**The Second Report of Deep Learning for Natural Language Processing**

Yue Zhao

zy2303607@buaa.edu.cn

**Abstract**

This paper explores the application of Zipf's Law and the concept of information entropy within the context of the Chinese language, employing both word-based and character-based N-Gram models. Zipf's Law, reflecting the inverse relationship between word frequency and rank, is validated through empirical analysis of a Chinese corpus. Additionally, the study delves into calculating the average information entropy for Chinese, illustrating the impact of context length (N) on the diversity of word combinations and the simplicity of text structure.

**Introduction**

**Zipf's Law**, proposed by the American scholar G.K. Zipf in the 1940s, reveals an inverse relationship between the frequency of word usage and its rank in natural languages: the most frequently used words appear with high frequency, while the frequency of other words decreases as their rank increases. This law is not only applicable in the field of linguistics but also prevalent in various domains such as urban population and corporate size, reflecting universal principles of information organization and social structure. The study of Zipf's Law offers significant insights into understanding the efficiency of language, distribution of information, and the organization of complex systems.

1. 主题模型（Topic Model）

对于一个文档集合来说，假如一篇文章是讲猫科动物的，那么可能会一部分讲猫， 一部分讲老虎， 一部分讲猎豹。那么讲猫的那一部分与猫有关的词语出现的频率应该高一些，比如“鱼”，“老鼠”等，讲老虎的那一部分与老虎有关的词语出现的频率应该高些，比如“森林之王”，“一山不容二虎”等，讲猎豹那一部分与猎豹有关的词语出现的频率应该高些，比如"速度"，“豹纹”等等。所以一篇文档应该有多个主题，每个主题的比例不同，每一个主题下面也应该有很多词语，每个词语的比例也不同。

主题模型就是用数学框架来体现出文档的这种特点，主题模型自动分析每篇文档，统计文档内的词语，根据统计的信息来断定当前文档含有哪些主题，以及每个主题所占的比例各为多少。

从上面的定义可以看出，主题模型其实主要在学习两个分布，文档-主题分布（doc-topic）和主题-词分布（topic-word）。既然是分布就要满足两个条件，第一是非负性，第二是积分或者求和为1。也就是doc-topic矩阵或topic-word矩阵中，任意一行元素均为非负数且元素和为1。

————————————————

版权声明：本文为博主原创文章，遵循 CC 4.0 BY-SA 版权协议，转载请附上原文出处链接和本声明。

原文链接：https://blog.csdn.net/weixin\_50891266/article/details/116273153

1、定义

**LDA（Latent Dirichlet Allocation）**是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。

所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA是一种非监督机器学习技术，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

2、生成过程

对于语料库中的每篇文档，LDA定义了如下生成过程：

1.对每一篇文档，从主题分布中抽取一个主题；

2.从上述被抽到的主题所对应的单词分布中抽取一个单词；

3.重复上述过程直至遍历文档中的每一个单词。

语料库中的每一篇文档与T（通过反复试验等方法事先给定）个主题的一个多项分布 相对应，将该多项分布记为θ。每个主题又与词汇表中的V个单词的一个多项分布相对应，将这个多项分布记为φ。

3、LDA整体流程

先定义一些字母的含义：文档集合D，主题（topic)集合T

D中每个文档d看作一个单词序列<w1,w2,…,wn>，wi表示第i个单词，设d有n个单词。（LDA里面称之为wordbag，实际上每个单词的出现位置对LDA算法无影响）

·D中涉及的所有不同单词组成一个大集合VOCABULARY，LDA以文档集合D作为输入，希望训练出的两个结果向量（设聚成k个topic，VOC中共包含m个词）：

·对每个D中的文档d，对应到不同Topic的概率θd<pt1,…,ptk>，其中，pti表示d对应T中第i个topic的概率。计算方法是直观的，pti=nti/n，其中nti表示d中对应第i个topic的词的数目，n是d中所有词的总数。

·对每个T中的topict，生成不同单词的概率φt<pw1,…,pwm>，其中，pwi表示t生成VOC中第i个单词的概率。计算方法同样很直观，pwi=Nwi/N，其中Nwi表示对应到topict的VOC中第i个单词的数目，N表示所有对应到topict的单词总数。

