Report of Deep Learning for Natural Langauge Processing

齐夫定律验证与中文信息熵计算

20376223 韩子轩
18210237202@163.com

Abstract

本次实验以金庸武侠小说作为中文语料库，在进行了数据集加载与预处理后使用结巴分词库对文本里的词语进行分割，统计词频并绘制频率分布图，成功验证了齐夫定律。之后分别使用 unigram，bigram 和 trigram 模型计算文本的信息熵。随着 N-gram 模型的 N 增大，计算得到的信息熵递减。

Introduction

齐夫定律是哈佛大学语言学家乔治·齐夫（George Zipf）于 1949 年发现的一项实验定律。该定律指出，在自然语言中，一个词语出现的频率与其在频率表中的排名成反比关系。对于数学家和统计学家来说，齐夫定律代表着一种典型的幂律分布，准确地说是帕累托分布的特例，可以形象地描述为「二八定律」，比如任何国家的 20%的人口拥有 80%的国民财富。在齐夫定律中，则是第 1 富有的是第 n 富有的 n 倍。

在他的著作中，齐夫提出了一个基于人类行为和经济学的解释。他认为，人类作为一个「能量-物质」系统，为了解决问题，会选择一条最省力、或成本最小的路径。这既包括系统本身的物质-能量流动，也包括整体单位在环境中的移动，他将这一原则称之为最省力原则（Principle of Least Effort）。为遵循最省力原则，特定语言的词汇不断优化，最终决定了词语中实际数量的词汇和意义。

信息熵的概念最早由香农于 1948 年借鉴热力学中的“热熵”的概念提出，旨在表示信息的不确定性。信息熵具有单调性、非负性和累加性。熵值越大，则信息的不确定程度越大。其数学公式可以表示为：

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

针对于联合分布的随机变量 $(X,Y) \sim P(X,Y)$，在两变量相互独立的情况下，其联合信息熵为：

$$H(X,Y) = -\sum_{x \in \chi} \sum_{y \in \gamma} p(x,y) \log p(x,y)$$

Methodology

1.数据集处理

选取金庸 16 本武侠小说作为中文语料库。首先将它们合并读取，并删除所有隐藏字符（如换行符）、非中文字符和广告标语，以保证语料库的纯净度。

2.齐夫定律的验证

使用 jieba 库对文本进行中文分词，并统计每个词语的词频，最后绘制词频分布的双对数图，观察是否符合反比关系。

3.中文信息熵的计算

N-gram 模型是一种基于前 N 个词出现频率的概率模型，常用于自然语言处理任务中。在 N-gram 模型中，N 表示模型考虑的上下文的长度。本实验分别采用 1-gram,2-gram 和 3-gram 模型计算信息熵；

– unigram model: $P(wi|w1,w2,\ldots wi-1) \approx P(wi)$

– bigram model: $P(wi|w1,w2,\ldots wi-1) \approx P(wi|wi-1)$

– trigram model: $P(wi|w1,w2,\ldots wi-1) \approx P(wi|wi-2,wi-1)$

N-gram 模型的核心假设是“马尔可夫假设”，即当前词出现的概率只与前 N-1 个词相关，而与其他词无关。这个假设简化了模型的复杂性，使得语言模型的建模和计算变得可行。

Experimental Studies

以金庸十六本小说合并作为中文语料库，得到的词频分布的双对数图如下图所示：

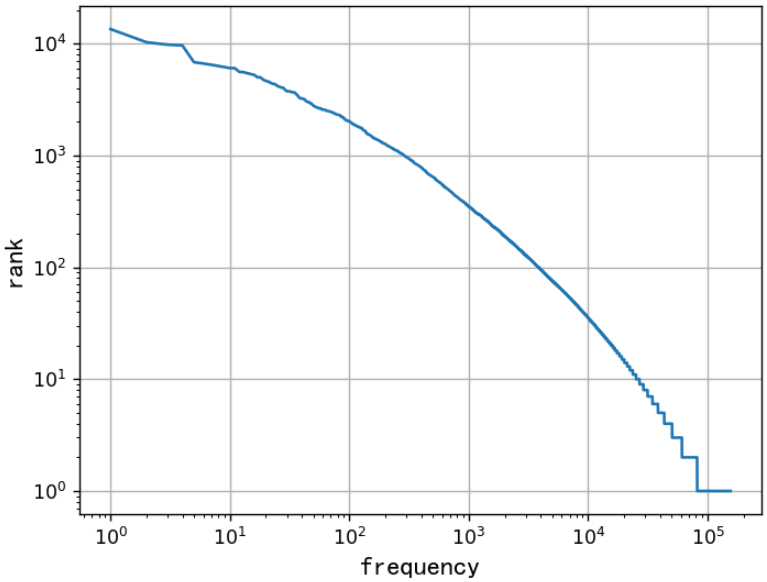

图 1: 频率-名次图

图中横坐标为词频，纵坐标为该词词频对应的排名。可以看到在双对数图中，图像近似一条直线，说明一个词语出现的频率与其在频率表中的排名成反比关系，符合齐夫定律。

	语料库字数	分词个数	信息熵（比特/词）
一元模型	7419437	4430459	12.01287
二元模型	7419437	4430459	6.89176
三元模型	7419437	4430459	2.41675

表 1: 中文信息熵

上表为使用 1-gram、2-gram 和 3-gram 三种语言模型计算中文信息熵的结果，我们可以观察到随着 N 值的增大（即考虑上下文关系的长度增加），文本的信息熵降低。这是因为较大的 N 值会导致分词后文本中的词组分布更加简单。较大的 N 值会使得固定数量的词出现频率更高，减少了由字或短词打乱文章顺序的可能性。这种有序性减少了由字组成词和句子的不确定性，从而降低了文本的信息熵，这也符合我们的实际认知。

Conclusions

本实验使用金庸小说集作为中文语料库，统计词频并进行排序，验证了齐夫定律的适用性。使用 N-gram 模型计算了文本的中文信息熵，发现较大的 N 值提供了更多的上下文信息，使得模型更准确地预测词语的出现顺序，从而减少了文本的不确定性，进而降低了信息熵。

References

[1] MAZZARISI, ONOFRIO, DE AZEVEDO-LOPES, AMANDA, ARENZON, JEFERSON J., et al. Maximal Diversity and Zipf's Law[J]. 2021,127(12):128301.1-128301.5. DOI:10.1103/PhysRevLett.127.128301.

[2] Python——jieba 优秀的中文分词库（基础知识+实例）
https://blog.csdn.net/m0_63244368/article/details/126837925

[3] 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.