作业 1 计算中文平均信息熵

一．实验目的

阅读文献，参考文章计算中文语料的平均信息熵。

二．实验原理

1. 熵

熵，泛指某些物质系统状态的一种量度，某些物质系统状态可能出现的程度。亦被社会科学用以借喻人类社会某些状态的程度。熵的概念是由德国物理学家克劳修斯于 1865 年所提出。最初是用来描述“能量退化”的物质状态参数之一，在热力学中有广泛的应用。但那时熵仅仅是一个可以通过热量改变来测定的物理量，其本质仍没有很好的解释，直到统计物理、信息论等一系列科学理论发展，熵的本质才逐渐被解释清楚，即，熵的本质是一个系统“内在的混乱程度”。它在控制论、概率论、数论、天体物理、生命科学等领域都有重要应用，在不同的学科中也有引申出的更为具体的定义，按照数理思维从本质上说，这些具体的引申定义都是相互统一的，熵在这些领域都是十分重要的参量。

2. 信息熵

信息熵的定义公式：

$$H(X) = -\sum p(x)\, log\, p(x)$$

并且规定：

$$0\, log\, 0 = 0$$

信息熵的三个性质：

（1）单调性，发生概率越高的事件，其携带的信息量越低；

（2）非负性，信息熵可以看作为一种广度量，非负性是一种合理的必然；

（1）累加性，即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件不确定性的量度的和，这也是广度量的一种体现。

3. N-Gram 语言模型

（1）一元模型信息熵计算公式为：

$$H(X) = -\sum_{x \in X} p(x)\, log\, p(x)$$

其中$p(x)$为每个词在语料库中出现的频率，可以看作单个词的信息熵。

（2）二元模型的信息熵公式为：

$$H(X|Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x|y)$$

其中联合概率 P(x,y)可近似等于每个二元词组在语料库中出现的频率，条件概率 P(x|y)可近似等于每个二元词组在语料库中出现的频率与以该二元词组的第一个词为词首的二元词组的频数的比值。

（3）三元模型的信息熵公式为：

$$H(X|Y,Z) = -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log p(x|y,z)$$

三．实验内容

1．数据预处理

本次给出的中文语料库为 16 篇金庸小说，首先读取文章内容，删除标点符号，将处理后的文本生成文件。

2．计算 N-Gram 模型下的中文信息熵

读取处理后的文本文件，使用 jieba 分词系统对读取到的文本进行分词，没有删除掉停用词表中包含的词。计算每个词出现的频数和频率，以此计算每个词的信息熵，最后对所有词的信息熵求和得到结果。其中二元模型和三元模型需要考虑上下文关系并使用条件熵计算。

四．实验结果

	总词数	不同词词数	信息熵
一元词	4313893	172477	12.1645
二元词	4254577	1949753	6.9463
三元词	4195510	3481501	2.3041