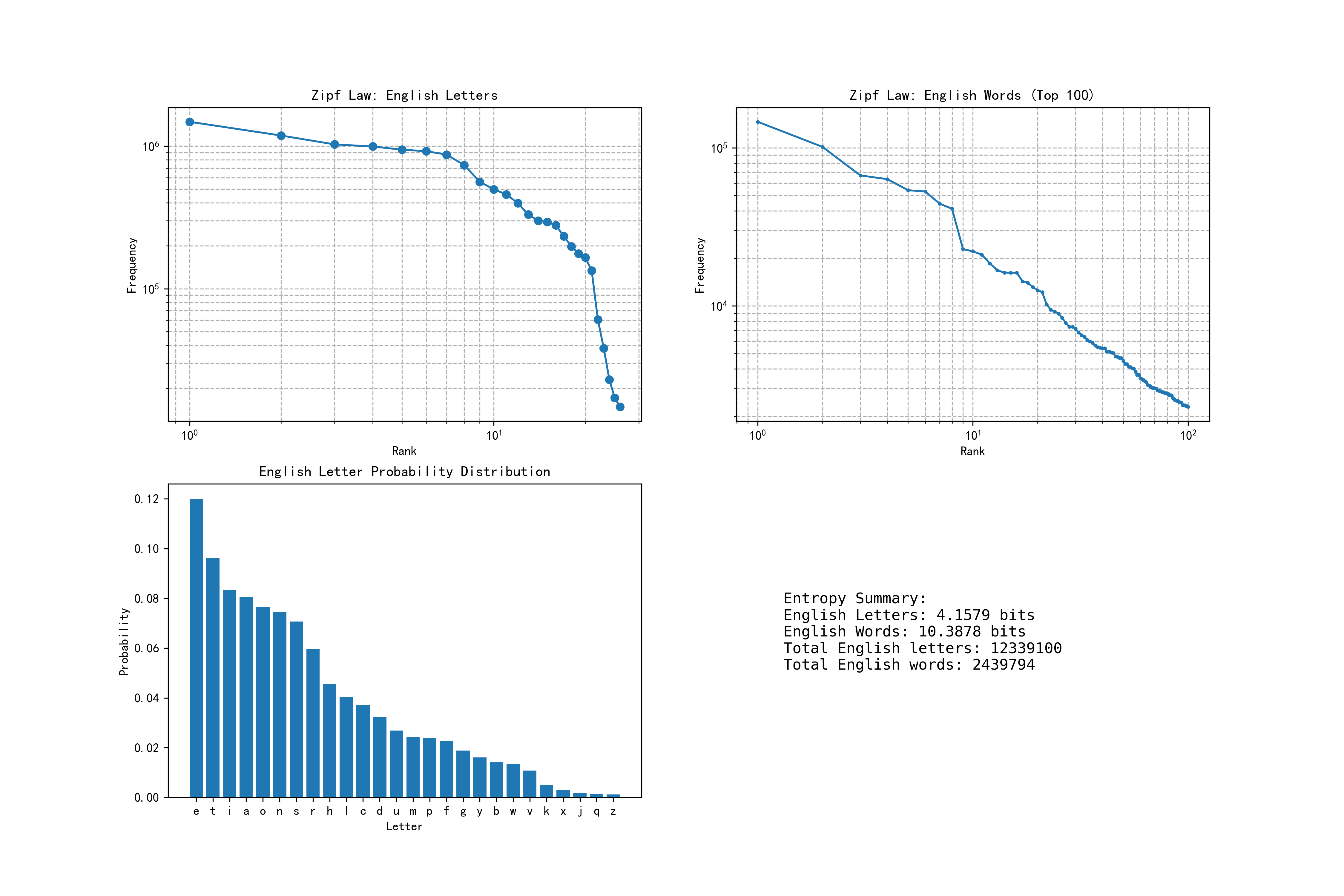


图 12 14MB英文语料分析图

在扩大语料库的过程计算熵并观察熵的变化。趋势如下所示：

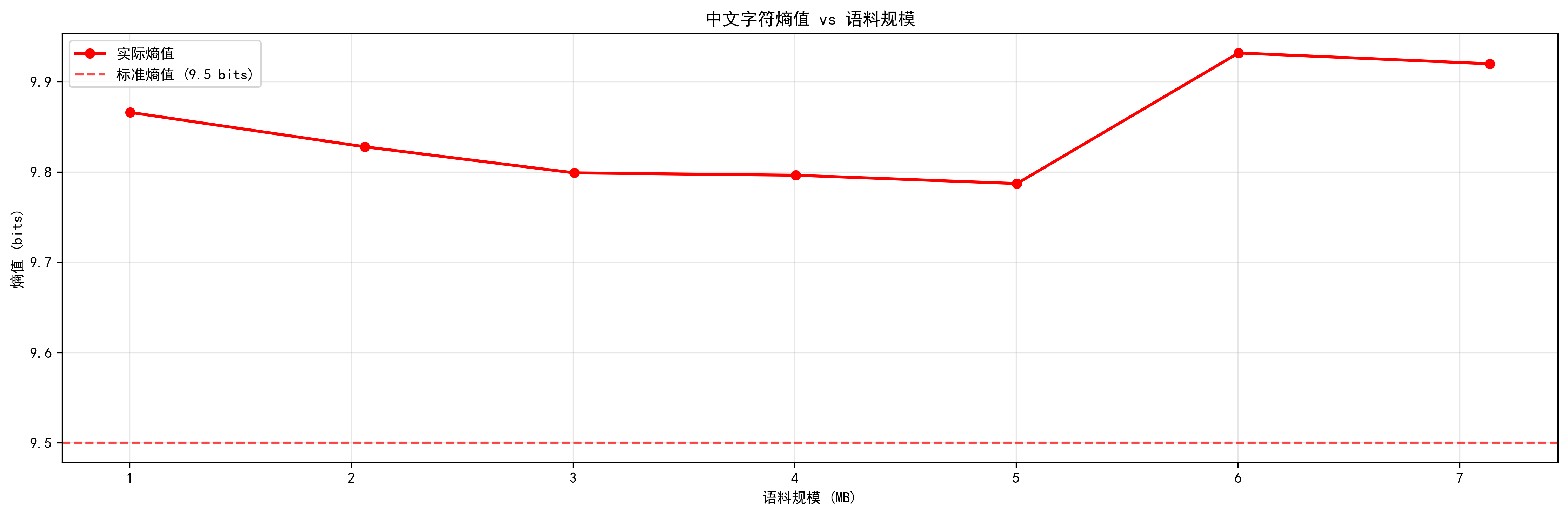


图 13 中文字符熵随语料规模的变化图

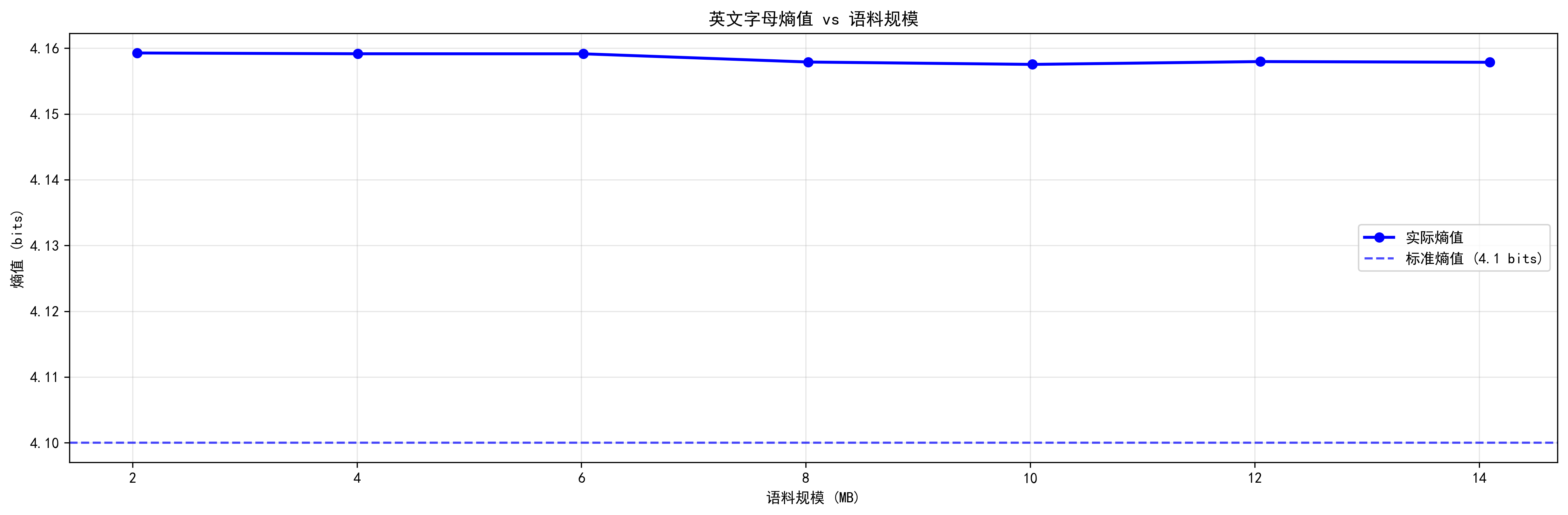


图 14 英文字母熵随语料规模的变化图

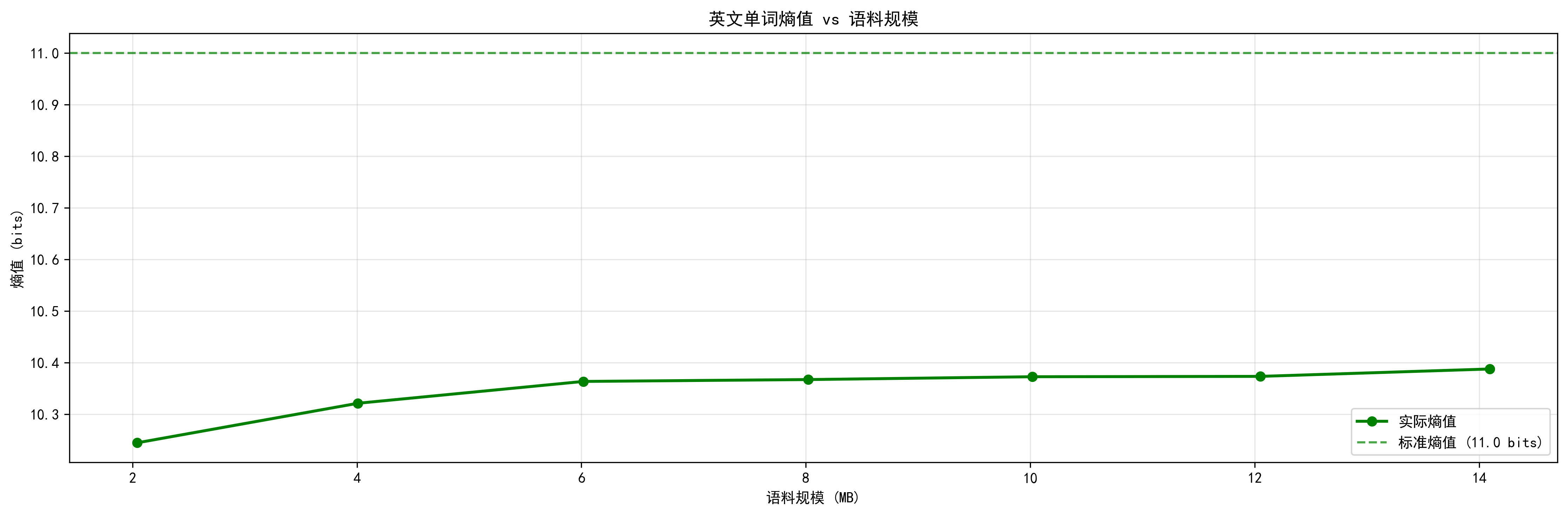


