Report of Deep Learning for Natural Langauge Processing

Baobao He
hebaobao@buaa.edu.cn

Abstract

本文通过对金庸的 16 部小说进行语料分析，验证了中文词频分布满足齐夫定律，并采用一元、二元和三元字词统计方法，计算了中文平均信息熵。

Introduction

齐夫定律（Zipf's law）是由哈佛大学的语言学家乔治·金斯利·齐夫于 1949 年发表的实验定律。它可以表述为：在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。所以，频率最高的单词出现的频率大约是出现频率第二位的单词的 2 倍，而出现频率第二位的单词则是出现频率第四位的单词的 2 倍。这个定律被作为任何与幂定律概率分布有关的事物的参考。

在信息论中，熵是接收的每条消息中包含的信息的平均量，又被称为信息熵、信源熵、平均自信息量。这里，“消息”代表来自分布或数据流中的事件、样本或特征。（熵最好理解为不确定性的量度而不是确定性的量度，因为越随机的信源的熵越大。）来自信源的另一个特征是样本的概率分布。这里的想法是，比较不可能发生的事情，当它发生了，会提供更多的信息。由于一些其他的原因，把信息（熵）定义为概率分布的对数的相反数是有道理的。事件的概率分布和每个事件的信息量构成了一个随机变量，这个随机变量的均值（即期望）就是这个分布产生的信息量的平均值（即熵）。熵的单位通常为比特，但也用 Sh、nat、Hart 计量，取决于定义用到对数的底。

Methodology

Zipf's law

利用 jieba 库对中文小说进行词频统计，验证 Zipf-Law。具体过程如下：使用 jieba 将中文文本分割成；计算每个词的出现次数；删除标点符号和其他不需要的字符；按频率对词语进行排序；将排序后的词及其频率写入文本文件；使用 Matplotlib 可视化 Zipf 定律。

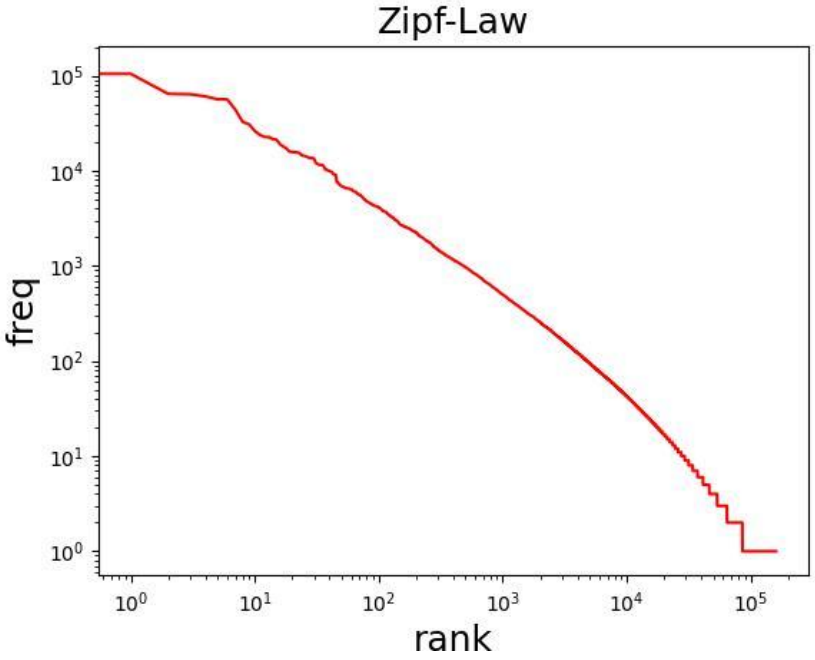

Figure 1：Zipf's law

Information Entropy

读取指定文件中的中文文本内容，并进行预处理，包括去除特定字符串和停用词，然后进行分词；计算单个字、单个词、二元词和三元词的词频（TF）；计算一元模型（Unigram Model）、二元模型（Bigram Model）和三元模型（Trigram Model）的信息熵；将处理后的文本内容写入总文件，并将各模型的信息熵输出到日志文件；绘制柱状图，展示不同模型在不同数据库上的信息熵情况，以便进行可视化分析。

