Report of Deep Learning for Natural Langauge Processing

Jiajie Wu
ZY2343402

Abstract

本文基于中文语料库中金庸的 16 部作品所有文本作为数据输入，通过结巴分词方式，统计每篇文本对应的词频，获取词-频率字典，按照频率-排次数据绘制表格及图像，同时取特定 rank 值直观感受 rank*frequency 的较小波动。观察对比发现，在可控的误差范围内，齐夫定律成立。同时计算得到以字和词为单位下的平均信息熵，

Introduction

Zipf's law，又名齐夫定律，即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。

信息熵在信息论中式接受每条信息中包含信息的平均量，又被称为信息熵、信源熵、平均自信息量。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 16 篇小说进行词频统计并验证齐夫定律，接着分别计算了字/词的信息熵，最后对实验结果进行了比较分析。

Methodology

Part 1:齐夫定律验证

验证齐夫定律需要对文本中的词和词频进行统计。具体方法如下：首先对文本进行预处理，删除所有的隐藏符号、非中文字符和标点符号：接着使用 jieba 库， 对每个 txt 文件中的文本进行分词：然后统计每个词出现的次数并绘图观察。

在 txt 文件中，在正文之中夹杂着大量空格、换行符、标点符号等，如图工所示，这些会影响分词的结果。因此再进行分词之前，首先对文本进行预处理，

只保留文本。

Jieba 库是一款优秀的用于中文分词的库，它利用一个中文词库，确定汉字之间的关联概率，汉字间概率大的组成词组，形成公词结果。Jieba 库支持精确、全模式、搜索引擎三种分词模式。本文中对经过预处理的文本使用 Jicba 精确模式，它将一段文本切分为若干个中文单词，不会产生冗余，是最适合词频统计的模式。

在得到分词结果之后，对词频(词出现的次数)和词序(词出现次数的排序，出现多的排在前)取对数画图。分别绘制了展示所有 txt 文件统计结果的总图和各个 txt 文件分图。

Part 2: 信息熵计算

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 $f(P_1,P_2)=f(P_1)+f(P_2)$，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

$$f(P)=\log\left\{\frac{1}{p}\right\}=-\log\{p\}$$

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：$U_1,\cdots,U_i,\cdots,U_n$，对应概率为：$P_1,\cdots,P,\cdots,P_n$，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$-\log p_i$的统计平均值(E)，可称为信息熵，即：

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

式中对数一般取 2 为底，单位为比特。

其中，P 为 X 的概率质量函数，E 为期望函数，I(X)为 X 的信息量。当样本数量有限时，公式可以表示为：

$$H(X)=\sum P(x_i)I(x_i)=-\sum P(xi)log_bP(x_i)$$

其中，b 取 2 的时候，熵的单位是 bit；b 取自然常数 e 时，熵的单位是 nat；b 取 10 时，熵的单位是 Hart。本文中 b 取 2。考虑语料中词与词之间的关系，可以分为一元语言模型、二元语言模型、三元语言模型、……、n 元语言模型。本文只考虑一元语言模型。

一元语言模型中，每个字\词出现的概率与其它字词无关。此时，字/词信息熵的计算公式为：

$$H(x)=-\sum_{x\in X}P(x)log_2P(x)$$

其中 P(x)可近似等于每个字或词在语料库中出现的频率

Experimental Studies

Part 1:齐夫定律验证

对 16 部作品分别分词统计，得到结果如图 1。

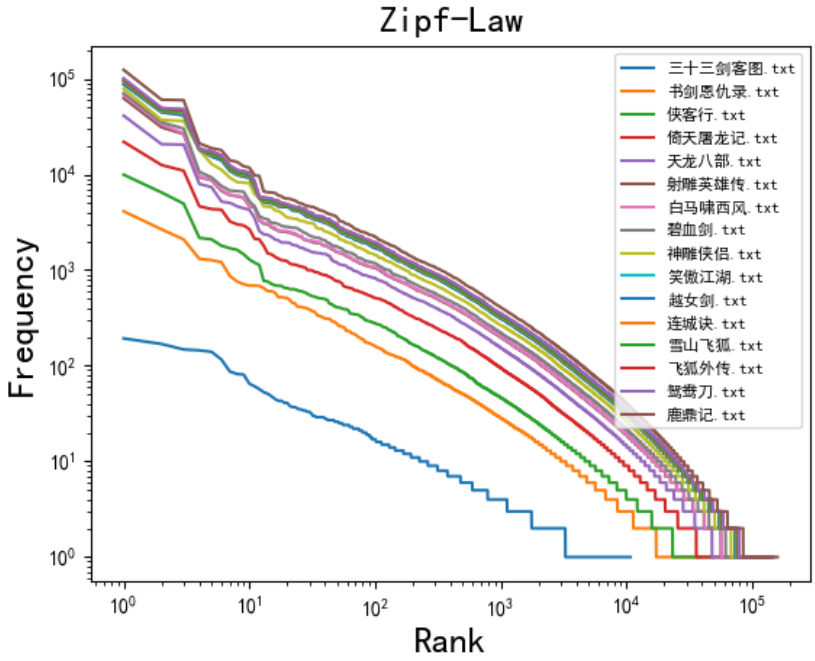


图 1 金庸作品词频统计

取特定$rank = \{10, 20, 30, 40, 50\}$，以特殊值的方式在较小范围内验证 Zipf_laws，各作品的特殊 rank-frequency 数据如表 1。

