Report of Deep Learning for Natural Langauge Processing

Yuxiang Cheng　ZY2303301
cyx1313@126.com

Abstract

First part: to indentify Zipf's Law through Chinese corpus

Second part:Read Entropy Of English and calculate the average information entropy of Chinese (in words and characters respectively)

Introduction

Task one

齐普夫定律最初是根据计量语言学来制定的，一般表述为：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。则最频繁出现的单词的频率大约是第二个最频繁单词的两倍，是第三个最频繁单词的三倍，依此类推。这个定律被作为任何与幂定律概率分布有关的事物的参考。

Task two

"An Estimate of an Upper Bound for the Entropy of English"，该论文由 Shannon 在 1948 年发表的论文，他在这篇论文中提出了信息论的概念，并尝试估计英语的熵上限（熵是信息论中用来衡量信息量的指标，表示一个随机变量的不确定性）。在英语中，每个字母或者单词出现的概率是不同的，因此可以通过统计英语文本中字母或者单词的频率来估计熵的上限。根据 Shannon 的估计，英语文本的熵上限约为 1.0 至 1.5 bits per character（每个字符约 1.0 至 1.5 比特）。文章还从 5.83 亿个训练文本中构建的词三元语言模型来测量其交叉熵。 Shannon's entroy 是最简洁同时用途最多的数学公式之一，其本意是去衡量一件事物信息量的多少，而"信息量的多少"又决定于其"不确定性 。

Methodology& Experimental Studies

Task one:

