Report of Deep Learning for Natural Langauge Processing

Weida Chen

908715799@qq.com

Abstract

本研究将金庸的十六部经典武侠小说作为语料库进行研究分析，验证了齐夫定律的适用性，并分别以单个字符和词为基本单位，对文本进行了一元、二元和三元的统计分析，从而计算得到了语料库中字和词的中文平均信息熵。

Introduction

齐夫定律（Zipf's Law），由美国语言学家乔治·齐夫于 1949 年提出，是描述自然语言中词频分布的一个经验定律。该定律指出，在给定的语料库中，一个词的频率与其在频率表中的排名成反比。换句话说，第二频繁出现的词出现的频率大约是最频繁出现的词频率的 1/2，第三频繁的是 1/3，依此类推。这意味着少数几个词汇在文本中出现的频率极高，而大多数词汇出现的频率则相对较低。齐夫定律在自然语言处理（NLP）、信息论、以及其他许多领域都有广泛的应用。例如，它被用来解释各种自然现象和社会现象中的分布规律，包括城市人口分布、网站访问量分布等。在语言学和文本分析中，齐夫定律帮助研究者了解和预测词汇使用的规律性，从而在文本挖掘、语言模型建构、以及其他 NLP 任务中发挥重要作用。

信息熵（Entropy）是衡量信息量的一个指标，最初由克劳德·香农在信息论中提出。对于语言模型而言，信息熵可以被视为平均每个词或字符携带的信息量。更高的信息熵意味着文本中的不确定性更大，预测下一个词的难度也更高。

N 元语言模型（N-gram language model）是自然语言处理（NLP）中一种基础而重要的模型，用于预测或估计一系列词汇（或字符）的出现概率。在这种模型中，一个词的出现概率假定只与它前面的 N-1 个词有关。通过统计和分析这些词序列（N-gram）的出现频率，N 元模型可以用来估计文本的信息熵，进而评估语言模型的复杂性和不确定性。

在本项研究中，我们选取了金庸十六部小说构建语料库，以此为基础探讨了齐夫定律在中文文本中的有效性。通过将分析的基本单位设定为单个字或词，本研究进一步细分为一元、二元以及三元的统计学分析，旨在估算出该语料库内字符和词汇的中文平均信息熵。

Methodology

下面将金庸的十六部经典武侠小说作为语料库，分别验证齐夫定律并计算中文平均信息熵。

M1：验证齐夫定律

根据齐夫定律的内容，需要验证一个词出现的频率与其在频率表中的排名成反比。因此首先需要统计每个词在语料库中出现的频率。由于原始文本中存在大量的空格以及标点符号等非中文字符；并且文本中存在大量的频繁出现的功能词以及无实际意义的词语，也即停用词，需要将其从语料库中删去。因此，实验首先对原始数据进行预处理，删除文本中的非中文字符。此后使用 jieba 库对中文语料进行分词，得到语义较为独立的词。然后导入停用词库，对分词后的语料库进行筛选，滤去语料库中所有的停用词。最后，使用 collections 库中的 Counter 计数器，对出现的每一个词统计词频，并绘制词频-排名图。

M2：计算中文平均信息熵

信息熵描述了接收的每条信息中包含的平均信息量，或者说，它度量了信息的混乱程度。在给定的消息集合中，信息熵越高，表明该消息集合的不确定性越大，每条消息携带的平均信息量也就越多。对于一个随机变量X，其信息熵$H(X)$定义为：

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

这个公式中的对数常用以 2 为底，意味着信息熵的单位通常是比特（bit）。

下面介绍 N-gram 语言模型。N-gram 语言模型通过统计的方式来预测或确定一个词语在给定的一系列词语之后出现的概率。在这个模型中，"N"表示的是数目，指的是在做预测时，考虑的上下文（即前面的词语）的数量。N-gram 模型根据前 N-1 个词来预测第 N 个词，因此，它是一种基于前面词语序列的条件概率模型。N=1 时为一元模型，模型仅考虑每个单独的词出现的概率，而忽略词与词之间的关系。N=2 时为二元模型，模型考虑一个词出现的概率依赖于它前面的一个词。N=3 时为三元模型，模型考虑一个词出现的概率依赖于它前面的两个词。