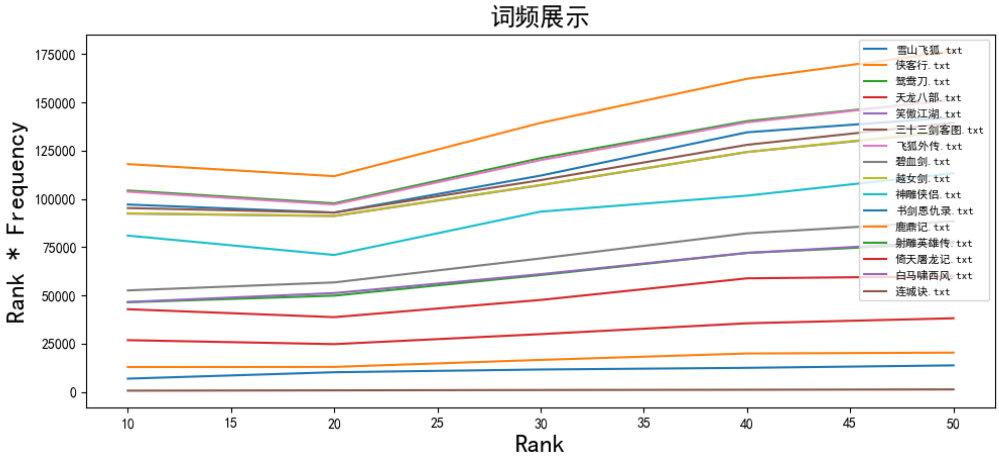
表 1 特定频率排次词频统计

书名	R	F	C=R*F	书名	R	F	C=R*F
三十三剑客图	10	65	650	神雕侠侣	10	8098	80980
	20	41	820		20	3547	70940
	30	33	990		30	3113	93390
书剑恩仇录	10	688	6880	笑傲江湖	10	9237	92370
	20	510	10200		20	4556	91120
	30	387	11610		30	3569	107070
侠客行	10	1290	12900	越女剑	10	9254	92540
	20	648	12960		20	4561	91220
	30	553	16590		30	3573	107190
倚天屠龙记	10	2682	26820	连城诀	10	9523	95230
	20	1238	24760		20	4645	92900
	30	998	29940		30	3657	109710
天龙八部	10	4284	42840	雪山飞狐	10	9707	97070
	20	1937	38740		20	4645	92900
	30	1588	47640		30	3735	112050
射雕英雄传	10	4645	46450	飞狐外传	10	10372	103720
	20	2493	49860		20	4854	97080
	30	2018	60540		30	4002	120060

白马啸西风	10	4659	46590	鸳鸯刀	10	10433	104330
	20	2560	51200		20	4885	97700
	30	2033	60990		30	4037	121110
碧血剑	10	5258	52580	鹿鼎记	10	11801	118010
	20	2836	56720		20	5588	111760
	30	2305	69150		30	4645	139350

注：表头中 F 为 frequency，即词频；R 为 rank，即词频排名；C=R*F 为齐夫定律的验证来源。由于数据量较大，在表格中仅展示$rank = \{10, 20, 30\}$部分的词频及排次，整体在图 2 中进行展示。

图 2 特定频率排次词频统计



Part 2: 信息熵计算

对 16 个 txt 文件分别计算字和词的信息熵，结果如表 2。

表 2 各文件字/词平均信息熵

File	Character Entropy	Word Entropy
三十三剑客图	9.194767	12.482304
书剑恩仇录	9.015893	12.628315
侠客行	8.737233	11.752044
倚天屠龙记	8.969869	12.561877
天龙八部	8.944395	12.560704
射雕英雄传	8.967772	12.355665
白马啸西风	8.299873	9.545214
碧血剑	9.028639	12.720257
神雕侠侣	8.942552	12.28667
笑傲江湖	8.810249	12.30987
越女剑	8.232131	9.142983
连城诀	8.726167	11.641214
雪山飞狐	8.783654	11.640543
飞狐外传	8.904812	12.457212
鸳鸯刀	8.426371	9.775895
鹿鼎记	8.813182	12.10208

Conclusions

本文对金庸的 16 部作品的 txt 文档进行了分词与词频统计分析，并分别计算了字和词的信息熵。图 1 显示了所有词频曲线变化趋势基本一致，且都与一条直线近似，验证了齐夫定律。由图 2 进一步分析，从$rank = 20$开始 C 呈较明显上升趋势的作品本身数据量较小，小于 500kb；而大于 1500kb 的作品的分词 rank*frequency 曲线基本持平。这说明字数越多的作品，其曲线越贴近横线，其规律越符合齐夫定律。

从表 2 中可见，不同作品字和词的信息熵基本都维持在类似水平，同时可以明显看出词的信息熵大于字的信息熵。

在使用 Jieba 库进行分词前，对文本进行的预处理中删除了标点符号，这可能会影响分词的准确度，是可以在后续工作中改进的点。

References

[1] Brown P F, Della Pictra S A. Dclla Pictra V J. ct al. An estimate of an upper bound for the entropy of English[Jl.Computational Linguistics, 1992.18(1):31-40.

[2] historyasamirror.Zipfs law[EB/OL].2008:[2024-04-07].https://blog.csdn.net/historyyasamirror/article/details/3125223