

深度学习与自然语言处理 作业 1 报告

魏才伦

1260034837@qq.com

Abstract

本文首先在提供的中文语料库中验证了 Zipf's Law, 然后学习了论文《Entropy of English》计算信息熵的方法, 在提供的中文语料库上, 分别计算了一元、二元、三元中文词组信息熵和中文字组信息熵, 结果依次为 12.316 比特/词、3.749 比特/词、0.473 比特/词, 以及 9.559 比特/字、5.197 比特/字、1.734 比特/字

Introduction

Zipf's law 是一个经验观察, 它指出一个词或术语的频率与它的排名成反比。它可以表述为: 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,, 频次最小的词等级为 D。

信息是个很抽象的概念。人们常常说信息很多, 或者信息较少, 但却很难说清楚信息到底有多少。比如一本五十万字的中文书到底有多少信息量。直到 1948 年, 香农提出了“信息熵”的概念, 才解决了对信息的量化度量问题。

Methodology

本文采用 python 中的 jieba 模块进行中文分词, 统计了中文语料库中的词频, 验证了 Zipf's Law, 然后采用统计语言模型的方法, 计算了一元、二元、三元中文信息熵, 下面将对方法进行详细介绍。

M1: jieba 模块

为验证 Zipf's Law, 本文采用了 python 的 jieba 模块。jieba 模块是一个 python 中文分词组件, 其利用一个中文词库, 确定汉字之间的关联概率, 汉字间概率大的组成词组, 形成分词结果。jieba 采用基于前缀词典实现高效的词图扫描, 生成句子中汉字所有可能成词情况构成的有向无环图, 然后采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合。

M2: 验证 Zipf's Law

通过 jieba 模块分词后，遍历所有分词，去除标点和停用词，得到了中文语料库的所有分词结果，构建一个字典数据结构，字典的 key 为中文语料库的分词结果，即语料库的词组，value 为词组出现的次数，即词频。根据词频对字典从大到小排序，得到词组的频率排名，最后横轴为词组频率排名，纵轴为词频，绘制曲线，为了更加直观的表现 Zipf's Law 对横轴和纵轴进行了 log 缩放，绘制出的曲线近似为直线，符合 Zipf's Law。

M3: 中文信息熵

获得信息可以被认为是不确定性的减小的过程，对于一个信源发送什么符号是不确定的，假定信源有 n 种取值 U_1, U_2, \dots, U_n ，对应的概率为 P_1, P_2, \dots, P_n ，且各种符号的出现彼此独立，对于每个符号定义一个不确定性函数 $f(P_i) = -\log P_i$ ，信源的平均不确定性为每

个符号不确定性的统计平均值 $-\sum_{i=1}^n P_i \log P_i$ ，称这个统计平均值为信息熵，记为

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

下面将信息熵的概念引入到自然语言中，假定 S 表示某个有意义的句子，由一连串特定顺序排列的词 w_1, w_2, \dots, w_n 组成， n 为句子的长度。计算信息熵需要已知 S 出现的概率，即 $P(S)$ 。利用条件概率公式， S 的概率等于每个词出现的条件概率相乘，则有

$$P(S) = P(w_1)P(w_2 | w_1)P(w_3 | w_1, w_2) \cdots P(w_n | w_1, w_2, \dots, w_{n-1})$$

此时一个词的概率由前面的 $N-1$ 个词决定，称其为 N 元模型，然而当句子过长时， $P(w_n | w_1, w_2, \dots, w_{n-1})$ 的可能性太多，无法估算。假设句子具有马尔可夫性，即任意一个词 w_i 出现的概率只同它前面的词 w_{i-1} 有关，则 $P(S)$ 变为

$$P(S) = P(w_1)P(w_2 | w_1)P(w_3 | w_2) \cdots P(w_n | w_{n-1})$$

其对应的统计语言模型为二元模型。当 $N=1$ 时，每个词出现的概率与其他词无关，称为一元模型，则 $P(S)$ 变为

$$P(S) = P(w_1)P(w_2) \cdots P(w_n)$$

如果统计量足够大，字、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

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

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

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

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概率，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。