LDA的核心公式如下：

p(w|d)=p(w|t)\*p(t|d)

直观的看这个公式，就是以Topic作为中间层，可以通过当前的θd和φt给出了文档d中出现单词w的概率。其中p(t|d)利用θd计算得到，p(w|t)利用φt计算得到。

实际上，利用当前的θd和φt，我们可以为一个文档中的一个单词计算它对应任意一个Topic时的p(w|d)，然后根据这些结果来更新这个词应该对应的topic。然后，如果这个更新改变了这个单词所对应的Topic，就会反过来影响θd和φt。

4、LDA学习过程

下面介绍一种LDA的学习过程：

LDA算法开始时，先随机地给θd和φt赋值（对所有的d和t）。然后上述过程不断重复，最终收敛到的结果就是LDA的输出。再详细说一下这个迭代的学习过程：

1.针对一个特定的文档ds中的第i单词wi，如果令该单词对应的topic为tj，可以把上述公式改写为：

pj(wi|ds)=p(wi|tj)\*p(tj|ds)

2.现在我们可以枚举T中的topic，得到所有的pj(wi|ds)，其中j取值1~k。然后可以根据这些概率值结果为ds中的第i个单词wi选择一个topic。最简单的想法是取令pj(wi|ds)最大的tj（注意，这个式子里只有j是变量），即argmax[j]pj(wi|ds)

3.然后，如果ds中的第i个单词wi在这里选择了一个与原先不同的topic，就会对θd和φt有影响了（根据前面提到过的这两个向量的计算公式可以很容易知道）。它们的影响又会反过来影响对上面提到的p(w|d)的计算。对D中所有的d中的所有w进行一次p(w|d)的计算并重新选择topic看作一次迭代。这样进行n次循环迭代之后，就会收敛到LDA所需要的结果了。

---------

1、LDA描述

LDA（Linear Discriminant Analysis），是一种文档主题生成模型，，它可以将文档中每篇文档的主题按照概率分布的形式给出。也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA是一种非监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

LDA的核心思想是寻找到最佳的投影方法，将高维的样本投影到特征空间(feature space)，使得不同类别间的数据“距离”最大，而同一类别内的数据“距离”最小。

2、LDA模型生成

首先定义文章集合为D o c DocDoc,文章主题集合为T o p i c TopicTopic，D o c DocDoc中的每个文档d o c docdoc可以看作为一个单词序列< w 1 , w 2 , . . . , w n > <w\_1,w\_2,...,w\_n><w

表示为第i ii个单词，d o c docdoc共有n nn个单词。

D o c DocDoc中的所有不同单词组成一个集合V o c VocVoc，LDA模型以文档集合D o c DocDoc作为输入，最终训练处两个结果向量，k kk表示Topi词，m mm表示V o c VocVoc中包含的词语数量。

对每个D o c DocDoc中对应到不同T o p i c TopicTopic的概率θ d = < p t 1 , . . . , p t k > \theta\_d=<p\_{t\_1},...,p\_{t\_k}>θ

，其中p t i p\_{t\_i}p

表示d o c docdoc对应T o p i c TopicTopic中第i ii个Topic词的概率。

其中

表示d o c docdoc中对应的第i ii个Topic的词的数目，n nn表示d o c docdoc中所有词的总数。

对每个T o p i c TopicTopic中的Topic，生成不同单词的概率φ t = < p w 1 , . . . , p w m > \varphi\_t=<p\_{w\_1},...,p\_{w\_m}>φ

，其中p w i p\_{w\_i}p

表示t tt生成V o c VocVoc中的第i ii个单词的概率。

其中

表示对应到Topic的V o c VocVoc中的第i ii个单词的数目，N NN表示所有对应到Topic的单词总数。

LDA 的核心公式如下所示：

P ( 词 ∣ 文 档 ) = P ( 词 ∣ 主 题 ) ∗ P ( 主 题 ∣ 文 档 ) P(词|文档)=P(词|主题)\*P(主题|文档)

P(词∣文档)=P(词∣主题)∗P(主题∣文档)

