Report of Deep Learning for Natural Language Processing

Siyuan Jin
20373680@buaa.edu.cn

Introduction

Zipf's Law

Zipf's Law（齐夫定律）是由哈佛大学的语言学家乔治 · 金斯利 · 齐夫（George Kingsley Zipf）于 1949 年发表的实验定律。该定律表明，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，如公式1所示。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。

$$\text{word freuency} \propto \frac{1}{\text{word rank}} \tag{1}$$

jieba 分词

jieba 是一个开源的中文分词工具，它使用了基于前缀词典实现的分词算法。该工具将中文文本作为输入，并尝试将其切分成有意义的词语，以便进行后续的文本分析或处理。

它的工作原理基于两种关键技术：字典匹配和概率统计。jieba 首先使用内置的字典进行精确匹配，将文本中的词语与字典中的词条进行比对，尽可能地将文本切分成已知的词语。然后，对于那些无法直接匹配的部分，jieba 会利用概率模型进行分词，根据文本中各个字的出现概率和常见词语的组合频率来进行判断，从而得出最优的切分方案。除了基本的分词功能外，jieba 还支持用户自定义字典，可以通过添加特定词汇来提高分词准确性。

由于其简单易用且效果良好，jieba 被广泛应用于各种中文文本处理任务，如文本分类、情感分析、信息检索等领域。

验证 Zipf's Law

利用提供的中文语料库，通过 jieba 对文本进行分词并统计每个汉字词语的出现频率，然后根据频率对词语进行排序。然后绘制排名与频率的关系图，横轴表示词语的排名，纵轴表示词语的频率，采用对数坐标轴。当图中的数据点大致落在一条直线上，且符合对数关系，则可说明该中文语料库满足 Zipf's Law，亦即可验证 Zipf's Law 的普遍性。

信息熵的概念

信息熵的概念最早由香农在《A mathematical theory of communication》中提出 [1]，他将热力学中熵的概念引入信息论中，将接收到的每条消息中包含的信息的平均量称为信息熵。信息熵在自然语言处理中有广泛的应用，通过对不同语言的信息熵的计算，我们可以得知各种语言每个词平均包含的信息量。如果我们可以准确的计算出某种语言的信息熵，那么我们就能得到这种语言的信息压缩下界。

信息熵的计算

本文将根据 Brown 等 [2] 提出的计算英语信息熵上界的方法计算中文的信息熵。

Methodology

M1：验证 Zipf's Law

1. 导入语料文件。
2. 使用 jieba 分词，统计频率。
3. 降序排序。
4. 绘制关系图（使用对数坐标轴，横坐标为词语的排名，纵坐标为词语的频率）。
5. 如果数据点整体规律在一条直线上，符合对数关系，则验证成功。

M2：计算信息熵

1. 导入语料文件。
2. 数据预处理，去除符号和乱码。
3. 以每个中文汉字或 jieba 分词后的每个中文词汇为单位，基于字或词计算得到中文平均信息熵。

Experimental Studies

M1：验证 Zipf's Law

验证 Zipf's Law 的普遍的实验结果如图1所示，图中蓝色线为未过滤停用词的词频与词语排名关系，绿色线为过滤停用词后的词频与词语排名间的关系，红色线为基准值，可以明显看出无论是否过滤停用词还是过滤停用词，词频与词语排名均明显在一条至直线上，符合对数关系，可以验证 Zipf's Law。

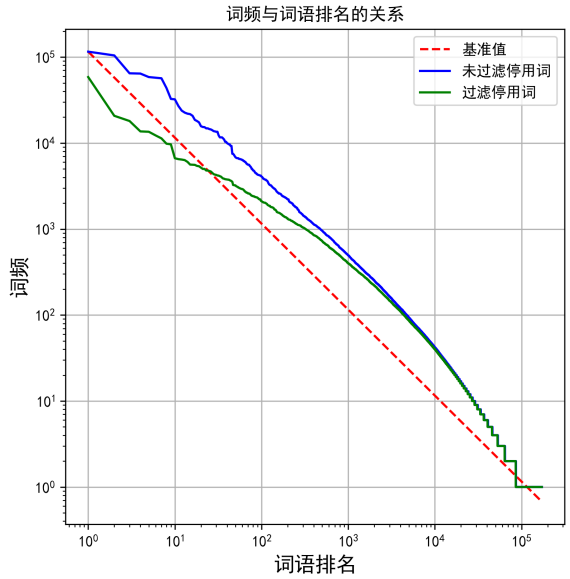

图 1: 验证 Zipf's Law 实验结果

M2：计算信息熵

计算信息熵的整体实验结果如表1所示,具体每本小说的信息熵只列举了《三十三剑客图》(表2)与《三十三剑客图》(表3)，其余因版面问题未列举，详细内容可见附件。

表 1: 语料库整体信息熵

参数	数值
运行时间	30.67
数据库总字数	7258004
数据库分词总个数	4264134
平均词长	1.702
基于字的中文平均信息熵	148.27
基于词的中文平均信息熵	180.67

表 2: 《三十三剑客图》信息熵计算结果

参数	数值
运行时间	0.63
小说总字数	53285
小说分词总个数	31175
平均词长	1.709
基于字的中文平均信息熵	9.67
基于词的中文平均信息熵	11.68

表 3: 《书剑恩仇录》信息熵计算结果

参数	数值
运行时间	1.75
小说总字数	435615
小说分词总个数	253082
平均词长	1.721
基于字的中文平均信息熵	9.46
基于词的中文平均信息熵	11.71

Conclusions

本次作业分为两项任务，第一项任务是验证 Zipf's Law，第二项任务是计算中文的平均信息熵。根据实验一与实验二的结果，两项任务均完成。

参考文献

[1] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[2] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992.