DNLP 作业 1

中文语料库的 Zipf’s Law 与 Entropy

张乃天 SY2302516
freshznt@buaa.edu.cn

Abstract

本次作业旨在利用若干小说组成的中文语料库验证 Zipf’s Law，并分别计算以词和字为单位的熵。结果表明，去除停用词后的中文的确满足 Zipf’s Law 所描述的规律；中文以“词”和“字”为符号的信息熵分别大致为 11bit 和 9bit。

The purpose of this assignment is to verify Zipf's Law using a Chinese corpus composed of several novels, and to calculate entropy in terms of words and characters respectively. The result shows that, after removing stopwords, the Chinese language indeed conforms to the Zipf's Law; The entropy of Chinese in terms of words and characters are roughly 11 bits and 9 bits.

PART1: Zipf’s Law

Introduction

Zipf's Law 是自然语言处理中的一个重要概念，由美国语言学家 George Kingsley Zipf 于 20 世纪提出，是一种经验定律，描述了自然语言中词频和词序之间的关系。词汇的频率与其排名之间存在着反比关系，换句话说，高频词的排名较低，低频词的排名较高，这是一种常识的、定性的描述。而 Zipf’s Law 却用了一个定量的数学公式严格描述了词频与排名之间的这种反比关系，即在对数尺度下，词频与排名大致呈一个斜率为负的一次函数。

Methodology

在此部分中，为了验证 Zipf’s Law 所描述的关系，首先利用 jieba 库对老师所提供的中文语料库（16 个小说文档）进行分词，去掉停用词；随后统计词频并排序，绘制对数尺度下词频与排序的函数曲线即可。

Result

所绘制的词频与排序的关系曲线如图 1 所示。可以看到，在对数尺度下，词频与排序确实大致呈现出一个斜率为负的一次函数。验证了 Zipf’s Law 的正确性。

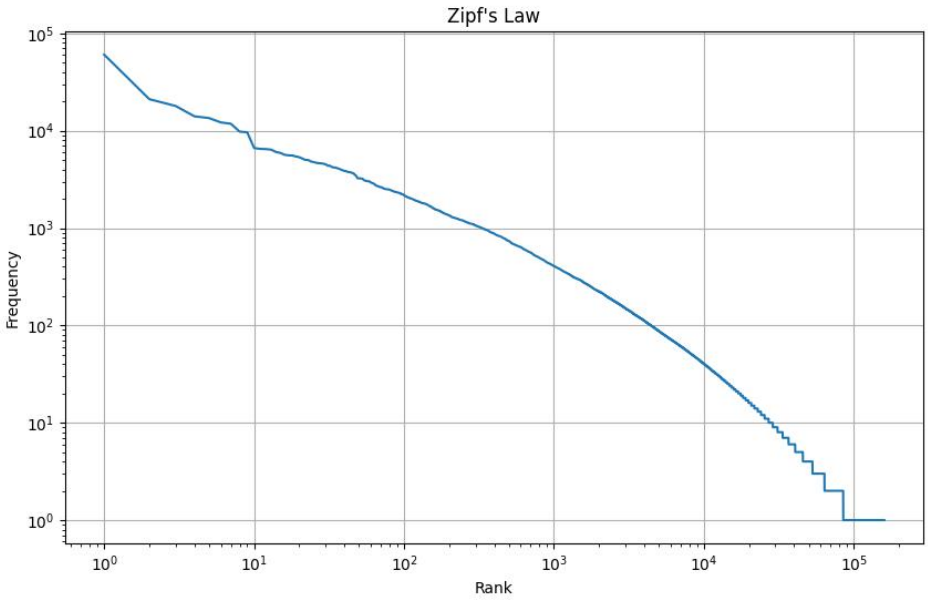

图 1 中文语料库的 Zipf's Law 曲线

PART2: Entropy

Introduction

熵是信息论中提出的概念，用于描述一个集合（随机变量）的平均信息量或不确定程度。其公式为：

$$\mathrm{H(X)}=\sum_{i}-p_i\log(p_i)$$

可以看到，熵完全由集合的概率分布决定。对于语言模型这种庞大的集合来说，需要对庞大的语料库进行统计，用以估计语言模型中各种字、词的概率分布。此外，考虑到相邻字、词之间的相关性，利用联合概率或条件概率计算的熵或许更加合理。基于上面的考虑，可以假设语言模型是一个 N-1 阶马尔可夫链，并利用 N-gram model 对语言的概率分布进行建模，即当前的字或词的概率分布只与前 N-1 个字或词有关。这样，就可以通过统计一个 N-1 阶的条件概率分布来计算条件熵 $H(X_N \mid X_{N-1}...X_2X_1)$。

Methodology

在此部分中，将分别假设 1-gram 2-gram 和 3-gram 模型成立。为了简化计算条件熵的难度，这部分将利用“平均符号熵”对 N-gram model 的条件熵进行近似。为了阐明上述近似的合理性，首先要引入极限熵的概念。在信息论中，极限熵的定义为平均符号熵的极限：

$$\mathrm{H}_\infty=\lim_{N\to\infty}\frac{1}{N}H(X_1X_2...X_N)$$

假设信源平稳，则平均符号熵是单调递减的有界数列，因此极限存在，极限熵存在。其物理意义就是一个符号（一个字或词）所携带的平均信息量。由于我们假设语言模型具有马尔可夫性质，则在进一步假定遍历性后可以证明语言模型一定是平稳信源，因此语言模型的极限熵存在。此外，还可以证明，在 N-gram model 的假设下，N-1 阶的条件熵就等于极限熵。

综上所述，N 阶平均符号熵就是 N-gram model 条件熵的一个上界。

$$H(X_1) = H(X_1)$$
$$\frac{1}{2}H(X_1X_2) \geq H(X_2 \mid X_1)$$
$$\frac{1}{3}H(X_1X_2X_3) \geq H(X_3 \mid X_2X_1)$$

Result

下面给出以字和词为符号单位的熵。其中以词为单位时使用 jieba 库进行分词，以字为单位时直接读取 txt 文档即可。

表 1 以“词”和“字”为单位的语料库信息熵

符号单位 / 模型	词	字
1-gram（符号熵/bits）	12.18	9.53
2-gram（二阶平均符号熵/bits）	9.56	8.13
3-gram（三阶平均符号熵/bits）	7.14	6.74

虽然表格中平均符号熵递减与理论相符，但本人认为表格中的 3-gram model（三阶平均符号熵）的结果太小，不准确。原因在于中文语料库是有限长的。当我们所研究的符号越大，其可能的组合情况将呈指数级增长。然而从语料库中可供选取大符号的个数却是有限的。这就会导致语料库太小，并没有足够多的大符号素材。这就好比，假设将语料库中所有的汉字看作一个超大符号，则整个语料库中只出现过这一种符号，熵为 0。这种结果显然是荒谬的。综上，本人认为 1 gram 和 2 gram model 的结果较为可信，因此中文的“词”熵大致为 11bit 左右，而“字”熵大致在 9bit 左右。

Conclusions

本次作业利用所提供的中文语料库，验证了著名的 Zipf’s Law，并估算了中文以词和字为单位的信息熵。其中中文的词语熵大致为 11bit 左右，字熵大致在 9bit 左右。

通过本次作业所完成的简单验证和计算，熟悉了自然语言处理的特点和其关心的问题。