即

P ( w ∣ d ) = P ( w ∣ t ) ∗ P ( t ∣ d ) P(w|d)=P(w|t)\*P(t|d)

P(w∣d)=P(w∣t)∗P(t∣d)

公式以Topic作为中间层，通过当前的θ d \theta\_dθ

给出了文档d dd中出现单词w ww的概率。其中的P ( t ∣ d ) P(t|d)P(t∣d)可通过θ d \theta\_dθ

计算得到，P ( w ∣ t ) P(w|t)P(w∣t)利用φ t \varphi\_tφ

计算得到。因此，我们利用当前的θ d \theta\_dθ

，我们可以为一个文档中的单词计算它对应任意一个Topic时的P ( w ∣ d ) P(w|d)P(w∣d)值，然后根据这些结果来更新这个词对应的Topic。相对应的，如果这个更新改变了这个单词所对应的Topic值，反过来也会影响θ d \theta\_dθ

**SVM介绍**

支持向量机（Support Vector Machine, SVM）是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面 。

SVM使用铰链损失函数计算经验风险并在求解系统中加入了正则化项以优化结构风险，是一个具有稀疏性和稳健性的分类器 。SVM可以通过核方法行非线性分类，是常见的核学习

**欧氏距离**

根据这些概率结果，区分每篇待分类文章究竟来自哪一本小说，此处采用的是**欧式距离的方式**，即比较待分类文章与已知的小说，两者对于各个Topic的概率向量之间的距离最近，即认为是来自该本小说。

**Methodology**

**M1: Zipf's Law**

Zipf's Law is a law of word frequency distribution. It can be articulated as follows: If one compiles the frequency of each word's occurrence in a lengthy text, arranges them in a descending order with high-frequency words at the beginning and low-frequency words following, and assigns natural numbers as rank indices to these words, such that the word with the highest frequency is assigned rank 1, the next highest frequency rank 2, and so on, until the word with the lowest frequency is assigned rank D. Letting  represent frequency and  represent the rank index, the relationship is as shown in equation (1).



Where  is a constant.

**M2: Average Information Entropy of Chinese**

The concept of information entropy was first introduced by Claude Shannon (1916-2001) in 1948, drawing on the concept of "thermal entropy" from thermodynamics, aimed at representing the uncertainty of information. The higher the entropy value, the greater the degree of uncertainty in the information.

According to [1], for text , its definition of information entropy is given by Equation (2).



According to the Law of Large Numbers, when the sample size is sufficiently large, the probability of occurrence of words, bigrams, and trigrams approximates their frequency of occurrence.

Thus, the information entropy formula for the unigram model is given by Equation (3),



where  can be approximated by the frequency of each word in the corpus.

The information entropy formula for the bigram model is given by Equation (4),



where the joint probability  can be approximated as the frequency of occurrence of each bigram in the corpus, and the conditional probability  can be approximated as the ratio of the frequency of occurrence of each bigram in the corpus to the frequency of bigrams starting with the first word of the bigram.

The information entropy formula for the trigram model is given by Equation (5),



where the joint probability  can be approximated as the frequency of occurrence of each trigram in the corpus, and the conditional probability  can be approximated as the ratio of the frequency of occurrence of each trigram in the corpus to the frequency of trigrams starting with the first two words of the trigram.

**Experimental Studies**

**E1: Verify Zipf's Law**

Based on the aforementioned principle and the available data, code is developed to utilize a Chinese corpus for the validation of Zipf's Law, yielding Figures 1 as results.

1、语料处理

题目要求均匀抽取1000个段落（每个段落可以有 K 个 token, K 可以取20，100，500, 1000, 3000）， 每个段落的标签就是对应段落所属的小说。

对给定语料库进行分析可知，语料库内共有16篇文章，从每一篇文章内抽取63个段落，共有1008个段落；对每篇文章分词之后，将一篇文章的总词数除以63，即每篇文章共有63个区间，所选取的段落为每个区间抽取前K个词。

