Report of Deep Learning for Natural Langauge Processing_1

ZY2303109 金子棋

Abstract

通过中文语料库验证 Zipf's Law，并依照文献 Entropy of English 计算中文的平均信息熵

Introduction

第一部分 Zipf's Law 指的是当对一篇较长文章统计词频，将频次最高的词等级记为 1，第二常见的词等级记为 2，以此类推，其等级与对应的频次相乘接近于一常数，即两者成反比关系，文章第一部分将对该定律进行验证。第二部分信息熵被用于描述信息源各可能事件发生的不确定性，具体意义是把信息中排除了冗余后的平均信息量，文章第二部分将基于中文语料库计算相应的信息熵

Methodology

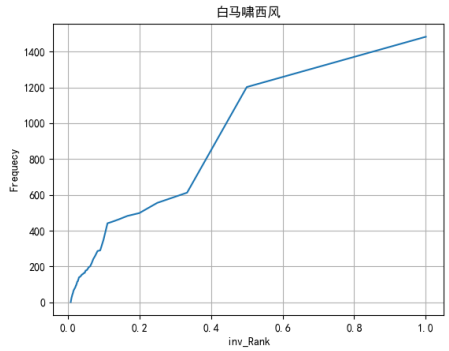
(1) 文本预处理

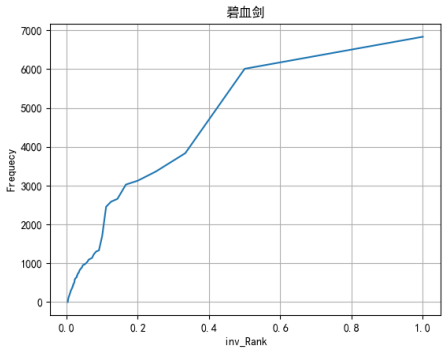
本次使用的中文语料库为金庸小说集，有效正文文本文件共 16 个，此外还有包含所有文件的 zip 包、文本文件来源网站 url 链接以及包含所有章节标题的文本文件 inf.txt，直接在文件夹中删去。

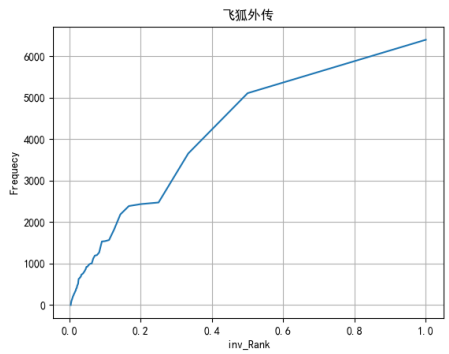
对于正文文本，一需要删除开头结尾网站附上的广告信息，二需要去除文本中非中文字符、空格换行等无效信息，三需要去除中文文本中的标点符号，最后仅剩中文文字作为后续处理的信息源。

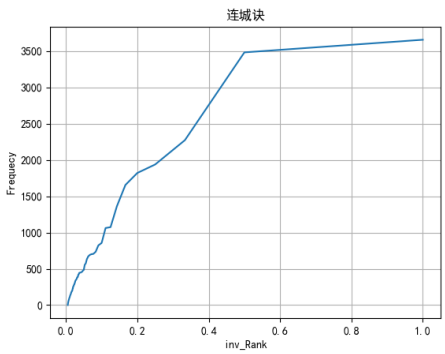
(2) Zipf's Law 验证

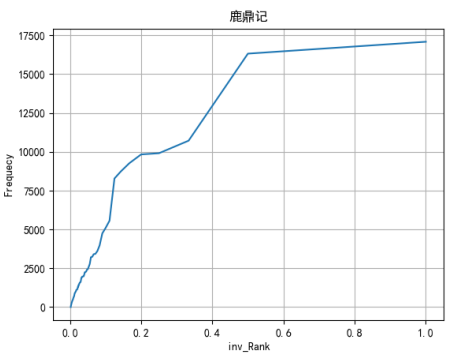
通过 jieba 分词库对文章采用精准模式进行分词，并利用字典得到词频表，排序后得到各词对应的等级。为验证两者成反比的关系，使用 1/等级作为横轴，频次作为纵轴画图，结果如下：