图 15 英文单词熵随语料规模的变化图

图9、图10、图11、图12显示**中文字符和英文单词满足齐夫定律**，其对数频率分布和频率排名近似为一条直线。其中中文字符的top10字符（的、中、国、一、人、是、年、和、在、有）与PPT中的top10字符（的、了、在、是、和、一、这、有、他、我）存在一定差异。究其原因是本文的中文语料的主要来源是各大新闻网站，尤其是时政类网站较多，并且在爬取过程中存在语料清洗不充分的情况，如在爬取人民网网站时有大量重复的“人 民 网 版 权 所 有 未 经 书 面 授 权 禁 止 使 用”等无效语料。

图13显示中文字符熵值一直高于标准值，检查中文语料文件发现其中存在一定数量的繁体字，使得整个概率分布变得更平坦，从而导致了更高的熵值。

图14显示英文字母熵值一直很稳定，且与标准熵值差距很小，这表明在几MB规模的语料下，已经见过了所有的字母，并且它们的频率分布已经基本稳定。

图15显示英文单词熵显示与标准熵值仍有差距，但是差距在不断减小，这表明14MB的语料太小，词汇量远未饱和。

在最大的语料库下，中文字符熵值、英文字母熵值、英文单词熵值分别约为9.92bit、4.16bit、10.39bit，**与标准熵值差距不大**，符合预期。