Report of Deep Learning for Natural Langauge Processing

种雨萌

2775965921@qq.com

Abstract

本篇报告对金庸的 16 篇小说进行了分析，首先验证了 Zipf’s Law，然后分别以字和词为单位，计算了每篇小说的平均信息熵。

Introduction

Zipf's law（齐夫定律）是美国学者 G.K.齐普夫于 20 世纪 40 年代提出的词频分布定律，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。如果把一篇较长文章中每个词出现的频次统计起来，按照高频词在前、低频词在后的递减顺序排列，并用自然数给这些词编上等级序号，即频次最高的词等级为 1，频次次之的等级为 2，以此类推；若用 f 表示频次，r 表示等级序号，则有 f×r=C(C 为常数)。

信息熵是信息论的基本概念，用于描述信息源各可能事件发生的不确定性。20 世纪 40 年代，香农（C.E.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”。在自然语言处理中，信息熵只反映内容的不确定性和编码情况，与内容本身无关。通过计算信息熵，能够衡量词表意的精确程度，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；信息熵越小，单个词提供的信息量也就越小，表意也就越精确。

本篇报告对金庸的 16 篇小说进行了分析，首先验证了 Zipf’s Law，然后分别以字和词为单位，计算了每篇小说的平均信息熵。

Methodology

1-1: 验证 Zipf’s Law

要验证 Zipf’s Law，需要对指定的文本文件进行文本处理和分词，统计词频并绘制词频分布图。具体步骤如下：首先对文本进行预处理，包括删除隐藏符号、非中文字符、停用词和标点符号；然后读取每个文件的内容，进行预处理和分词，并统计词频，将词频数据写入到指定文件中；最后将词频数据按照频次从大到小

排序，绘制词频分布曲线。

在验证 Zipf's Law 的过程中，使用到了 jieba 库。Jieba 库是一款用于中文分词的第三方库，由于中文文本之间每个汉字都是连续书写的，需要通过特定的手段——也就是分词，来获得其中的每个词组。这一过程可以通过 jieba 库来完成。Jieba 库支持精确、全模式、搜索引擎三种分词模式，本次实验中对预处理后的文本使用 jieba 精确模式，它能够更好地控制切词的准确性，适用于大多数文本分析任务。

1-2：计算平均信息熵

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：$x_1, x_2, \cdots x_n$，对应概率为：$P_1, P_2, \cdots P_n$ ，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的统计平均值，称为信息熵，即：

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

式中对数一般取 2 为底，单位为比特。

对于文本信息来说，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：

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

其中，$P(x)$可近似等于每个字或词在语料库中出现的频率。

二元模型的信息熵计算公式为：

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

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：

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

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。

Experimental Studies

1-1: 验证 Zipf's Law

对金庸的 16 部作品分别进行分词统计后，得到的实验结果如下：

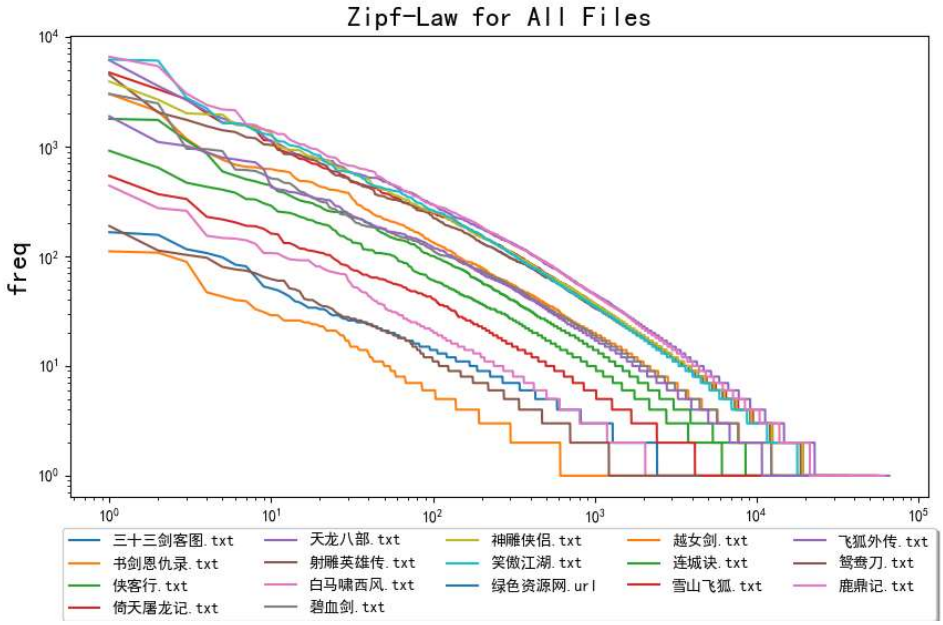

图 1 金庸 16 部作品词频统计结果

1-2: 计算平均信息熵

计算以上 16 部作品基于字/词的一元、二元、三元模型的信息熵，得到的实验结果如下：

表 1 金庸作品信息熵统计结果

语料库名称	字数	信息熵（比特/词）					
		一元字	二元字	三元字	一元词	二元词	三元词
白马啸西风	34177	9.2280	4.0872	1.2105	11.1961	2.8786	0.3540
碧血剑	262028	9.7560	5.6750	1.7951	12.8856	3.9617	0.4307
飞狐外传	233447	9.6300	5.5691	1.8650	12.6259	4.0404	0.4609
连城诀	116752	9.5152	5.0902	1.6390	12.2066	3.5890	0.3685