下面介绍基于 N-gram 模型计算中文的平均信息熵。当统计量达到一定的规模时，自然语言中的单个字符（字）、单词、以及多字词组（如二元词组和三元词组）出现的概率可以通过其在文本中出现的频率来近似估计。这一原理基于大数定律，即在足够大的样本中，某一事件的频率趋向于其概率。由此可得，字和词的一元模型信息熵计算公式为：

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

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$P(x,y)$可可近似等于每个二元词组在语料库中出现的频率，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

E1：验证齐夫定律

本实验将金庸的十六部经典武侠小说作为语料库，对语料进行预处理后，得到的词频-排名对数图如图一所示。

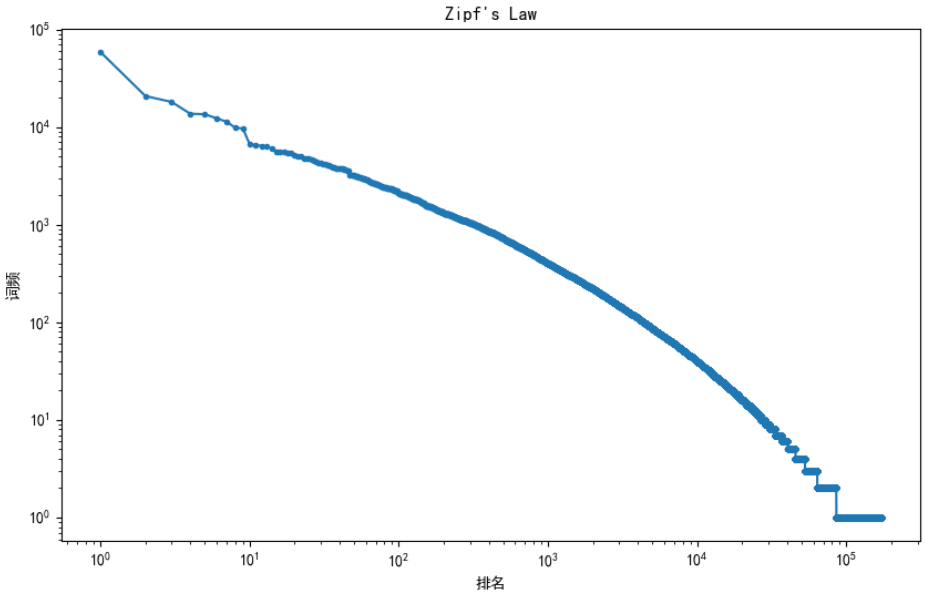

图 1：基于金庸十六部经典武侠小说的词频-排名对数统计图

从词频-排名对数统计图中可以看出，分别对词频和排名取对数后，它们在对数-对数图上呈现为一条斜率近似为-1 的直线。这恰好说明了在金庸十六部经典武侠小说中，一个词出现的频率与其在频率表中的排名成反比。因此，本实验成功地基于金庸小说验证了齐夫定律。

E2：计算中文平均信息熵

本实验部分以金庸小说作为语料库，去除标点符号和停用词等无用信息，分析的基本单位设定为单个字或词，基于一元、二元和三元模型计算中文平均信息熵，并分别绘制柱状图如图 2 和图 3 所示。