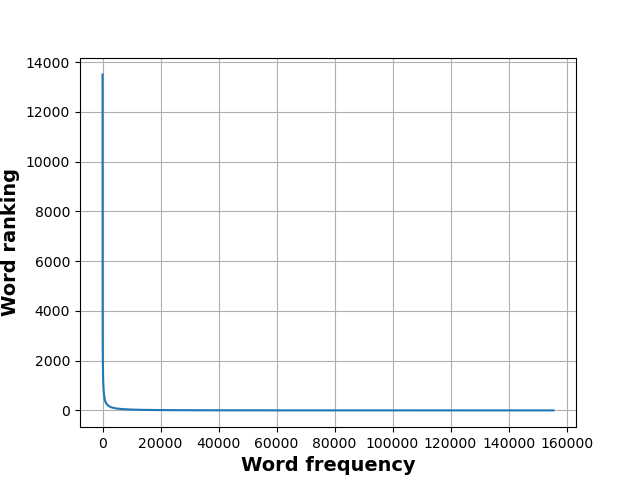


Figure 1：The curve of Word ranking-Word frequency

In Figure 1, we observe that due to the extensive range and dramatic variations in the data, it is challenging to intuitively determine whether the relationship between rank and frequency conforms to Zipf's Law.

To address this issue, we take the logarithm of both sides of Equation (1), resulting in Equation (6),



where theoretically, the relationship between rank and frequency should present a linear correlation.

Redrawing the graph according to Equation (6) yields Figure 2, where the curve approximates a straight line, indicating a near-linear relationship between rank and frequency, thereby validating Zipf's Law.

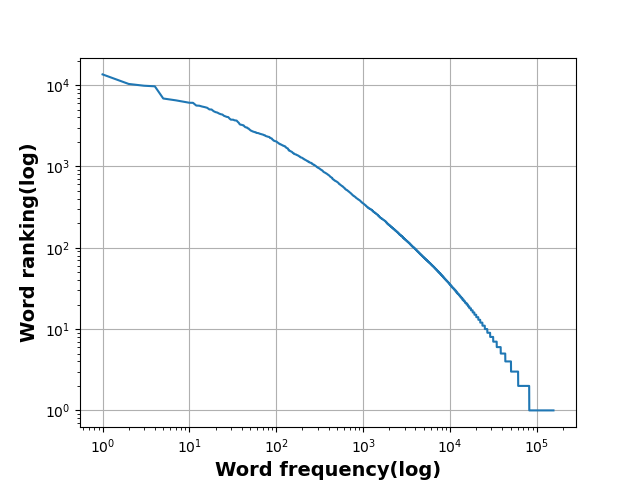


Figure 2：The logarithmic curve of Word ranking-Word frequency

**E2: Calculate the Average Information Entropy of Chinese**

**Preprocessing of the Chinese corpus:** The corpus in the database is in txt format, which includes portions of code and irrelevant symbols. Initially, it is necessary to preprocess the data, which involves removing hidden characters (such as newline characters, page breaks, etc.); eliminating irrelevant information and characters obtained through web crawling; and deleting punctuation marks from the text, as punctuation does not contribute meaningfully to the information in a Chinese corpus.

**Average Information Entropy of Chinese by word basis:** Following the preprocessing of the Chinese corpus, code is written based on the principles introduced in Section M2 to calculate the average information entropy of Chinese under the N-Gram model. The statistical language model employed in this paper is based on words, necessitating the segmentation of Chinese sentences into words. For this purpose, the jieba segmentation system—a Chinese word segmentation system for Python—is utilized to tokenize the sentences, with the results summarized in Table 1.

Table 1: Average Information Entropy of Chinese by word basis

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Combined-gram, Frequency | | |
| 1-Gram | 2-Gram | 3-Gram |
| 1 | '的', 115604 | '道你', 5738 | '只听得', 1613 |
| 2 | '了', 104527 | '叫道', 5009 | '忽听得', 1140 |
| 3 | '他', 64712 | '道我', 4953 | '站起身来', 731 |
| 4 | '是', 64457 | '笑道', 4271 | '哼了一声', 580 |
| 5 | '道', 58623 | '听得', 4202 | '笑道你', 572 |
| 6 | '我', 57483 | '都是', 3905 | '吃了一惊', 539 |
| 7 | '你', 56679 | '了他', 3638 | '啊的一声', 523 |
| 8 | '在', 43691 | '他的', 3497 | '点了点头', 505 |
| 9 | '也', 32606 | '也是', 3201 | '说到这里', 476 |
| 10 | '这', 32199 | '的一声', 3102 | '了他s的', 459 |
| Total number of combined-grams | 4267805 | 4208528 | 4149591 |
| The number of different combined-grams | 172209 | 1940075 | 3448842 |
| Entropy  (bits per word) | 12.180902382994336 | 6.932954107965453 | 2.2941584704204274 |