Experimental Studies

本文以金庸小说集为中文语料库，编写 python 程序，验证 Zipf's Law，并统计一元、二元和三元信息熵，程序代码详见。

验证 Zipf's Law 的结果如图 1 所示，图中横坐标为词频数排名，纵坐标为词频，为便于展示规律，横纵坐标轴进行了 log 缩放，可以看出曲线图近似为一条递减的直线，符合 Zipf's Law。

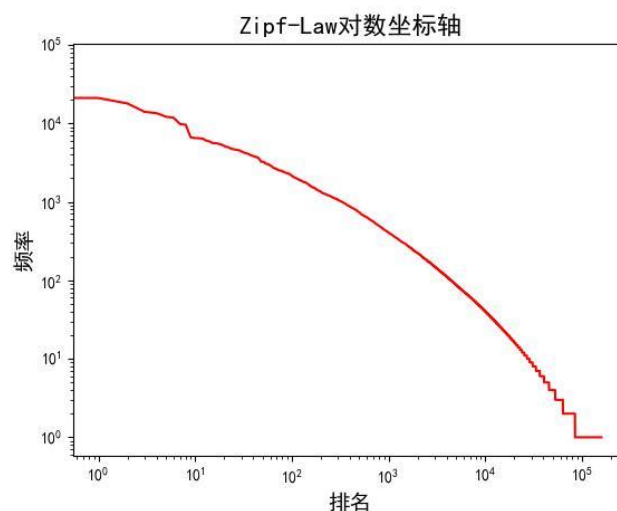
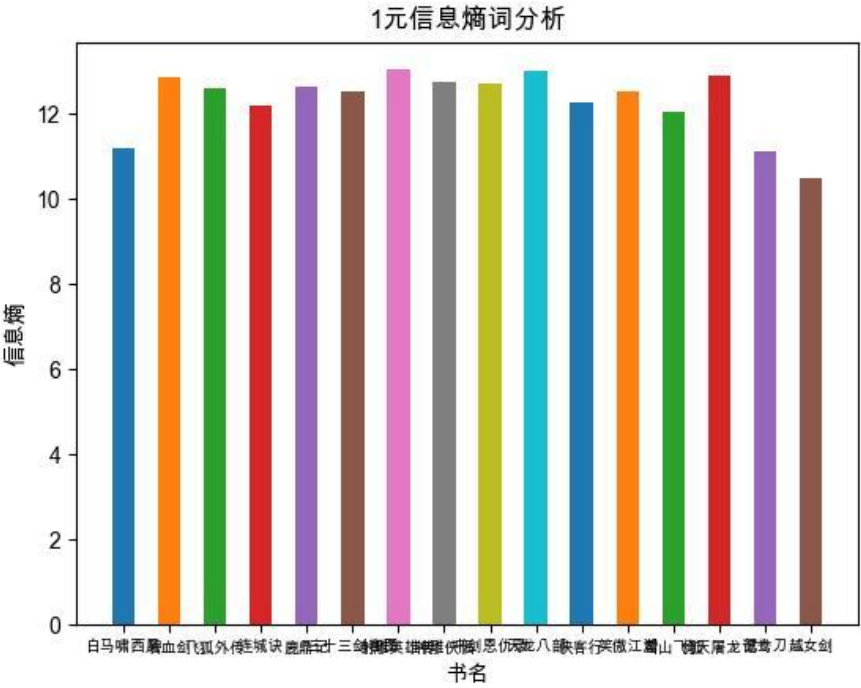


图 1 Zipf's Law 曲线图

去除标点符号和停用词，统计金庸小说集中每部小说的一元、二元和三元词组信息熵和字组信息熵，一元、二元和三元词组信息熵的计算结果如表 1 所示，图 2 为表 1 的结果所绘制的柱状图，图 3 为字组信息熵的柱状图结果。