基于 jieba 库对 Zipf’s Law 的验证

数据采用金庸小说集的语料库，通过 Python 提供的 jieba 库实现中文分词，具体步骤如下：

- **文本读取**

通过 read 函数，并用 gb18030 解码，将文件转化为字符串，并要删除其中的无意义字串。

- **词频统计**

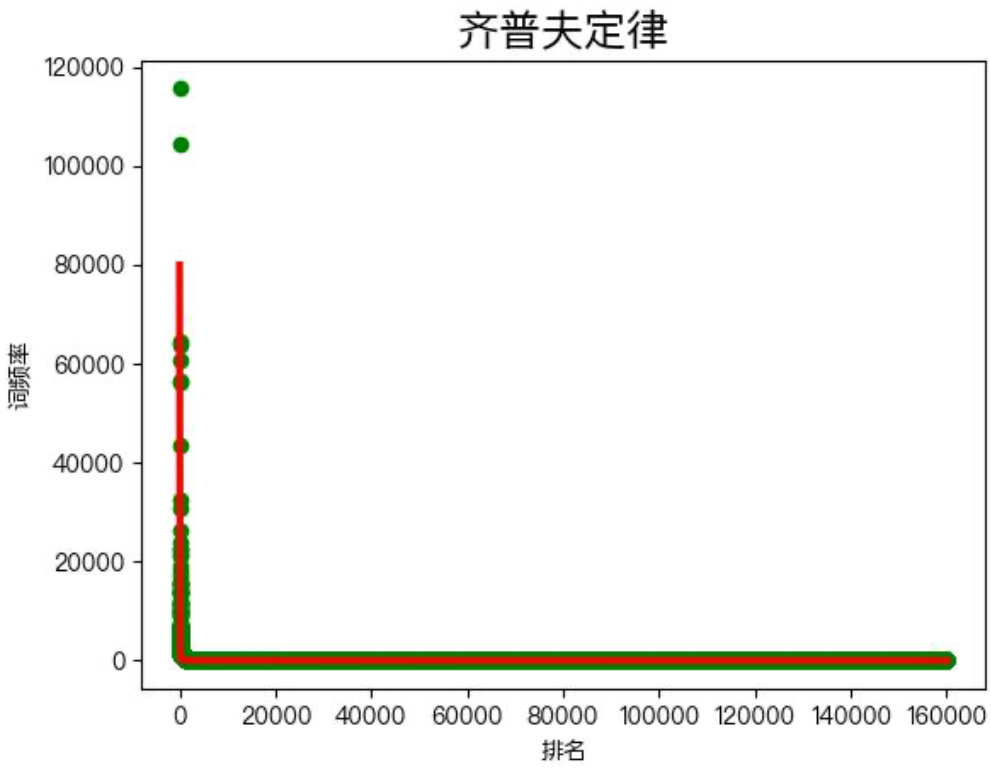
通过 jieba 的 lcnt 函数，进行词频统计。

- **词频排序**

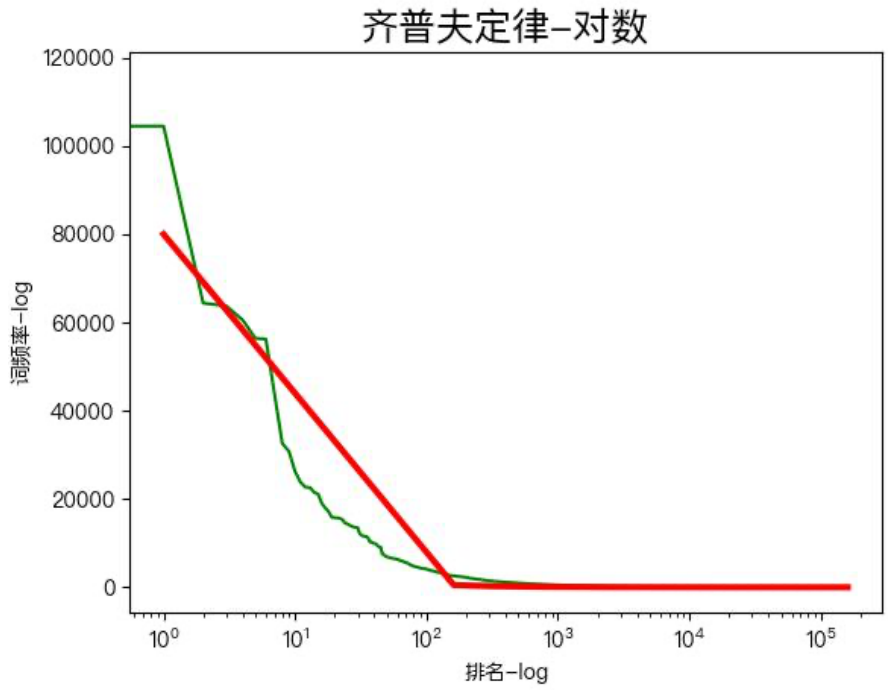
通过 sort 函数按照词频进行倒序排序

- **可视化**

通过 matplotlib 进行可视化词频、排序之间的关系，并绘制参考曲线 y = 80000 / x，同时为了更加直观的观察，采用对数坐标系。



图一：Zipf’s Law 的验证



图二：Zipf's Law 的验证（对数坐标系）

Task two:

本作业采用了最基本的 Unigram 模型对中文信息熵进行了字、词分别计算信息熵，步骤如下：

- **文本读取**

  以 gb18030 编码方式读取小说语料库，并删除所有无意义字串。

- **词频统计**

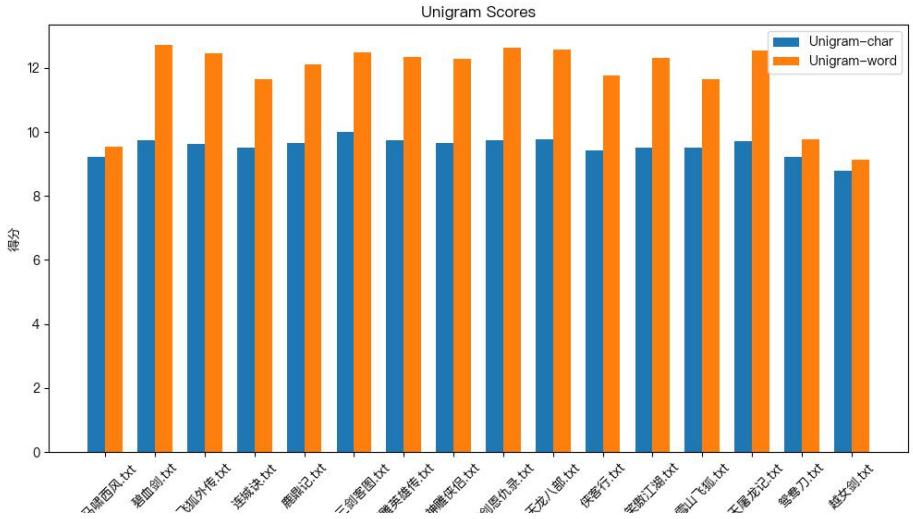
  按照 Unigram 对汉字数组与词组数组进行词频统计。

- **信息熵计算**

  以频率估计概率的方式来计算信息熵

- **绘制图表**

  通过 matplotlib 函数将各个小说文本的字、词信息熵进行柱状图绘制



图三：各个小说文本的字、词信息熵可视化图

Conclusions

在 Task one 中，通过 jieba 库对中文语料库进行了分词与词频统计，利用统计结果验证了 Zipf’s Law；在 Task two 中，基于 Unigram 模型在中文语料库以字、词语两个角度计算了中文信息熵的平均值，词比字的信息熵普遍要高， 分析可以知道字的不确定性要高于词。

References

[1] 用 Python 正则实现词频统计并验证 Zipf-Law_如何判断是否符合 zipf's law python-CSDN 博客

[2] https://blog.csdn.net/qq_43552032/article/details/125614648

[3] Zipf G K. The psychology of language[M]. Encyclopedia of psychology. Philosophical Library, 1946: 332 341.

[4] Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.