学　　期 2023-2024（2）

北京航空航天大學
BEIHANG UNIVERSITY

NLP 第一次大作业

院（系）名称	自动化科学与电气工程学院
专业名称	控制工程
学生姓名	胡正皓
学　　号	ZY2303205

2024 年 4 月

一、研究背景

第一部分：通过中文语料库来验证 Zipf's Law. 第二部分：阅读 Entropy Of English（链接如上），计算中文(分别以词和字为单位) 的平均信息熵。 报告与代码分别用两个链接，报告的形式请看相应的报告链接。

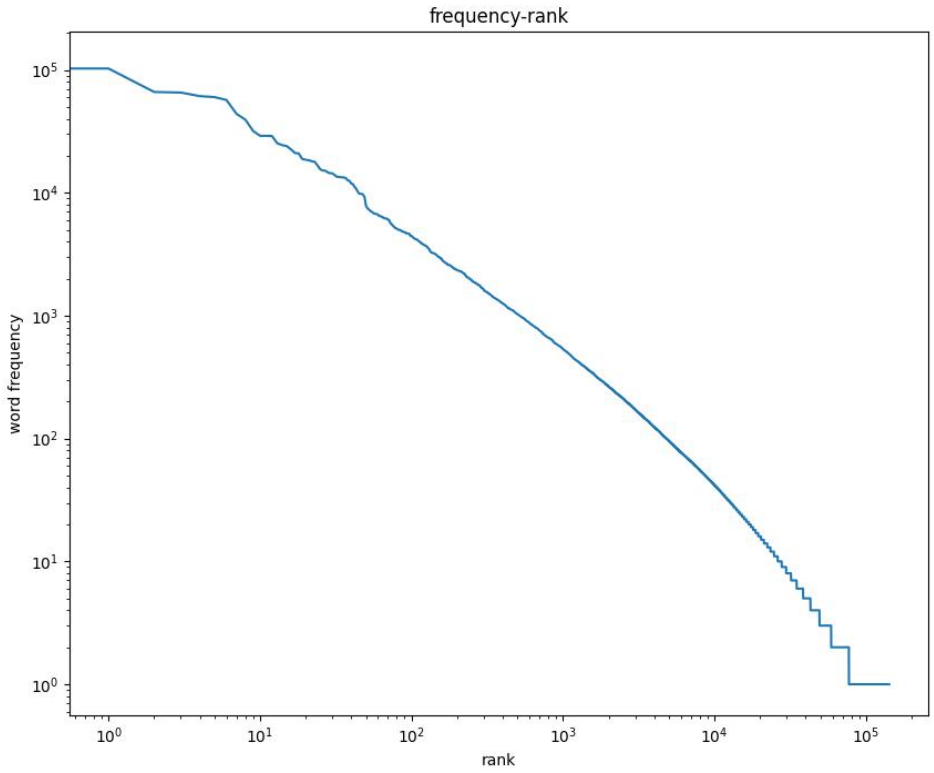
二、Zipf 定律

齐普夫定律是美国学者 G.K.齐普夫于 20 世纪 40 年代提出的词频分布定律。它可以表述为：如果把一篇较长文章中每个词出现的频次统计起来，按照高频词在前、低频词在后的递减顺序排列，并用自然数给这些词编上等级序号，即频次最高的词等级为 1，频次次之的等级为 2，……，频次最小的词等级为 D。若用 f 表示频次，r 表示等级序号，则有 fr=C(C 为常数)。人们称该式为齐普夫定律。

齐普夫定律(Zipf’s law) 的表述为: 当文章作者给出的文献语料库中的词汇足够多时，单词出现频率呈现出一定的分布规律.

研究发现: 不同的作者的用词取向和用词频度是不同的，这种规律被称为“语言指纹”.

所谓用词频度(词频) 是指每一个词在一定长度文件中出现的频率占总词数的比， 如对一个由 K 个词组成的总长度为 L 的语料库中， 词的出现频率由高到低排序为 r 的词频为 Pr. 而依词频从高到低将词排序的序号则是计量的另一个最基本的数量指标.



三、中文平均信息熵计算

对于文本$X=\{\ldots X_{-2}X_{-1}X_0X_1X_2\ldots\}$来说，它的信息熵定义为:

其中信息熵具有三个基本性质：

1、单调性，发生概率越高的事件，其携带的信息量越低；

2、非负性，信息熵可以看作为一种广度量，非负性是一种合理的必然；

3、累加性，即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件不确定性的量度的和，这也是广度量的一种体现。

N_Gram 语言模型

根据大数定理，当统计量足够大的时候，词、二元词组、三元词组出现的概率大致等于其出现的频率。

则有一元模型的信息熵计算公式为：

$$H(x)=-\sum_{x\in X}P(x)logP(x)$$

其中 P(x)可近似于每个词在语料库中的出现频率

二元模型的信息熵计算公式为：

$$H(X|Y) = -\sum_{x\in X, y\in Y} P(x,y) log P(x|y)$$

其中联合概率 P(x,y)可近似等于每个二元词组在语料库中出现的频率，条件概率 P(x|y)可近似等于每个二元词组在语料库中出现的频率与以该二元词组的第一个词为词首的二元词组的频数的比值。

三元模型的信息熵计算公式为：

$$H(X|Y,Z) = -\sum_{x\in X, y\in Y, z\in Z} P(x,y,z) log P(x|y,z)$$

其中联合概率 P(x,y,z)可近似等于每个三元词组在语料库中出现的频率，条件概率 P(x|y,z)可近似等于每个三元词组在语料库中出现的频率与以该三元词组的前两个词为词首的三元词组的频数的比值。

语料库	一元模型	二元模型	三元模型
射雕英雄传	13.088	3.782	0.859
白马啸西风	13.088	2.458	0.335
碧血剑	12.917	3.658	0.434
飞狐外传	12.659	3.609	0.581
连城诀	12.241	3.243	0.618
鹿鼎记	12.677	3.634	0.396
三十三剑客图	12.570	3.165	0.280
神雕侠侣	12.547	3.829	0.794
书剑恩仇录	12.767	3.525	0.373
天龙八部	13.068	3.876	0.675
侠客行	12.325	3.395	0.381
笑傲江湖	12.561	3.588	0.496
雪山飞狐	12.080	3.121	0.307
倚天屠龙记	12.937	3.776	0.481
鸳鸯刀	11.157	2.484	0.091

越女剑	10.517	2.020	0.152