表 1 金庸小说集中文词组信息熵

小说名称	一元信息熵 比特/词	二元信息熵 比特/词	三元信息熵 比特/词
《白马啸西风》	11.199	2.876	0.354
《碧血剑》	12.886	3.962	0.431
《飞狐外传》	12.626	4.040	0.461
《连城诀》	12.207	3.589	0.369
《鹿鼎记》	12.639	4.992	0.834
《三十三剑客图》	12.534	1.808	0.091
《射雕英雄传》	13.045	4.593	0.533
《神雕侠侣》	12.760	4.687	0.626
《书剑恩仇录》	12.727	4.136	0.497
《天龙八部》	13.019	4.838	0.663
《侠客行》	12.288	3.992	0.512
《笑傲江湖》	12.524	4.838	0.795
《雪山飞狐》	12.058	3.064	0.290
《倚天屠龙记》	12.893	4.685	0.642
《鸳鸯刀》	11.141	2.142	0.232
《越女剑》	10.511	1.728	0.232
平均	12.316	3.749	0.473



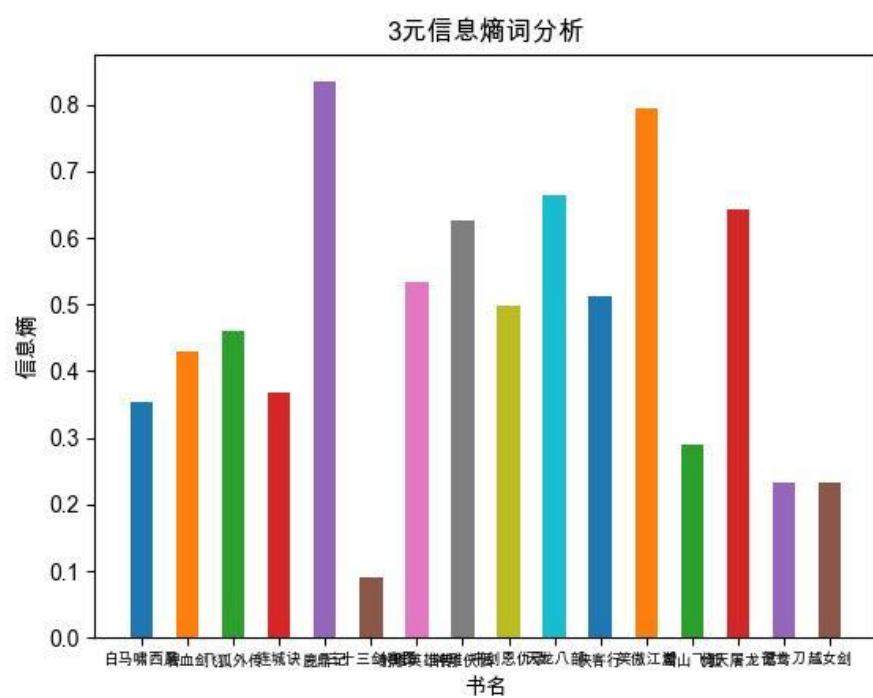
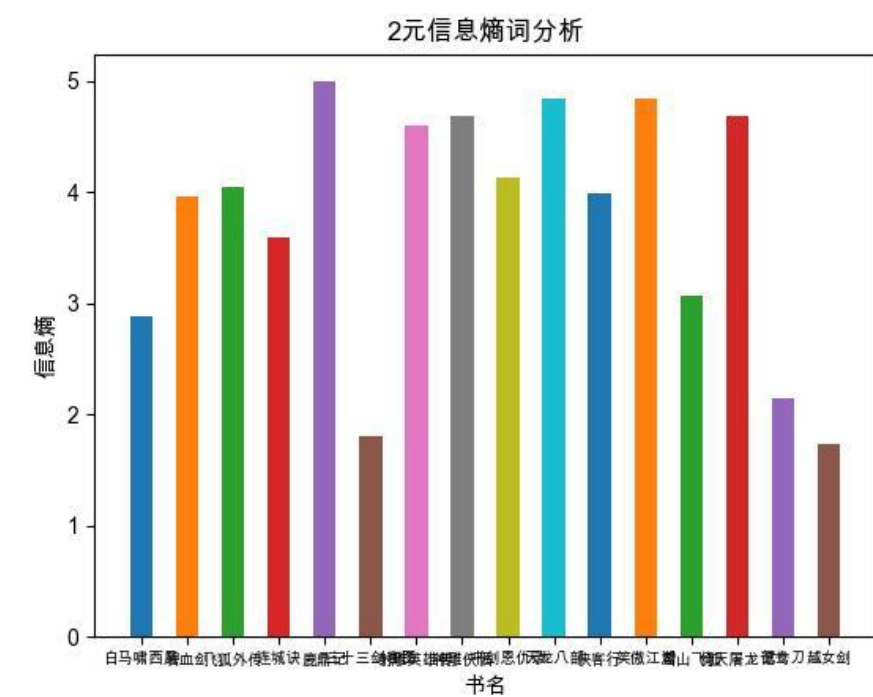
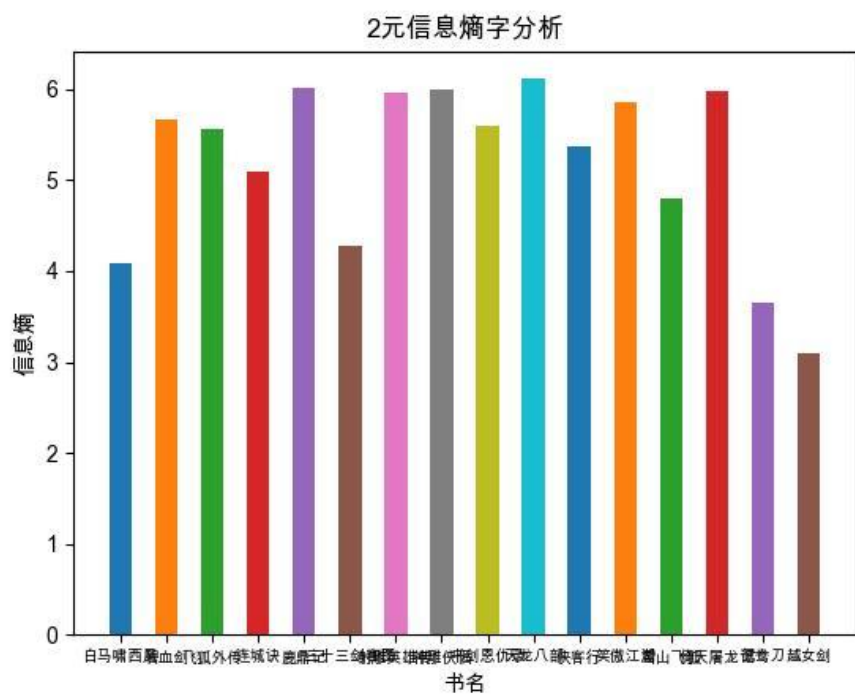
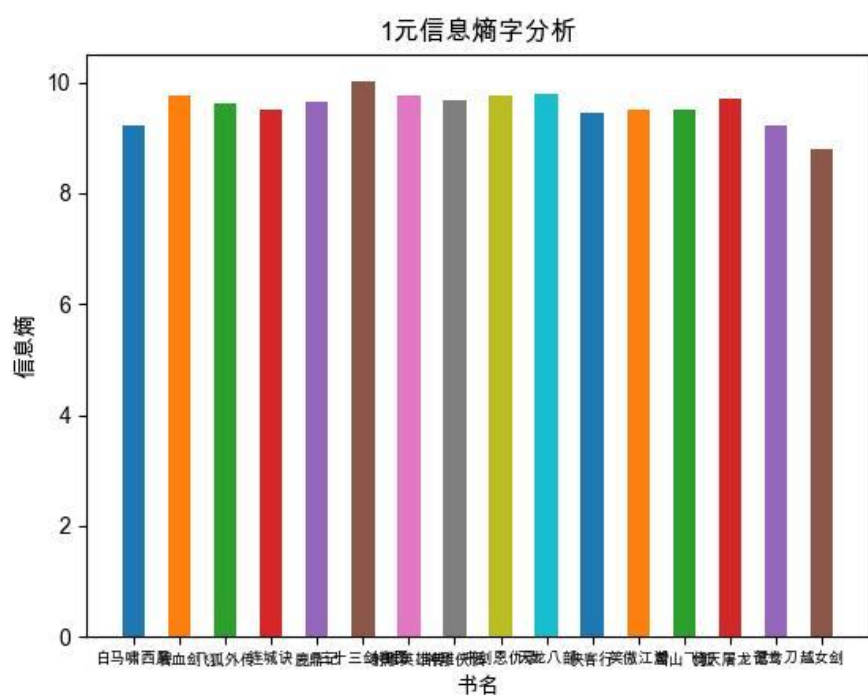


