一、问题描述

阅读文章 *Entropy of English PeterBrown*，参考该文章，分别以字为单位和以词为单位，计算金庸小说全集的平均信息熵。

二、实验原理

信息熵的概念最早由香农（1916-2001）于 1948 年借鉴热力学中的“热熵”的概念提出，旨在表示信息的不确定性。熵值越大，则信息的不确定程度越大。其数学公式可以表示为：

$$H(x)=\sum_{x\in X}P(x)\log\frac{1}{P(x)}=-\sum_{x\in X}P(x)\log P(x)$$

其中，$P(x)$可近似等于每个词在语料库中出现的频率。

对于联合分布的随机变量$(X,Y)\sim P(X,Y)$，在两变量相互独立的情况下，其联合信息熵为：

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

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

该联合信息熵可以用于二元模型与三元模型的计算。

二元模型的信息熵计算公式为：

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

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数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。

三元模型的信息熵计算公式为：

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

其中，联合概率$P(x,y,z)$可近似等于每个三元词组在语料库中出现的频率，条件概率$P(x|y,z)$可近似等于每个三元词组在语料库中出现的频数与以该三元词组的前两个词为词首的三元词组的频数的比值。

三、实验过程与结果

1. 文本预处理

本实验采用的文本数据中有大量非中文字符，会影响对中文信息熵的计算。包括：小说下载站的广告“本书来自 www.cr173.com 免费 txt 小说下载站”、“更多更新免费电子书请关注 www.cr173.com”，非必要的换行“\n”，英文空格“ ”，中文空格“\u3000”，特殊符号“~!@#$%^&*()_+`-=”，英文字母和数字等。

为了更准确地计算中文平均信息熵，需要对文本进行预处理，将上述无用字符去除。之后得到的纯中文文本作为语料库。分别以单个汉字为单位，或以 jieba 精确模式划分得到的分次为单位，统计词频，进而计算平均信息熵。

2. 实验结果

	汉字个数	平均信息熵(比特/字)			分词个数	平均词长	平均信息熵(比特/词)		
		1-Gram	2-Gram	3-Gram			1-Gram	2-Gram	3-Gram
白马啸西风	59277	8.868	4.533	1.635	37000	1.602	10.065	3.975	0.879
碧血剑	416273	9.457	5.865	2.350	242596	1.716	11.747	4.997	0.993
飞狐外传	375434	9.312	5.753	2.405	221146	1.698	11.528	4.996	1.054
连城诀	194489	9.172	5.398	2.157	117203	1.659	11.042	4.692	0.955
鹿鼎记	1023944	9.294	5.987	2.949	604733	1.693	11.446	5.753	1.621
三十三剑客图	53293	9.670	4.834	0.982	31183	1.709	11.683	2.939	0.272
射雕英雄传	772422	9.443	6.057	2.758	456216	1.693	11.798	5.473	1.291
神雕侠侣	827645	9.351	6.016	2.836	494787	1.673	11.584	5.493	1.453
书剑恩仇录	435692	9.462	5.782	2.393	253170	1.721	11.708	5.035	1.040
天龙八部	1021156	9.405	6.125	2.949	604072	1.690	11.734	5.678	1.481
侠客行	309717	9.153	5.591	2.369	183198	1.691	11.188	4.958	1.127
笑傲江湖	824586	9.207	5.897	2.871	482233	1.710	11.398	5.620	1.527
雪山飞狐	117148	9.171	5.119	1.773	71045	1.649	10.913	4.135	0.844
倚天屠龙记	818274	9.395	6.020	2.805	474831	1.723	11.760	5.518	1.328
鸳鸯刀	30432	8.978	4.183	1.151	18329	1.660	10.240	3.155	0.586
越女剑	13649	8.823	3.642	0.911	8047	1.696	10.069	2.529	0.327
所有文件	7293431	9.536	6.723	3.946	4299789	1.696	12.165	6.941	2.313

四、结果分析

1. 以词为单位的平均信息熵除以词长，再与以字为单位的信息熵进行比较，可以看出按字进行计算得到的信息熵更大。即单个汉字比词语有更大的不确定性。

2. 以字为单位和以词为单位时，都可以发现，从一元模型到二元模型到三元模型，信息熵的数值在减小。即随着关联程度的增加，信息的不确定性减小。汉语中常用词组的存在以及作者本人的语言习惯都会造成这种现象。

五、参考资料

[1] 深度学习与自然语言处理实验——中文信息熵的计算

https://blog.csdn.net/weixin_42663984/article/details/115718241

[2] 深度学习与自然语言处理第一次作业——中文平均信息熵的计算

https://blog.csdn.net/weixin_50891266/article/details/115723958

[3] 中文信息熵的计算

https://blog.csdn.net/qq_37098526/article/details/88633403