Report of Deep Learning for Natural Langauge Processing

Pan Yao
1239388514@qq.com

Abstract

本次实验总共分为两个部分，第一部分：通过中文语料库来验证 Zipf's Law. 第二部分：计算中文(分别以词和字为单位) 的平均信息熵。

Introduction

一.Zipf 定律

Zipf 定律（Zipf's Law）是一种在自然语言处理、信息科学、统计学等领域中常见的经验规律。这一定律由美国语言学家乔治·金斯利·齐普夫（George Kingsley Zipf）提出，描述了在各种自然语料库中词汇的频率分布。Zipf 定律指出，一个词的频率与它在频率表中的排名反比。换言之，某一自然语言文本中第 r 位排名的词出现的频率 f，与它的排名成反比。

二.信息熵

信息熵（Entropy），在信息论中，是用来衡量信息量或者不确定性的一个量度。它最初由克劳德·香农（Claude Shannon）在 1948 年提出，作为信息论的基础概念之一。信息熵的概念在统计学、物理学、信息理论、以及与数据处理相关的众多领域都有应用。

在信息论中，信息熵定义为一个消息集合的平均信息量，用来描述一个信息源产生的数据平均所包含的信息量。如果一个信息源产生的数据越是不确定或者随机，那么它的信息熵就越高；反之，如果数据越是确定和有序，信息熵就越低。

信息熵可以被看作是衡量信息的不确定性的度量。举个例子，如果我们抛一枚完全均匀的硬币，结果有正面和反面两种可能，每种可能发生的概率都是 0.5，那么这个事件的信息熵是 1 比特。这表示我们在知道抛硬币结果之前，存在的不确定性量。相比之下，如果一枚硬币总是正面朝上，那么抛这枚硬币的信息熵就是 0，因为结果是完全确定的，没有不确定性。

信息熵的概念不仅用于信息理论和通信领域，还被广泛应用于数据压缩、密码学、机器学习、语言模型构建、生态学种群多样性的度量等多个领域。在这些领域，信息熵帮助人们量化和理解数据的不确定性和复杂性。

Methodology

一.验证 Zipf's Law

验证 Zipf 定律通常涉及统计一个文本集合中单词的频率，并与其排名进行比较。这可以通过以下步骤实现：

步骤 1: 准备数据
首先，需要一个足够大的文本数据集来确保结果的有效性。可以使用小说、新闻文章、甚至整个文本库。
步骤 2: 文本分词
对文本进行分词处理，统计每个单词的出现次数。这在英文中比较直接，但在中文或其他使用复合文字的语言中，需要用到分词软件如 jieba（对于中文）。
步骤 3: 计算频率和排名
将单词按频率排序，最常见的单词排在最前面。记录每个单词的频率以及它的排名。
步骤 4: 数据可视化
通常，验证 Zipf 定律的最直观方法是通过可视化。可以绘制一个图表，横轴为单词的排名（对数尺度），纵轴为单词的频率（同样对数尺度）。如果数据遵循 Zipf 定律，图表中的点应该近似于一条直线。
步骤 5: 分析结果
对图表进行观察和分析，看看它是否符合 Zipf 定律的预期——即频率与排名的反比关系。

二. 信息熵计算

信息熵的计算公式为:

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

二元模型的信息熵计算公式为:

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

三元模型的信息熵计算公式为:

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

Experimental Studies

Figure 1：验证 Zipf's Law

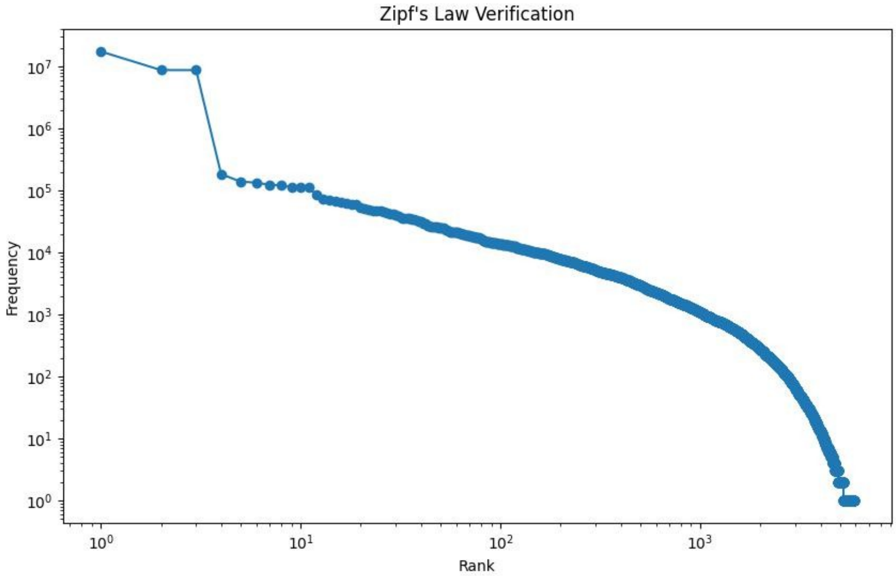

Table 1: this is the table 1

	字(比特/字)	词(比特/词)
一元	3.812458816	3.756804399
二元	2.189820479	2.206645453
三元	1.79800048	1.8126641

Conclusions

通过本次实验，我们不仅验证了 Zipf's Law 在中文语料库中的普适性，同时还通过对于中文语料库信息熵的计算，得到了当 N-gram 模型的 N 值越低，信息熵就越大，数据的不确定性越大的结论。

References

[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] C. E. Shannon, "A mathematical theory of communication," in The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, July 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.