北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理（NLP）第一次课后作业

院（系）名称	自动化科学与电气工程学院
专业名称	自动化
学号	ZY2303808
学生姓名	牛晨然

2024 年 04 月

1 内容介绍

齐夫定律（英语：Zipf's law）是由哈佛大学的语言学家乔治·金斯利·齐夫（George Kingsley Zipf）于1949年发表的实验定律。它可以表述为：在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。所以，频率最高的单词出现的频率大约是出现频率第二位的单词的2倍，而出现频率第二位的单词则是出现频率第四位的单词的2倍。这个定律被作为任何与幂定律概率分布有关的事物的参考。

信息是个很抽象的概念，人们常说信息很多，或者信息很少，但却很难说清楚信息到底有多少。比如一本书到底有多少信息量。信息论之父Claude Elwood Shannon第一次用数学语言阐明了概率和信息冗余度的关系。Shannon指出，任何信息都存在冗余，冗余大小与信息中每个符号的出现概率或者说不确定性有关。Shannon借鉴了热力学的概念，把信息中排除了冗余后的平均信息量称为“信息熵”，并给出了计算信息熵的数学表达式。在自然语言处理中，信息熵只反映内容的随机性（不确定性）和编码情况，与内容本身无关。信息熵越大，单个词提供的信息量也就越大，不确定性也就越大。通过计算信息熵，能够衡量词表意的精确程度，信息熵越小，表意越精确。

本文首先对金庸的10篇小说进行词频统计并验证齐夫定律，接着按照一元、二元、三元语言模型分别计算了字/词的信息熵，最后对实验结果进行了比较分析。

2 实验原理

第一部分：验证齐夫定律

首先对金庸的小说进行预处理，删除所有的隐藏独好、非中文符号和标点符号；然后利用 jieba 库进行文本分词处理；接着统计单词出现的频率，并按照频率从高到低进行排序；最后绘制频率-排名的对数图，观察结果是否符合 Zipf's Law 的预期关系。

第二部分：计算信息熵

1948 年，香农提出了“信息熵”的概念，解决了对信息的量化度量问题。一条信息的信息量大小和它的不确定性有直接的关系。比如说，我们要搞清楚一件非常非常不确定的事，或是我们一无所知的事情，就需要了解大量的信息。

通常，一个信源发送出什么符号是不确定的，衡量它可以根据其出现的概率来度量。概率大，出现机会多，不确定性小；反之不确定性就大。不确定性函数f是概率P的减函数；两个独立符号所产生的不确定性应等于各自不确定性之和，即：

$$f(P1,P2) = f(P1) + f(P2)$$

这称为可加性。同时满足这两个条件的函数f是对数函数，

即:

$$f(P1) = log\frac{1}{p} = -logp$$

在信源中，考虑的不是某一单个符号发生的不确定性，而是要考虑这个信源所有可能发生情况的平均不确定性。若信源符号有n种取值：$U1 \dots Ui \dots Un$，对应概率为：$P1 \dots Pi \dots Pn$，且各种符号的出现彼此独立。这时，信源的平均不确定性应当为单个符号不确定性$-logPi$的统计平均值E，可称为信息熵。即

$$H(U) = E[-log\ p_i] = -\sum_{i=1}^{n} p_i \log p_i$$

信息熵的具体计算公式如下：

一元模型信息熵：

如果统计量足够大，字、词、二元词组或三元词组出现的概率大致等于其出现的频率。由此可得，字和词的信息熵计算公式为：

$$H(X) = -\sum_{x\in X} P(\mathrm{x})\log_2 P(\mathrm{x})$$

其中，$P(\mathrm{x})$可近似等于每个字或词在语料库中出现的频率。

二元模型信息熵：

$$H(X|Y) = -\sum_{x\in X, y\in Y} P(\mathrm{x},\mathrm{y})\log_2 P(\mathrm{x}|\mathrm{y})$$

其中，联合概率$P(\mathrm{x},\mathrm{y})$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$P(\mathrm{x}|\mathrm{y})$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。

三元模型信息熵：

$$H(X|Y,Z) = -\sum_{x\in X, y\in Y, z\in Z} P(\mathrm{x,y,z})\log_2 P(\mathrm{x|y,z})$$