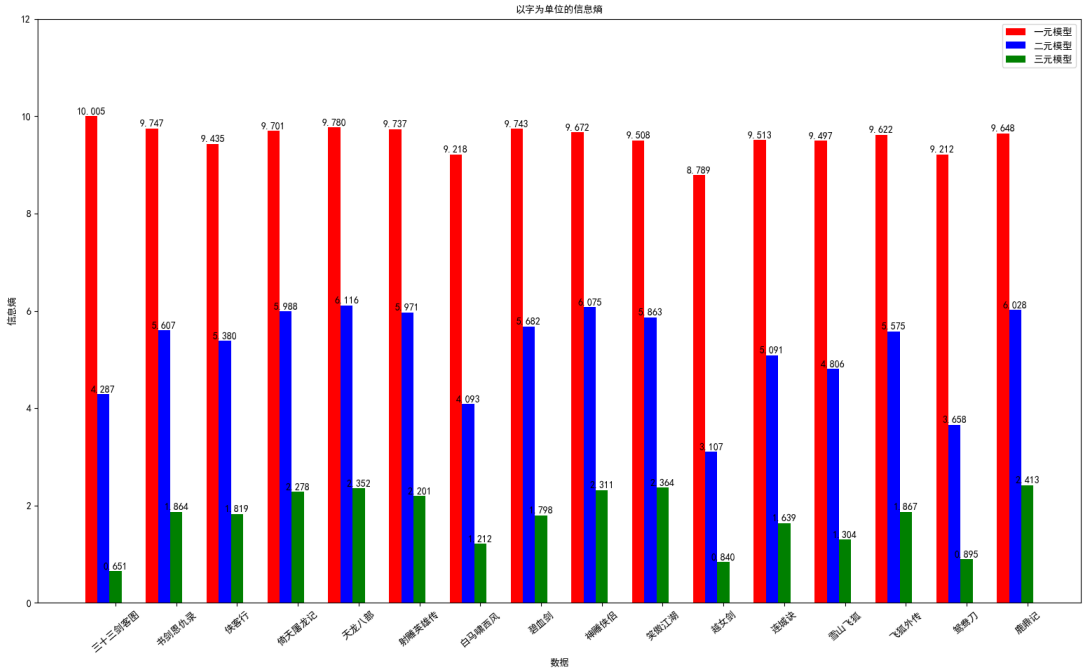

图 2：基于一元、二元和三元模型计算中文平均信息熵（以字为单位）

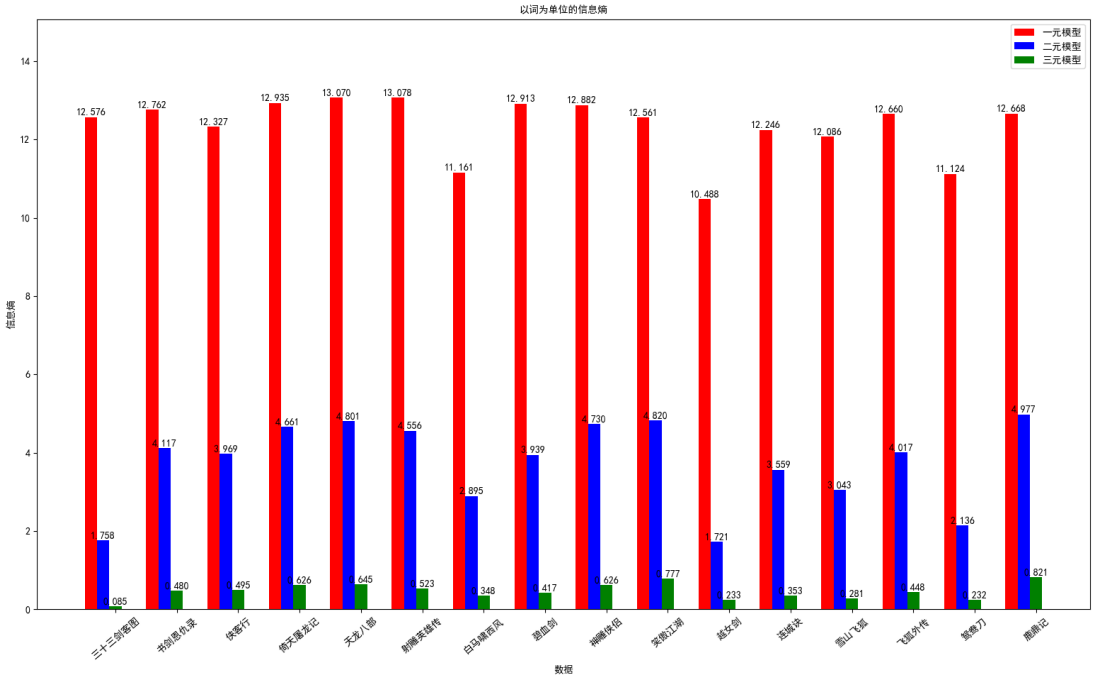

图 3：基于一元、二元和三元模型计算中文平均信息熵（以词为单位）

对比一元模型、二元模型和三元模型可以看到，无论是以字为单位还是以词为单位，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。

Conclusions

本文首先分别介绍了齐夫定律、信息熵和 N 元语言模型的定义，并使用金庸的十六部经典武侠小说作为语料库，验证了齐夫定律并基于一元、二元和三元语言模型分别计算了以字/词为单位中文平均信息熵，并针对各实验结果进行了分析。

References

[1] Brown P. F., Della Pietra S. A. et al.(1992), An estimate of an upper bound for the entropy of English[J].Computational Linguistics, Vol. 18: 1: pp. 31-40.

[2] https://zhuanlan.zhihu.com/p/658563402?utm_id=0