Deep NLP 第一次作业

ZB2303019　汪婧伶

Abstract

本报告旨在通过对中文语料库的分析，验证 Zipf 定律，并进一步计算字和词的平均信息熵，以探讨中文文本的信息特征。通过实验结果的分析，加深对 Zipf 定律的理解，并了解中文文本的词频分布规律和信息熵特征，为之后进行自然语言处理和文本挖掘领域的研究提供参考和启示。

Introduction

Zipf 定律是描述自然语言词频分布的经典规律之一，它指出一个词的频率与其在频率排名表上的排名成反比。这一定律不仅适用于英文，也被认为适用于其他语言，包括中文。通过使用中文语料库，验证文本是否符合 Zipf 定律。另一方面，信息熵是衡量信息内容不确定性的指标，对于文本的信息量和复杂度有着重要的意义。因此，通过计算字和词的平均信息熵，可以更全面地了解文本的信息特征。

Methodology

M1：使用中文语料库，验证 Zipf’s Law

Zipf's Law（即齐普夫定律）是由美国语言学家乔治·金斯利·齐普夫（George Kingsley Zipf）在 20 世纪早期提出的经验定律，用于描述自然语言中单词使用频率与其排名的关系。齐普夫定律的核心观点是：在自然语言中，排名第 n 位的词的频率与其排名的倒数成反比关系。

具体来说，如果将语料库中的所有词按照使用频率从高到低排列，那么排名第 n 位的词的频率 f(n)将约等于一个常数 k 除以排名 n：

$$f(n) = \frac{k}{n}$$

这里，k 是一个正常数，通常称为 Zipf 系数。齐普夫定律适用于各种自然语言，包括英语、中文、法语等。

齐普夫定律的重要性在于它揭示了自然语言中单词使用的不平衡性。尽管语言中有成千上万个词，但只有很少一部分词被广泛使用，而大多数词很少被使用。这种不平衡性对于语

言理解、信息检索、自然语言处理等领域都有重要的影响。

验证齐普夫定律的一种方法是通过绘制词频与排名的对数图，如果数据点近似形成一条直线，那么可以认为该语言遵循齐普夫定律。

M2：计算字和词的平均信息熵

使用信息熵公式计算每个字和每个词的信息熵：

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

Experimental Studies

实验一中，先根据记录了小说名字的 inf.txt 进行小说内容的读取，然后对小说文本进行预处理，去除无关部分。使用分词工具（如 jieba）对文本进行分词处理，将文本拆分成单个词的列表。加载中文停词列表，以排除常见停用词的影响，然后删除停词和非中文字符，保留有效的中文字符。最后统计计算词频，并取词频排名前 1000 名的词绘制对数图显示。由图 1 可发现，数据点近似形成一条直线，可以认为文本遵循齐普夫定律。

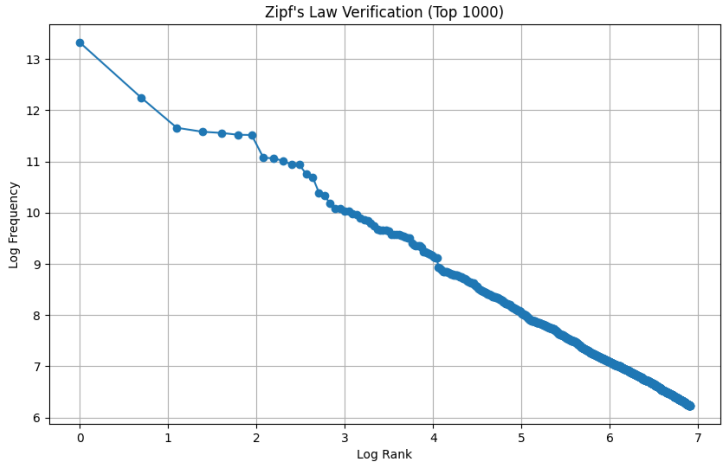

图 1：词频对数图

实验二中，分别计算各个小说文本中的字和词的平均信息熵。发现在给定的文本中，词的平均信息熵比字的平均信息熵要高。可能原因如下：

1. 词的平均信息熵较高可能表示文本中的词汇较为丰富和复杂，其中可能包含了大量具有不同语义的词汇。这可能是因为文本所涉及的主题较为复杂，需要使用多种不同的词汇来表达。
2. 高词单位信息熵可能反映了文本中的信息量较丰富，其中包含了大量不同的概念、想法或描述，需要使用多样化的词汇来表达。这样的文本通常具有较高的信息密度和丰富的内涵。

小说名称	字单位信息熵	词单位信息熵
白马啸西风	8.2973	9.3333
碧血剑	9.0287	10.3935
飞狐外传	8.9048	10.2325
连城诀	8.7262	9.8449
鹿鼎记	8.8131	10.0743
三十三剑客图	9.1948	10.2919
射雕英雄传	8.9612	10.4753
神雕侠侣	8.9426	10.4035
书剑恩仇录	9.0072	10.2879
天龙八部	8.9434	10.3475
侠客行	8.7372	9.9736
笑傲江湖	8.8102	10.0862
雪山飞狐	8.7837	9.9196
倚天屠龙记	8.9700	10.4142
鸳鸯刀	8.4264	9.4145
越女剑	8.2321	9.1296

图 2：各小说文本的字/词平均信息熵

Conclusions

通过实验，我对自然语言处理有了初步了解，并加深了对 Zipf's Law 的理解，学习了分析文本字、词的平均信息熵及其结果原因分析。