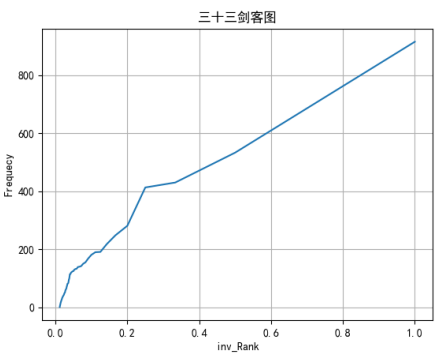

白马啸西风
Frequecy
inv_Rank

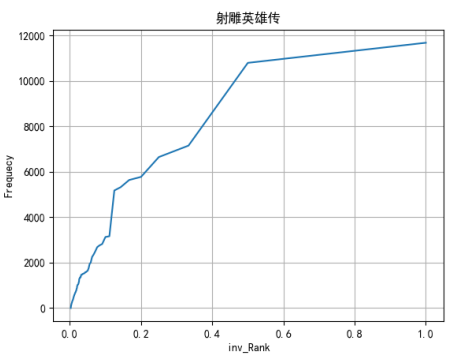

碧血剑
Frequecy
inv_Rank

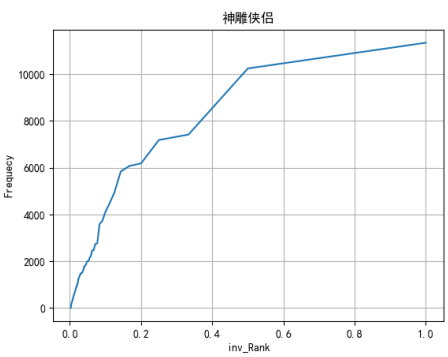

飞狐外传
Frequecy
inv_Rank

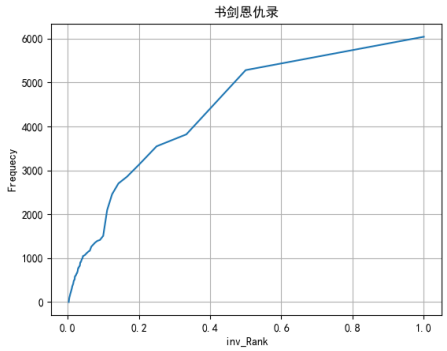

连城诀
Frequecy
inv_Rank

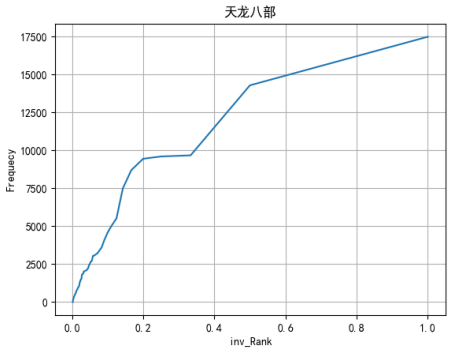

鹿鼎记
Frequecy
inv_Rank

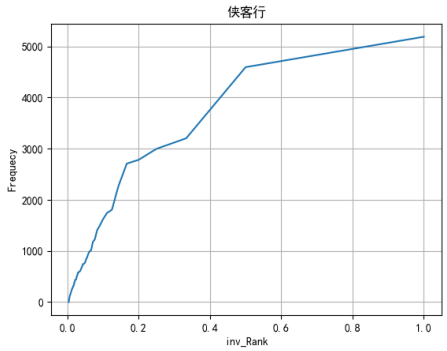

三十三剑客图
Frequecy
inv_Rank

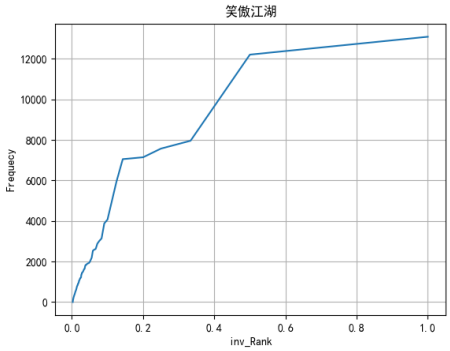

射雕英雄传
Frequecy
inv_Rank

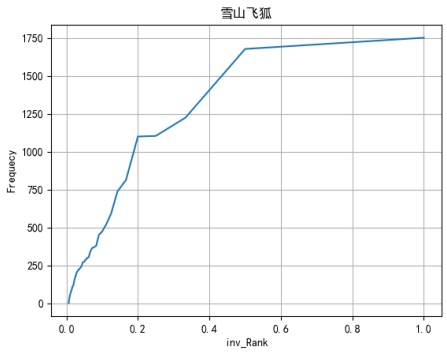

神雕侠侣
Frequecy
inv_Rank

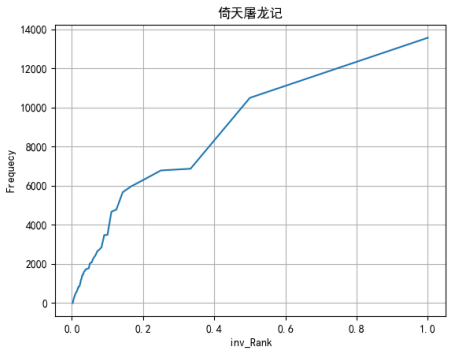


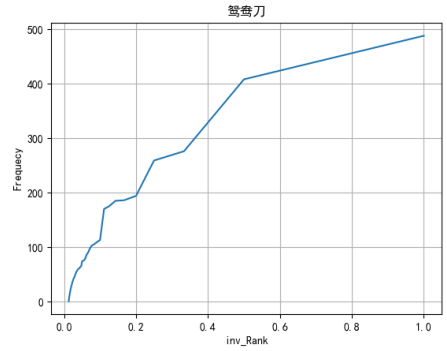


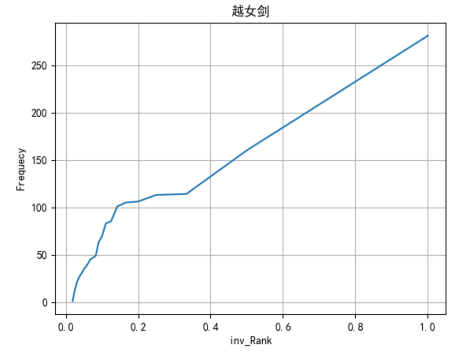












(3) 中文信息熵计算

首先，一个信源发送出什么符号是不确定的，衡量它可以根据其出现的概率来度量。概率大，出现机会多，不确定性小；反之不确定性就大。因此可以

用一个概率 P 的减函数作为不确定的衡量函数，常见的是$f(P) = -\log p$，由此信源的平均不确定性可以记为$H(P) = -\sum_{i=1}^{n} p_i \log p_i$。将文章中出现的词语$\{w_1, w_2, \cdots, w_k, \cdots\}$视为随机过程，则$p_i$表示词语$w_i$在文章中出现的概率，其服从于概率分布$P$，则文章的概率可表示为

$$p(w_1 w_2 \cdots w_n) = p(w_1)p(w_2|w_1)p(w_3|w_1 w_2)\cdots p(w_n|w_1 w_2 \cdots w_{n-1})$$

显然，条件概率$p(w_k|w_1 w_2 \cdots w_{k-1})$很难计算，所以通常使用马可洛夫模型对其进行近似，n 阶马可洛夫模型认为$p(w_i)$与$p(w_{i-1} w_{i-2} \cdots w_{i-n})$有关，即

$$p(w_1 w_2 \cdots w_k)$$
$$= p(w_1)p(w_2)\cdots p(w_n)p(w_{n+1}|w_1 w_2 \cdots w_n)\cdots p(w_k|w_{k-1} w_{k-2} \cdots w_{k-n})$$

