Report of Deep Learning for Natural Langauge Processing

Haolei Zhang
2531563133@qq.com

Abstract

本实验旨在通过实际的中文金庸小说语料库数据，验证著名的 ZipfsLaw，并进一步计算以词和字为基本单位的中文文本的平均信息熵，以揭示语言的内在统计规律及其信息复杂度。

Introduction

Zipf's Law 是语言学中的一个重要定律，它描述了一个词在文本中出现的频率与其排名成倒数的关系。Zipf's Law 表明，在一个大的文本语料库中，一个词的频率与它在频率排名中的位置成反比。这意味着，排名最高的词出现的频率是排名第二的词的两倍，排名第三的词的四倍，以此类推。Zipf's Law 在自然语言处理、信息检索和文本分析等领域有着广泛的应用。它提供了一种简单而有效的方法来描述和预测文本中的词频分布。此外，Zipf's Law 还为语言学研究提供了一种重要的工具，用于探索语言的结构和演变。本报告通过使用提供的金庸小说集作为中文语料库来验证 Zipf's Law。

信息熵是量化语言复杂性的一种重要指标，它起源于信息论，由克劳德·香农提出，用于衡量信息的不确定性或随机性。在自然语言处理中，信息熵可以用来评估一个文本的词汇多样性、信息含量和语言结构的复杂性。一个文本的信息熵越高，表示它的词汇使用越丰富，信息含量越大，语言结构越复杂。本报告的目的是计算中文的平均信息熵，分别以词和字为单位进行分析。

Methodology

准备一个金庸小说集作为中文语料库，为了处理这些文本数据，使用了 jieba 分词工具。jieba 是一个流行的中文分词工具，它能够有效地将中文文本分割成词语。通过分词，可以得到文本中每个词语的出现频率，从而进一步分析词频与排名之间的关系。在中文文本处理中，停用词是一些频繁出现但不含实际意义的词语，如“的”、“了”、“和”等。为了更准确地分析词频与排名的关系，需要从文本中过滤掉这些停用词。为此，创建了一个停用词列表，并将其应用于分词后的文本数据。在过滤掉停用词后，统计了每个词语在文本中的出现频率。这可以通过 Python 的 Counter 类来实现。Counter 类是一个简单的计数器工具，它可以快速统计元素的出现次数。为了直观地观察词频与排名之间的关系，绘制了 Zipf 图。在 Zipf 图中，以排名为横坐标，以词频为纵坐标，将每个词语的频率与排名绘制在图上。通过观察 Zipf 图，可以判断词频与排名之间的关系是否符合 Zipf's Law 的预测。

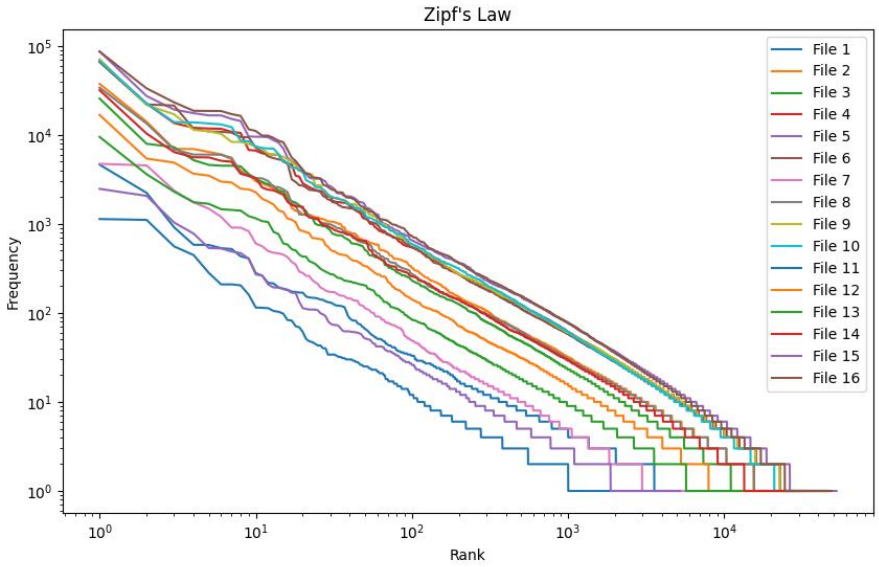
信息熵的计算公式如下：

$$H(X) = -\sum \mathrm{p(x)} \log \mathrm{p(x)}$$

其中，H(X)表示信息熵，p(x)表示随机变量 X 取值为 x 的概率。在文本分析中，可以将每个词语的出现频率视为其概率，从而计算整个文本的信息熵。

为了计算中文的信息熵，分别以词和字为单位进行分析。对于以词为单位的分析，统计了每个词语的出现频率，并计算了整个文本的信息熵。对于以字为单位的分析，将每个字符视为一个词，并统计了每个字符的出现频率，同样计算了整个文本的信息熵。

Experimental Studies



通过绘制 Zipf 图，可以看到词频与排名之间的关系大致符合 Zipf's Law 的预测。

语料库	字单位信息熵	词单位信息熵
白马啸西风	8.279318844	8.768358046
碧血剑	9.012272587	10.35766779
飞狐外传	8.886975216	10.19709589
连城诀	8.709937336	9.702584996
鹿鼎记	8.795503591	9.94054922
三十三剑客图	9.175880897	10.2851387
射雕英雄传	8.943051308	10.27353686
神雕侠侣	8.924081402	10.27064598
书剑恩仇录	8.988460792	10.27008752
天龙八部	8.92504561	10.23335685
侠客行	8.721138406	9.835085141
笑傲江湖	8.795156391	10.04271344
雪山飞狐	8.769212279	9.807626307
倚天屠龙记	8.949696537	10.3287924
鸳鸯刀	8.416950478	8.926473679
越女剑	8.214000986	8.587963638

计算了多部小说的字和词的信息熵，并计算了平均信息熵：平均词信息熵: 9.864229777742839，平均字信息熵: 8.7816676663193。

Conclusions

具体来说，观察到排名较高的词的频率较低，而排名较低的词的频率较高。这与 Zipf's Law 的预测一致。信息熵结果显示，不同文本间的信息熵存在一定的差异，这可能与文本的主题、风格和语言使用有关。