From Table 1, comparing the results obtained from the 1-Gram, 2-Gram, and 3-Gram language models, it is evident that the larger the value of N, meaning the greater the length of context considered, the more numerous the distinct words that appear (corresponding to the “The number of different combined-grams” in the table). This is attributed to the fact that as the length increases, so does the number of combinations of characters forming words, resulting in a greater variety of distinct words.

Furthermore, a comparison among the three models reveals that as the value of N increases, the information entropy of the text decreases. This is reasoned to be due to the fact that, with larger values of N, the distribution of word groups in the text obtained after segmentation becomes simpler. Upon analysis, it is found that with larger N, the number of possible fixed words that can be formed is fewer. Fixed words reduce the chances of characters or short words disrupting the coherence of the article, thereby making the article more ordered. The uncertainty involved in forming words from characters and sentences from words decreases, consequently lowering the text's information entropy.

**Average Information Entropy of Chinese by character basis:** Furthermore, by segmenting Chinese sentences into characters as units, the average information entropy of Chinese under the N-Gram model is recalculated, with the results summarized in Table 2.

Table 2: Average Information Entropy of Chinese by character basis

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Combined-gram, Frequency | | |
| 1-Gram | 2-Gram | 3-Gram |
| 1 | '一', 139396 | '说道', 13525 | '韦小宝', 9803 |
| 2 | '不', 134149 | '了一', 12174 | '令狐冲', 5889 |
| 3 | '的', 121668 | '一个', 10571 | '张无忌', 4645 |
| 4 | '是', 112707 | '自己', 10319 | '的一声', 3478 |
| 5 | '了', 111916 | '道你', 10262 | '袁承志', 3037 |
| 6 | '道', 111055 | '小宝', 9942 | '小宝道', 2417 |
| 7 | '人', 84302 | '韦小', 9856 | '陈家洛', 2115 |
| 8 | '他', 73573 | ' 也不', 9304 | '小龙女', 2081 |
| 9 | '这', 68993 | '道我', 8472 | '石破天', 1818 |
| 10 | '我', 67000 | '笑道', 8140 | '不由得', 1803 |
| Total number of combined-grams | 7299807 | 7240491 | 7181175 |
| The number of different combined-grams | 5782 | 731519 | 3173250 |
| Entropy  (bits per character) | 9.53836558846301 | 6.715805200021205 | 3.9380548669875464 |

Compared to Table 1, under the same underlying principles, Table 2 exhibits a greater number of word combinations for each N-Gram model. This increase is attributable to the fact that, in this instance, each word in the combinations consists of only a single character. Consequently, the length of word combinations derived from segmentation by characters does not exceed that of combinations derived from segmentation by words, resulting in a higher overall count. For this reason, segmenting text into combinations of characters is less simple and efficient than using combinations of words. This disadvantage becomes more pronounced with the increase of N. Hence, the difference in information entropy for the 3-Gram model between Tables 2 and 1 is significant.

**Conclusions**

The investigation confirms Zipf's Law's applicability to the Chinese corpus, with graphical representations substantiating the expected linear relationship between word rank and frequency on a logarithmic scale. The study further reveals that as the length of context considered increases (higher N values in N-Gram models), not only does the variety of distinct word combinations augment, but also the information entropy of the text decreases, indicating a transition towards a more ordered and predictable structure. This reduction in entropy is more pronounced when analyzing the text on a character basis, due to the increased combination possibilities offered by individual characters. These findings underscore the significant impact of context length on the distribution and predictability of language, providing valuable insights into the efficiency and organization of linguistic information, as well as offering implications for the fields of computational linguistics, information theory, and language modeling.

**References**

[1] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C., & Mercer, R. L. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, *18*(1), 31-40.