5. 习题

2-1. 作业 分别收集尽量多的英语和汉语文本，编写程序计算这些文本中英语字母和汉字的熵，对比本章课件第12页上表中给出的结果。然后逐步扩大文本规模，如每次增加2M，重新计算文本规模扩大之后的熵，分析多次增加之后熵的变化情况。

要求：

① 利用爬虫工具从互联网上收集样本，并对样本进行处理，如清洗乱码等；

② 设计算法并编程实现在收集样本上字母/汉字的概率和熵的计算；

③ 当改变样本规模时，重新计算字母/汉字的概率和熵的计算，并分析计算结果；

④ 完成一份技术报告，在报告中写明利用什么爬虫工具从哪些网站上收集的样本，如何进行的样本清洗，清洗后样本的规模，在不同样本规模下计算的结果等。