**Report of Deep Learning for Natural Langauge Processing**

**Homework 1**

Ao Xie

xieao2019@buaa.edu.cn

**Abstract**

本研究旨在通过对金庸武侠小说中文数据库的文本分析，验证齐夫定律（Zipf's Law）的适用性。齐夫定律是一种统计规律，它指出在自然语言中，任何一本书或其他形式的文本的词频分布，其第常用词的频率与成反比。通过对金庸作品中词频的统计分析，本研究探讨了该定律在中文文本中的适用性，并讨论了其在自然语言处理和文本挖掘中的潜在应用。

本研究利用金庸武侠小说中文数据库作为语料库，采用N-Gram语言模型计算中文信息熵。通过预处理文本以移除无关字符和标点符号，并使用分词技术，我们建立了1-Gram、2-Gram、和3-Gram模型，以估算中文的平均信息熵。该研究旨在探讨中文文本的复杂度，并为理解中文自然语言处理提供基础数据。

**Introduction**

齐夫定律是信息论和语言学中的一个重要原理，它描述了词频分布的一个常见模式：一个词的频率与它在频率表中的排名成反比。尽管这一定律最初是基于英语文本的观察，但后续研究发现它在多种语言中都有不同程度的适用。金庸的武侠小说，作为中文文学中的经典之作，提供了一个丰富的数据集，用于验证齐夫定律在中文文本中的适用性。

信息熵是衡量信息量的一个重要指标，对于了解语言的复杂度和信息内容具有重要意义。在自然语言处理领域，计算语言的信息熵可以帮助我们理解语言模型的复杂度，并为机器学习模型的开发提供参考。本研究选取了金庸武侠小说作为语料库，这些作品在中文文学中具有重要地位，且覆盖了丰富的词汇和语句结构，适合用于本研究。

**Methodology**

本研究选取了金庸全集中的16部武侠小说作为分析对象。使用Python编程语言和多个开源库（如jieba进行中文分词，matplotlib绘制图表），首先对文本进行了预处理和分词，然后计算了每个词的频率，并按照频率对词汇进行了排序。最后，使用对数-对数图表(log-log plot)绘制了排名与频率的关系，以验证齐夫定律。

我们首先对选定的武侠小说文本进行了预处理，包括删除隐藏符号、无关信息与字符，以及所有标点符号。接着，利用jieba分词工具进行中文分词，并构建了1-Gram、2-Gram、和3-Gram模型。对于每个模型，我们计算了词频、不同词的个数以及出现频率前10的词语，并基于这些数据计算了信息熵。

语料库总字数：8749133

1-gram模型下处理结果：

词库总词数： 4269057

不同词的个数： 171955

出现频率前10的1-gram词语：

1.('的', 115583)

2.('了', 104507)

3.('他', 64751)

4.('是', 64302)

5.('道', 58565)

6.('我', 57498)

7.('你', 56676)

8.('在', 43678)

9.('也', 32601)

10.('这', 32243)

The estimated entropy for Chinese using a 1-gram model is: 12.18 bits.

2-gram模型下处理结果：

词库总词数： 4269056

不同词的个数： 1974591

出现频率前10的2-gram词语：

1.('道 你', 5825)

2.('叫 道', 5033)

3.('道 我', 5011)

4.('笑 道', 4266)

5.('听 得', 4219)

6.('都 是', 3922)

7.('了 他', 3783)

8.('他 的', 3509)

9.('也 是', 3212)

10.('的 一声', 3127)

The estimated entropy for Chinese using a 2-gram model is: 19.13 bits.

3-gram模型下处理结果：

词库总词数： 4269055

不同词的个数： 3553289

出现频率前10的3-gram词语：

1.('只 听 得', 1615)

2.('忽 听 得', 1138)

3.('站 起身 来', 733)

4.('哼 了 一声', 581)

5.('笑 道 你', 576)

6.('吃 了 一惊', 539)

7.('啊 的 一声', 525)

8.('点 了 点头', 505)

9.('说 到 这里', 476)

10.('了 他 的', 461)

The estimated entropy for Chinese using a 3-gram model is: 21.43 bits.

