Report of Deep Learning for Natural Langauge Processing

Homework 1

Ao Xie
xieao2019@buaa.edu.cn

Abstract

本次作业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）的适用性并采用 N-Gram 语言模型计算中文信息熵。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 r 常用词的频率与 r 成反比。通过对金庸作品中词频的统计分析，本次作业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。同时，本次作业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，建立了 1-Gram、2-Gram、和 3-Gram 模型，以估算中文的平均信息熵。本次作业旨在探讨中文文本的复杂度，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。本次作业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，适合用于本次作业。

Methodology

本次作业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。

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$1/r$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。

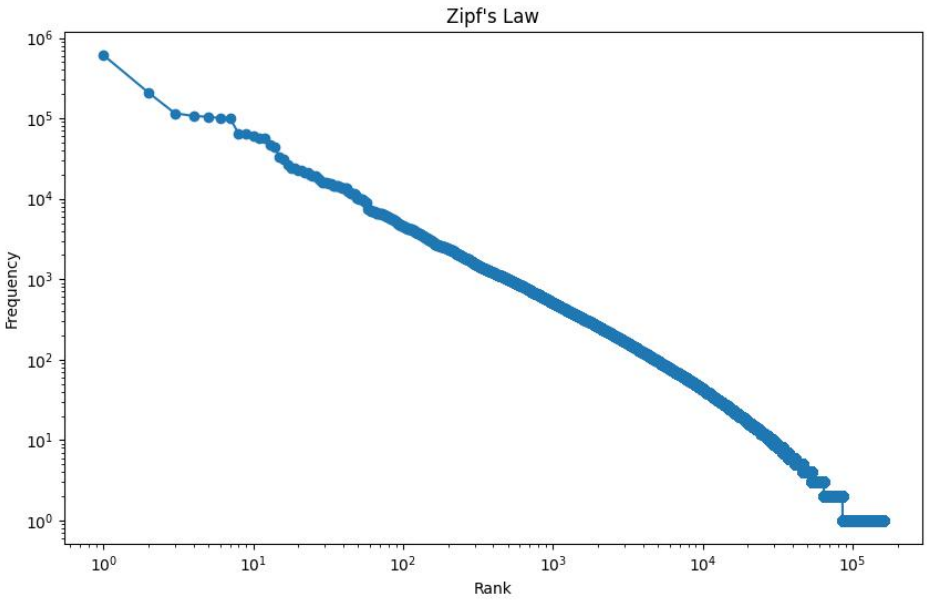

Figure 1：Zipf’s Law 验证结果

在实验中，我处理了包含数百万字的大型文本数据集。在进行文本预处理后，统计了文本中的字符数、词数、以及不同词的数量等关键信息。由实验结果，我发现随着 N 值的增加，词库中不同词的个数显著增加，展示了中文文本在更长的上下文中表现出的丰富性和多样性。同时，信息熵的逐渐降低反映了随着上下文长度的增加，语言模型的预测能力得到了显著提升。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: n-gram 模型下处理结果

出现频率前 10 的 1-gram 词语										
序号	1	2	3	4	5	6	7	8	9	10
词语	的	了	他	是	道	我	你	在	也	这
词频	115583	104507	64751	64302	58565	57498	56676	43678	32601	32243
出现频率前 10 的 2-gram 词语										
序号	1	2	3	4	5	6	7	8	9	10
词语	道你	叫道	道我	笑道	听得	都是	了他	他的	也是	的一声
词频	5825	5033	5011	4266	4219	3922	3783	3509	3212	3127
出现频率前 10 的 3-gram 词语										
序号	1	2	3	4	5	6	7	8	9	10
词语	只听得	忽听得	站起身来	哼了一声	笑道你	吃了一惊	啊的一声	点了点头	说到这里	了他的
词频	1615	1138	733	581	576	539	525	505	476	461

Table 2: 信息熵计算结果

语料库	**1-gram**		**2-gram**		**3-gram**	
总字数	**不同词的个数**	**信息熵**	**不同词的个数**	**信息熵**	**不同词的个数**	**信息熵**
8749133	171955	12.18	1974591	6.95	3553289	2.3

Conclusions

本次作业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。同时，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，随着考虑的上下文长度（N-Gram 模型的 N 值）增加，中文文本的信息熵显著降低，这说明了在更长的上下文中，文本的不确定性降低，语言模型能更准确地预测接下来的词语。让我对自然语言处理的理解更加深刻。