Report I of Deep Learning for Natural Langauge Processing

牛华坤
ZY2303315

Abstract

这份报告主要包括了两部分内容。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，计算中文(分别以词和字为单位)的平均信息熵。

Introduction

I1: Zipf's Law

齐普夫定律（Zipf's Law）是由美国学者 G.K.齐普夫在 20 世纪 40 年代提出的一种词频分布定律。该定律可以表述为：

在一篇较长的文章中，如果把每个词出现的频次统计起来，并按照高频词在前、低频词在后的递减顺序排列，然后用自然数给这些词编上等级序号。例如，频次最高的词等级为 1，频次次之的等级为 2，依此类推。如果用 freqs 表示频次，rank 表示等级序号，那么二者相乘近似等于一个常数 C。这个关系可以表示为 freqs*rank=C，其中 C 是常数。

齐普夫定律最初是在对英语文献中单词出现频次的统计研究中提出的，该定律还提出了对于词频分布原因的假说，包括“省力法则”假说和“成功产生成功”假说。这些假说探讨了导致词频分布呈现特定形状的可能原因。

I2: Information entropy

信息熵通常用于描述信息源中事件的随机性或不确定性。一个系统的不确定性越高，其信息熵也越高；反之，系统的不确定性越低，其信息熵也越低。

信息熵的公式是：

$$H(X) = E(-\log p(x_i)) = -\sum_{x_i \in X} p(x_i) \log p(x_i)$$

这个公式是香农在 20 世纪 40 年代提出的，用于度量信息源的平均信息量。在这个公式中，$p(x_i)$ 表示随机事件 X 为 x_i 的概率，$\log$ 是以 2 为底的对数。因此，信息熵本质上是最优信息压缩的界限，即使用最少的信息单位（如比特）来编码消息的期望值。

Methodology

M1: N-gram model

N-gram 模型是一种语言模型（Language Model，LM），语言模型是一个基于概率的判别模型，它的输入是一句话（单词的顺序序列），输出是这句话的概率，即这些单词的联合概率（joint probability）。

假设有一个由 *n* 个词组成的句子 $S=(w_1,w_2,\cdots,w_n)$，为了衡量它的概率，不妨假设每一个单词 w_i 都要依赖于从第一个单词 w_1 到它之前一个单词 w_{i-1} 的影响：

$$p(S)=p(w_1w_2\cdots w_n)=p(w_1)p(w_2\mid w_1)\cdots p(w_n\mid w_{n-1}\cdots w_2w_1)$$

然而随着单词数的增加，参数空间过大导致计算变慢，如 $p(w_n\mid w_{n-1}\cdots w_2w_1)$ 的参数有 *n* 个，故可引入马尔科夫假设（Markov Assumption）：一个词的出现仅与它之前的 *N* 个词有关：

$$p(w_1w_2\cdots w_n)=\prod p(w_i\mid w_{i-1}\cdots w_2w_1)\approx\prod p(w_i\mid w_{i-1}\cdots w_{i-N+1})$$

当 *N*=1 时，每个词的出现与其他词无关，称之为 unigram 模型（一元模型）：

$$p(w_1w_2\cdots w_n)=p(w_1)p(w_2)\cdots p(w_n)$$

其信息熵计算公式如下：

$$H(X)=-\sum_{w_i\in S}p(w_i)\log p(w_i)$$

当 *N*=2 时，一个词的出现仅依赖于它前面出现的一个词，称之为 bigram 模型（二元模型）：

$$p(w_1w_2\cdots w_n)=p(w_1)p(w_2\mid w_1)\cdots p(w_n\mid w_{n-1})$$

其信息熵计算公式如下：

$$H(X)=-\sum_{w_i\in S}p(w_{i-1}w_i)\log p(w_i\mid w_{i-1})$$

当 *N*=3 时，一个词的出现仅依赖于它前面出现的两个词，称之为 trigram 模型（三元模型）：

$$p(w_1w_2\cdots w_n)=p(w_1)p(w_2\mid w_1)p(w_3\mid w_1w_2)\cdots p(w_n\mid w_{n-2}w_{n-1})$$