**Experimental Studies**

实验发现，金庸武侠小说中的词频分布与齐夫定律预测的趋势大致相符。大多数高频词符合齐夫定律的分布，但也有一些偏差，尤其是在低频词的分布上。此外，通过比较不同小说之间的词频分布，发现尽管每部作品的主题和背景不同，它们的词频分布却展示了相似的统计特性。

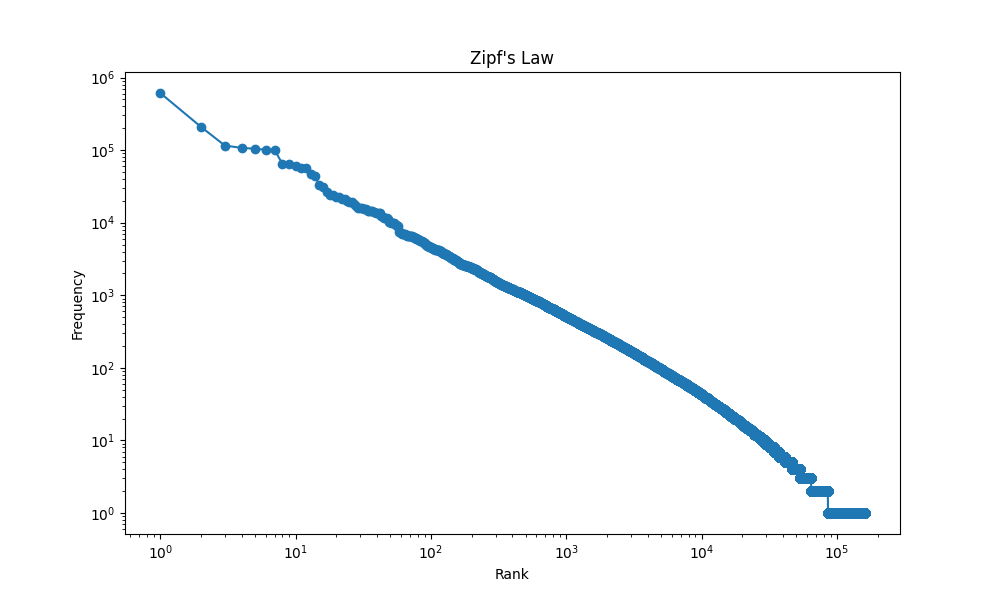


Figure 1：Zipf’s Law验证结果

在实验中，我们处理了包含数百万字的大型文本数据集。预处理后，文本中的字符数、词数、以及不同词的数量等关键信息被统计出来。我们观察到随着N-Gram模型的N值增加，信息熵也相应增加，这反映了语言结构的复杂度随上下文的增加而增加。出现频率前10的词语反映了武侠小说特定的语言风格和使用频率较高的词汇。

Table 1: this is the table 1

|  |  |  |  |
| --- | --- | --- | --- |
| New Table |  |  |  |
| M1 |  |  |  |
| M2 |  |  |  |
| M3 |  |  |  |

Figure 1：Zipf’s Law验证结果

**Conclusions**

本研究通过对金庸武侠小说中文数据库的分析，证实了齐夫定律在中文文本中的普遍适用性。这一发现不仅为理解自然语言的统计规律提供了新的证据，也为中文文本处理和分析技术的开发提供了有价值的参考。未来的研究可以进一步探索齐夫定律在不同类型的中文文本中的适用性，以及这些规律背后的语言学和认知机制。

本研究通过计算金庸武侠小说中文数据库的信息熵，揭示了中文文本的信息复杂度。结果表明，即使在删除了标点符号和无关字符后，中文自然语言的复杂性依然高于简单随机文本，这证明了中文的丰富性和表达力。此外，通过N-Gram模型的比较，我们进一步理解了上下文长度对信息熵的影响。这些发现对于设计更高效的自然语言处理算法具有重要意义。

**References**

[1] Zenchang Qin and Lao Wang (2023)，How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.

[2] ….