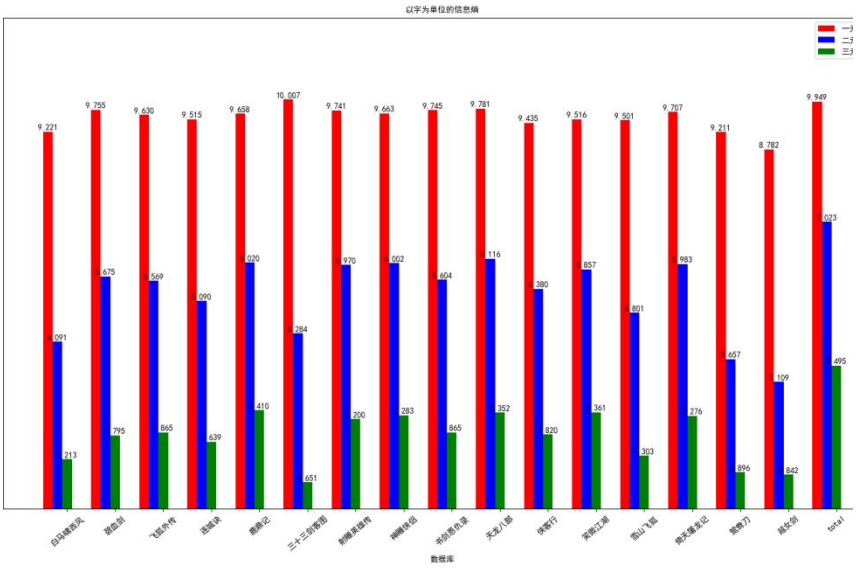

Figure 2：以字为单位的信息熵

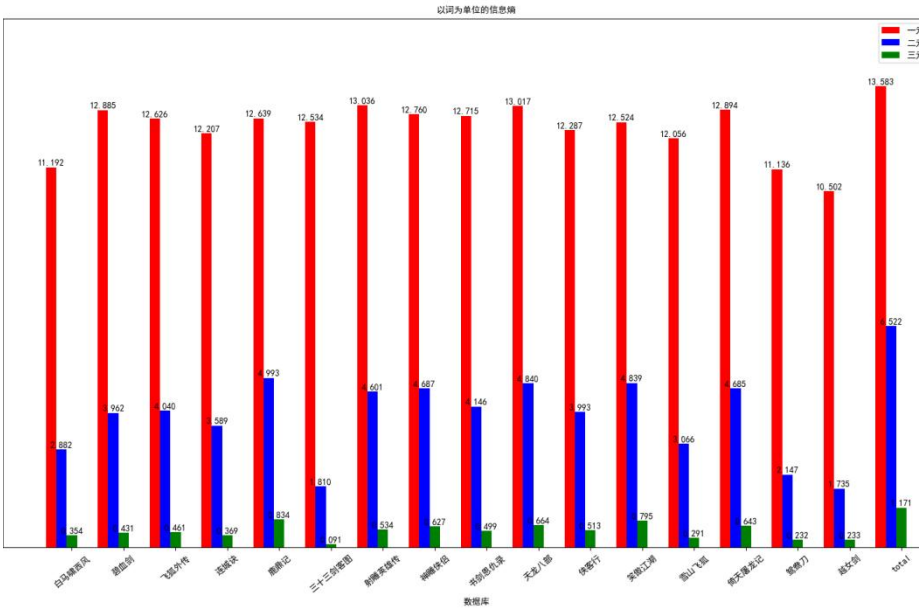

Figure 3：以词为单位的信息熵

Conclusions

对金庸的 16 部作品进行了分词与词频统计分析，并分别计算了字和词的信息熵。验证了中文词频分布满足齐夫定律

无论是一元、二元、三元语言模型，字\词的信息熵在每个作品间的变化趋势是相同的，同时可以发现，一元语言模型信息熵大于二元语言模型，二元语言模型信息熵大于三元语言模型，说明字数越多，表意越精确。值得注意的是，在一元语言模型下，词的信息熵大于字的信息熵，但在二元、三元语言模型中，词的信息熵小于字的信息熵。

References

[1] Alexander Gelbukh, Grigori Sidorov. Zipf and Heaps Laws' Coefficients Depend on Language. Proc. CICLing-2001, Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISSN 0302-9743, ISBN 3-540-41687-0, Springer-Verlag, pp. 332–335.

[2] Brown P F, Della Pietra SA, 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.