图 2 金庸小说集词组信息熵柱状图



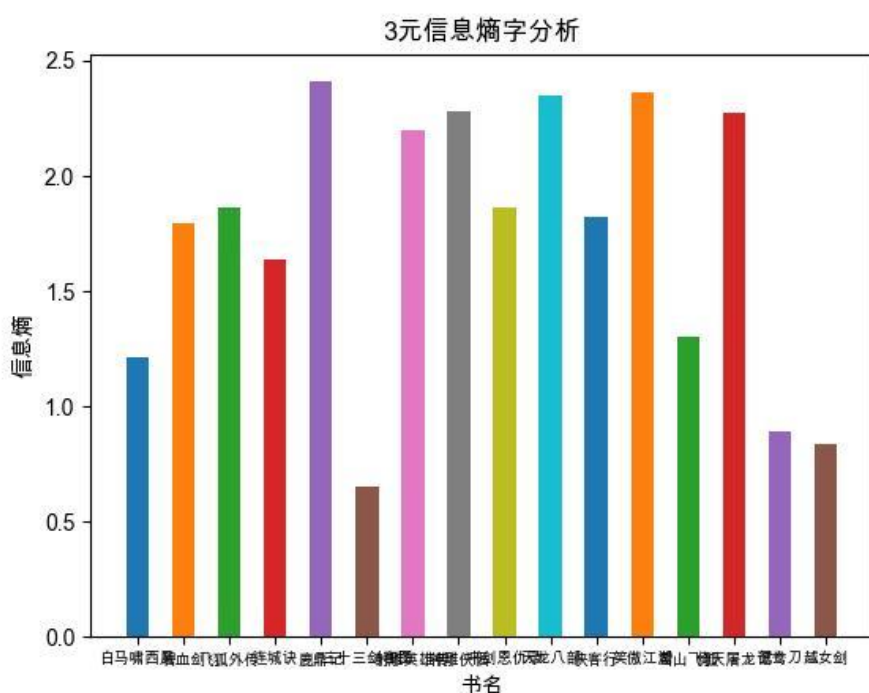


图 3 金庸小说集字组信息熵柱状图

Conclusions

Zipf's Law 可以表示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，当坐标轴采用 log 缩放时，曲线近似为一条递减的直线。

对比一元模型、二元模型、三元模型可以看到， N 取值越大，即考虑前后文关系的长度越大，文本的信息熵越小，这是因为 N 越大，组成该词组的词越多，其冗余度也就越小，使用的特定场景越小，出现在文章中的不确定性越小。

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.
- [2] <https://zhuanlan.zhihu.com/p/658563402>
- [3] https://blog.csdn.net/qq_37098526/article/details/88633403
- [4] https://blog.csdn.net/weixin_43353612/article/details/105147148
- [5] https://blog.csdn.net/qq_45688080/article/details/130669630