其中，联合概率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。

3 实验结果与分析

第一部分：验证齐夫定律

挑选了金庸的 10 部作品，进行词频统计后的结果如图 1 所示。

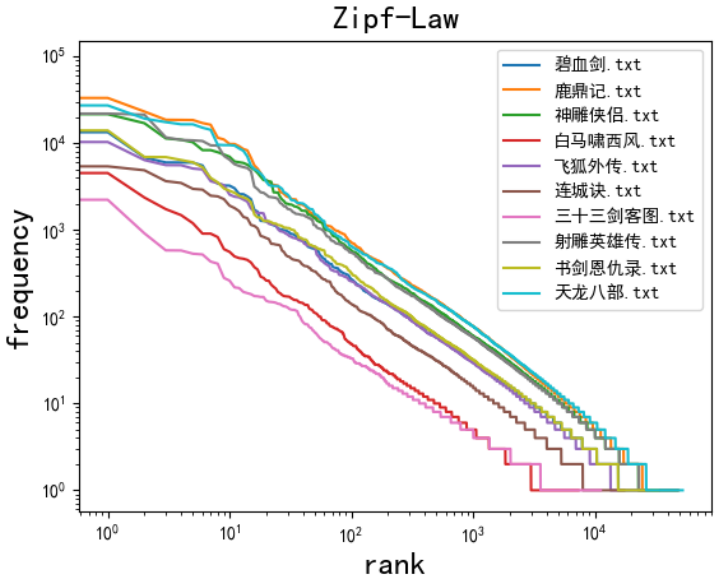

图 1 词频统计结果

可以看出，各个文件的词频曲线都近似于一条直线，由此可以验证齐夫定律。

第二部分：计算信息熵

对金庸的 10 部作品，分别计算一元、二元、三元的字/词信息熵，结果如表 1 和图 2 所示。

表 1 信息熵计算结果

文件名	一元信息熵	二元信息熵	三元信息熵
碧血剑	11.73664	5.01008	0.9961
鹿鼎记	11.65089	5.67041	1.41739

神雕侠侣	11.73136	5.53525	1.35125
白马啸西风	10.29235	4.02103	0.73543
飞狐外传	11.52548	5.00146	1.05597
连城决	11.03699	4.70805	0.9582
三十三剑客图	11.68292	2.94066	0.27189
射雕英雄传	11.83668	5.49114	1.26965
书剑恩仇录	11.71253	5.03519	1.03967
天龙八部	11.73321	5.68599	1.48729

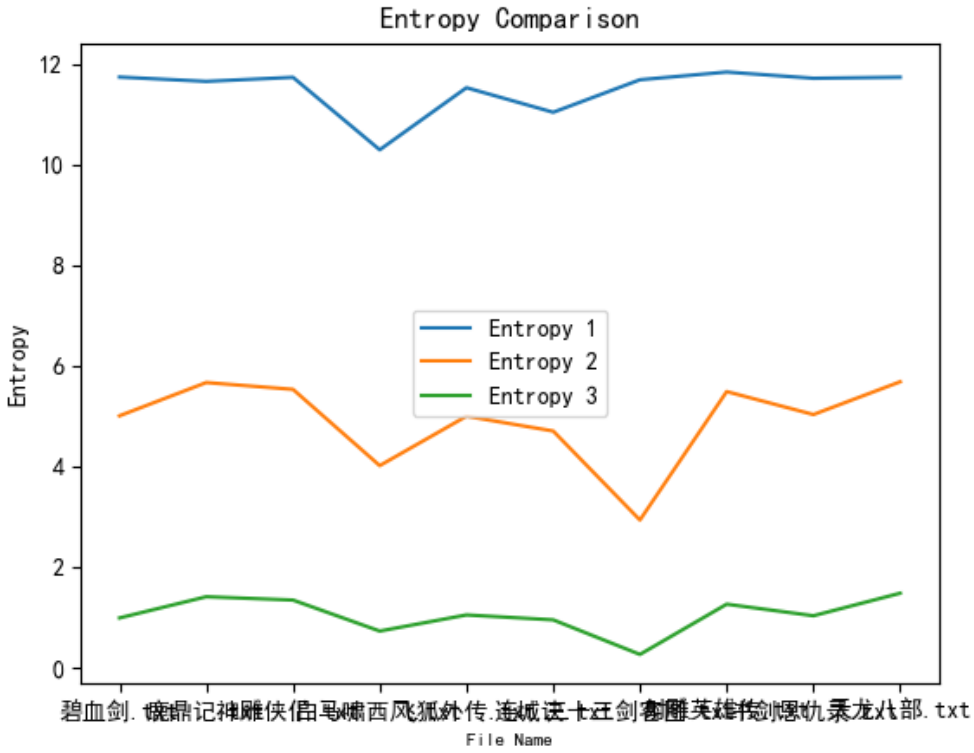

图 2 作品信息熵对比

由图 2 可以看出，无论是一元、二元还是三元的语言模型，字/词的信息熵在每个作品间的变化趋势是相同的，可以说明金庸不同作品的语言风格是相似的。同时，在同一部作品中，一元信息熵大于二元信息熵大于三元信息熵，在估计时考虑的词数越多，则上下文之间的联系越多，不同词组合出现的种类个数也会越多，则文本的信息熵则越小。之所以出现这样的情况，是因为元数越大，通过分词后得到的文本中词组的分布就越简单，元数越大使得固定的词数量越多，固定的词能减少由字或者短词打乱文章的机会，使得文章变得更加有序，减少了由字组成词和组成句的不确定性，也即减少了文本的信息熵。

4 参考文献

[1] Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. Comput. Linguist. 18, 1 (March 1992), 31–40.

[2] https://blog.csdn.net/weixin_42663984/article/details/115718241

[3] https://blog.csdn.net/GWH_98/article/details/117001985