Report of Deep Learning for Natural Langauge Processing

卢田雨
2589402656@qq.com

Abstract

本次 NLP 作业旨在通过一个以金庸小说为内容的中文语料库数据，通过代码实现来验证 Zipf“s Law，在此基础上分别计算以词为单位和以字为单位的中文文本平均信息熵，通过实验有助于加深对语言内在统计规律以及其包含信息复杂程度的理解。

Introduction

本研究以金庸小说为对象，探究其文本数据是否符合 Zipf 定律，并计算信息熵。通过预处理文本、分析词频与排名关系，验证了 Zipf 定律在金庸小说中的适用性，并计算了其信息熵。结果显示，金庸小说的词频分布呈现出明显的幂律分布特征，信息熵较高，反映了其中文本信息的丰富性。本研究对中文语料库特性的理解具有重要意义，也为文本处理与自然语言处理领域提供了实证分析的示例。

Part 1

Zipf‘s Law 是由美国语言学家乔治·金斯利·希普夫（George Kingsley Zipf）提出的一种经验定律，描述了自然语言中词频与词序之间的关系。该定律表明，一个单词在语料库中的频率与其在频率排序表中的排名成反比关系。换言之，频率第 n 高的单词出现的频率大约是频率最高的单词的 1/n。这一定律在很多自然语言的文本中都有显著表现，不仅适用于英语，也适用于其他语言。Zipf 定律的发现对语言学、信息论、文本处理等领域有着重要的理论和实际意义，为语言和文本的研究提供了重要的线索。

根据中文语料库，本文读取金庸小说文本数据，使用 jieba 进行分词和停用词去除，统计词频并排序，绘制词频-排名图，验证 Zipf 定律，最终保存词频数据和图表。

Experiment

使用金庸小说文本，验证 Zipf 定律。通过分词、词频统计，绘制词频-排名图表，词频-排名图表如图 1 所示。结果表明词频与排名呈现幂律分布，符合 Zipf 定律。

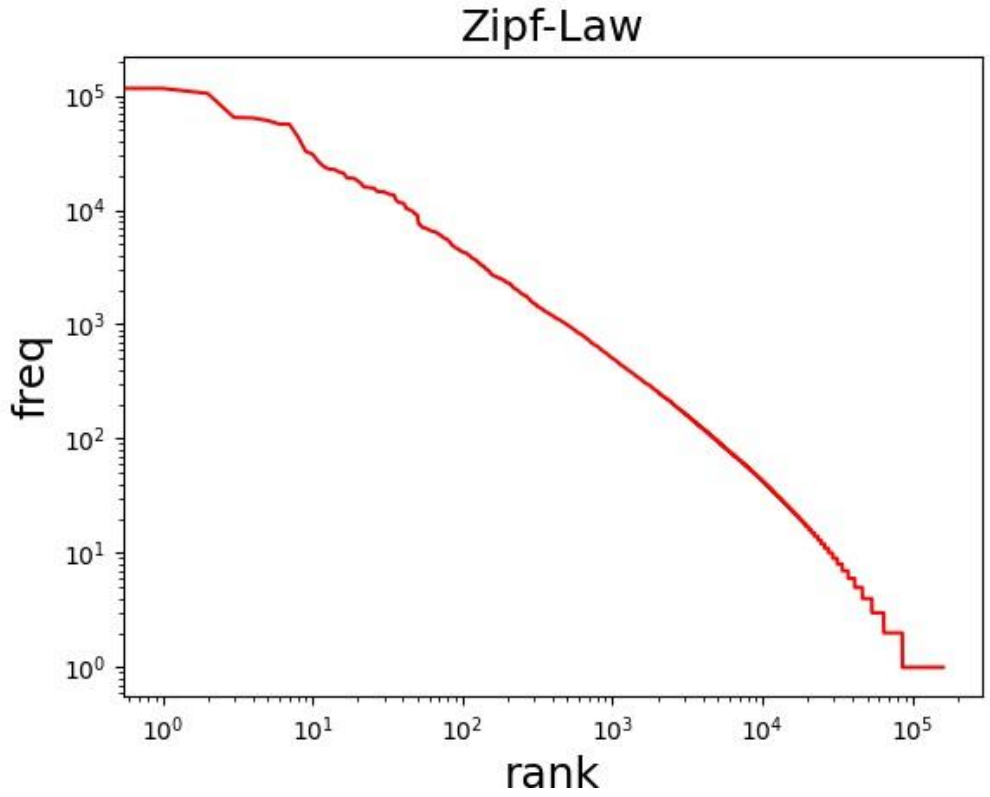

图 1 词频-排名图表

Part 2

《An Estimate of an Upper Bound for Entropy of English》提出了印刷英语字符熵上限的估计，为 1.75 位，通过构建词三元模型并计算该模型与一份平衡的英语文本样本之间的交叉熵获得。此外，Shannon 基于英语字母的概率分布，推导了信息熵的计算方法，并通过英语文本的统计数据进行了实证分析。研究结果为信息论和通信领域提供了重要理论基础，对于理解自然语言的信息内容及其传输具有重要意义。

本文计算中文语料库的字符和词语单位的平均信息熵。首先加载指定的过滤字符文件，然后遍历文件夹中的文本文件，对每个文件进行分词和频率统计，并计算字符和词语单位的信息熵。最后输出结果。

Experiment

中文语料库中 16 个文本信息熵如下：

语料库	词单位平均信息熵	字单位平均信息熵
笑傲江湖	12.309870310994583	9.470611334514883
白马啸西风	9.54521387527253	8.466042584750351
三十三剑客图	12.482304129023827	10.006718182224784
越女剑	9.142983286975124	8.196214694753174
碧血剑	12.720256521039476	9.725055802507779
射雕英雄传	12.355665139522426	9.542337188491967
倚天屠龙记	12.561876590429376	9.623054195771717

侠客行	11.752043558864168	9.27557754876074
鸳鸯刀	9.77589468160301	8.630639195643223
鹿鼎记	12.102079732687063	9.495520556164221
飞狐外传	12.457212354261763	9.596269267429856
神雕侠侣	9.528827212179042	12.286669727570953
书剑恩仇录	9.744555009073421	12.628314653617014
天龙八部	12.560704192419761	9.654094219418674
雪山飞狐	11.640542928759265	9.374737071211268
连城诀	11.641213954513587	9.332673515238321

Conclusions

本文通过信息熵和 Zipf 定律的分析，更深入地理解文本数据的特征。信息熵提供了衡量文本信息量的有效指标，而 Zipf 定律则揭示了词频分布的幂律规律。结合两者可以帮助发现文本中的重要信息和规律性特征，进一步指导自然语言处理和文本分析任务。因此，信息熵和 Zipf 定律的结合应用有助于深入挖掘文本数据的内在结构和含义。

References

[1]Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. Comput. Linguist. 18, 1 (March 1992), 31–40.