鹿鼎记	618267	9.6588	6.0199	2.4095	12.6394	4.9926	0.8343
三十三剑客图	34813	10.0111	4.2817	0.6499	12.5349	1.8089	0.0912
射雕英雄传	485336	9.7416	5.9699	2.1993	13.0363	4.6004	0.5340
神雕侠侣	514317	9.6633	6.0023	2.2826	12.7608	4.6871	0.6265
书剑恩仇录	278667	9.7456	5.6040	1.8651	12.7153	4.1457	0.4986
天龙八部	624629	9.7812	6.1153	2.3520	13.0175	4.8397	0.6635
侠客行	186334	9.4366	5.3793	1.8191	12.2884	3.9922	0.5126
笑傲江湖	497888	9.5162	5.8564	2.3612	12.5241	4.8385	0.7954
雪山飞狐	69790	9.5045	4.7998	1.3021	12.0581	3.0647	0.2905
倚天屠龙记	511460	9.7071	5.9828	2.2759	12.8941	4.6849	0.6425
鸳鸯刀	18688	9.2210	3.6512	0.8937	11.1413	2.1429	0.2321
越女剑	8924	8.8051	3.0955	0.8371	10.5114	1.7286	0.2324

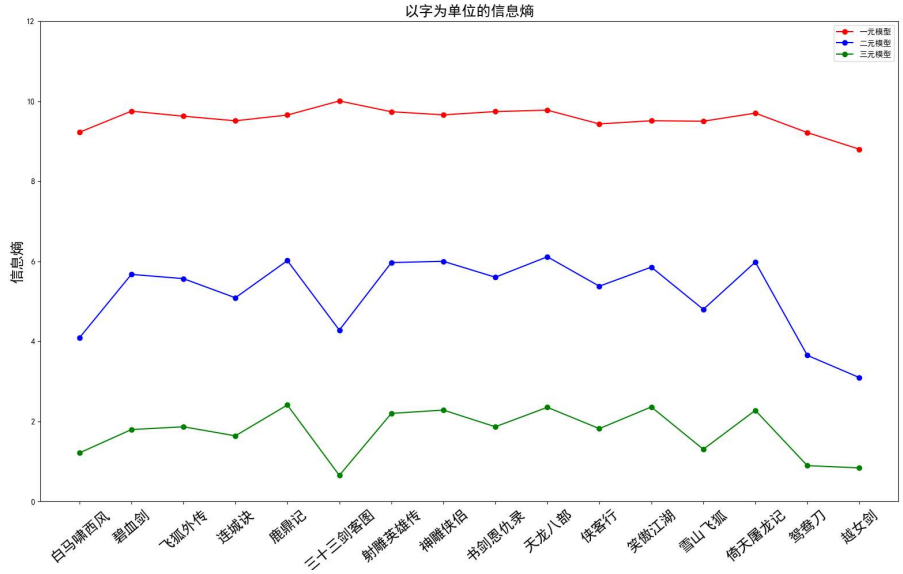

图 2 金庸 16 部作品以字为单位的平均信息熵

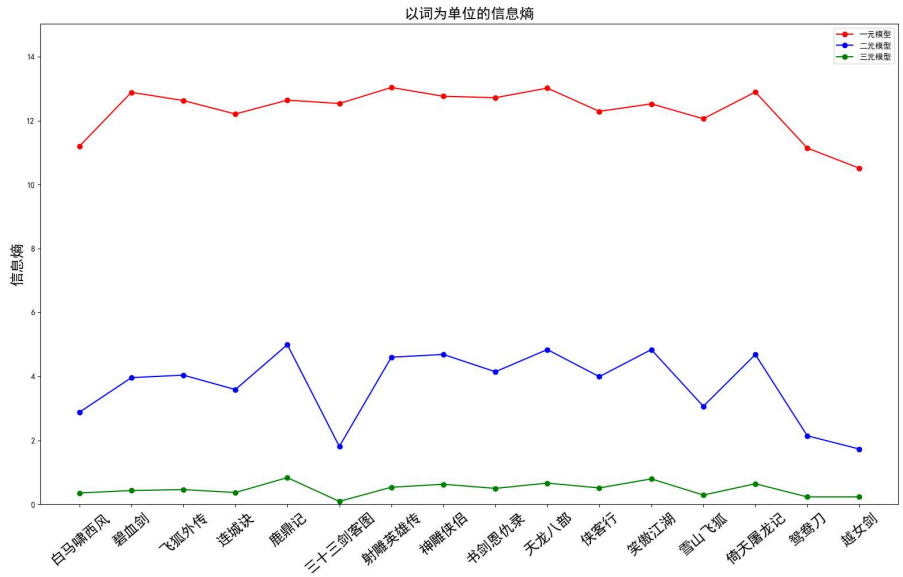

图 3 金庸 16 部作品以词为单位的平均信息熵

Conclusions

对金庸的 16 部作品进行了分词与词频统计分析，并分别计算了以字和词为单位的信息熵。

通过图 1 可以看出，这些作品中每个词出现的频率与其在频率表里的排名基本呈现出成反比的情况；再结合表 2 中每部作品的字数，可以发现，字数越多的作品，其规律越符合 Zipf's Law。

通过图 2 和图 3 可以看出，无论是一元、二元、三元语言模型，字/词的信息熵在每个作品间的变化趋势是相同的；同时可以发现，信息熵呈现一元模型＞二元模型＞三元模型的趋势，说明随着字数增多，字/词的表意愈发精确。

References

[1] Brown P F, Della Pietra S A, Della Pietra V J, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.