由于不可能知道真正的概率分布P，所以用有限样本中各个词语的频次近似表示，记为近似概率模型M，由文献[1]可知由其计算得的信息熵满足

$$H(P) \le H(M) \le H(P, M)$$

在本次实验中分别使用一阶/二阶马克洛夫模型计算，对应的信息熵计算公式为

$$H(M_1) = -\sum_{i=1}^{n} p(w_i)\log p(w_i)$$

$$H(M_2) = -\sum_{i=1}^{n-1} p(w_i w_{i+1})\log p(w_{i+1}|w_i)$$

计算结果如下：

指标 / 文章名	一阶模型		二阶模型	
	词(比特/词)	字(比特/字)	词(比特/词)	字(比特/字)
三十三剑客图	11.6542	9.6335	2.9568	4.8485
书剑恩仇录	11.6657	9.4219	5.0621	5.7970
侠客行	11.1552	9.1147	4.9800	5.6054
倚天屠龙记	11.7269	9.3522	5.5413	6.0428
天龙八部	11.7075	9.3662	5.6958	6.1520
射雕英雄传	11.7820	9.3987	5.5002	6.0728
白马啸西风	10.1997	8.8725	4.0218	4.6036
碧血剑	11.7016	9.4094	5.0258	5.8799

神雕侠侣	11.6890	9.3362	5.5574	6.0738
笑傲江湖	11.3504	9.1773	5.6465	5.9483
越女剑	10.0455	8.7941	2.5481	3.6534
连城诀	11.0147	9.1337	4.7095	5.4206
雪山飞狐	11.0772	9.1633	4.1325	5.1804
飞狐外传	11.4903	9.2681	5.0206	5.7721
鸳鸯刀	10.4365	8.9982	3.1161	4.2304
鹿鼎记	11.4151	9.2446	5.7816	6.0046

Conclusion

第一部分总的来说，所有文章词语频次与词语等级的倒数成正相关关系，以三十三剑客图与越女剑两篇文章的线性关系较好，其他文章则有所偏移，可见 Zipf's Law 本身还是经验法则而非严格的反比关系。

第二部分整体来看，采用二阶马可洛夫模型计算的信息熵无论是按字还是按词计算都相对较低，可见采用信息之间存在相关性的先验假设会使得信息冗余度降低。同时特别注意到三十三剑图/越女剑两篇文章在按词计算的二阶模型下信息熵特别低，这两者在 Zipf's Law 的验证中线性度也是最好的，可见两者存在一定的对应关系。

Reference

[1] Brown P F, Della Pietra S A, Della Pietra V J, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.