其信息熵计算公式如下：

$$H(X)=-\sum_{w_i\in S}p(w_{i-2}w_{i-1}w_i)\log p(w_i\mid w_{i-2}w_{i-1})$$

对条件概率，采用极大似然估计法计算（即查相邻词出现频数），如：

$$p(w_n\mid w_{n-2}w_{n-1})=\frac{C(w_{n-2}w_{n-1}w_n)}{C(w_{n-2}w_{n-1})}$$

Experimental Studies

E1：验证 Zipf's Law

根据金庸小说语料库验证 Zipf's Law 的基本程序流程图如下图所示：

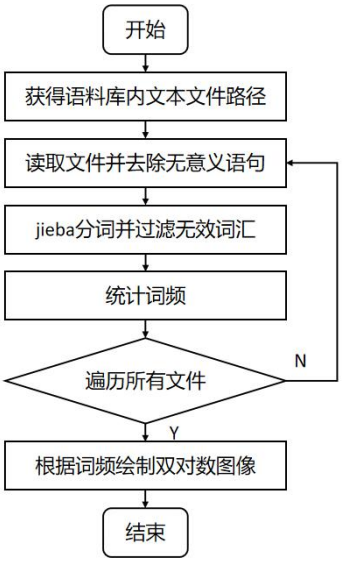

Figure 1：验证 Zipf's Law 的程序流程图

最终得到的排名-词频双对数图像如下图所示：

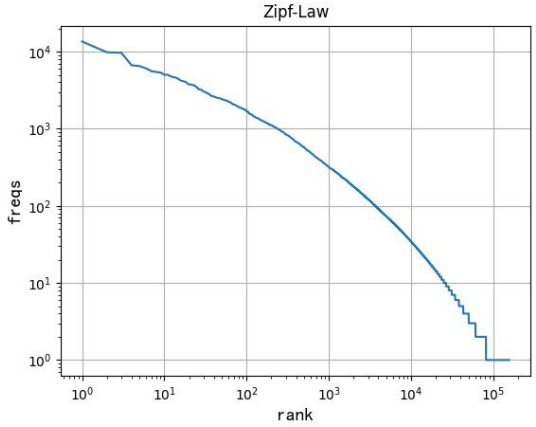

Figure 2：验证 Zipf's Law 的 rank-freqs 图

E2：计算平均信息熵

根据金庸小说语料库计算中文信息熵的基本程序流程图如下图所示：

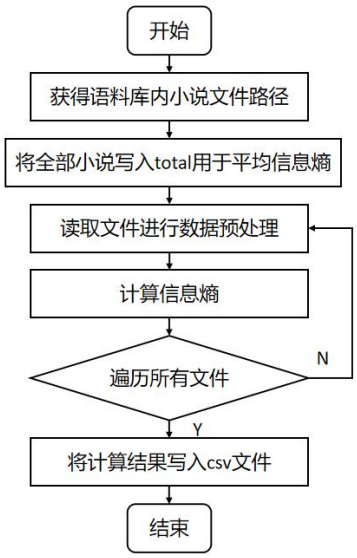

Figure 3：计算中文信息熵的程序流程图

最终得到的信息熵计算结果如下表所示，char 代表按字计算，word 代表按词计算，最后一行 total 即为所有小说汇总后得到的平均信息熵：

Table 1: this is the table 1

FileName	Char-Unigram	Word-Unigram	Char-Bigram	Word-Bigram	Char-Trigram	Word-Trigram
白马啸西风	9.232165974	9.545213875	4.085092485	3.030396976	1.209502728	1.094960372
碧血剑	9.755965809	12.72025652	5.674985012	3.943674211	1.795147874	0.517609113
飞狐外传	9.630063598	12.45721235	5.569077169	4.010437645	1.864972036	0.551207413
连城诀	9.515225215	11.64121395	5.090139273	3.533264397	1.638949452	0.639037315
鹿鼎记	9.659006321	12.10207973	6.019779789	4.856069814	2.40937911	1.069256385
三十三剑客图	10.01105911	12.48230413	4.281669157	1.831707273	0.649843759	0.119357703
射雕英雄传	9.752142991	12.35566514	5.965059818	4.530091118	2.195892502	0.873078375
神雕侠侣	9.663303823	12.28669073	6.002331579	4.880712343	2.28255384	0.839586685
书剑恩仇录	9.758488469	12.62831465	5.597777178	4.113069965	1.861289487	0.552920855
天龙八部	9.782788191	12.56070419	6.114711899	4.72328184	2.351452331	0.884042498
侠客行	9.436598028	11.75204356	5.379339446	3.882354559	1.819103092	0.759270645
笑傲江湖	9.516218771	12.30987031	5.856436219	4.760809211	2.36119401	0.897065964
雪山飞狐	9.504523243	11.64054293	4.799800891	3.053852164	1.302099487	0.495780564
倚天屠龙记	9.706869731	12.56187659	5.982905109	4.615630104	2.275938696	0.804311339
鸳鸯刀	9.220968953	9.775894682	3.651094057	2.429917685	0.893597867	0.842910037
越女剑	8.805137528	9.142983287	3.095291917	2.055526173	0.836950911	0.823473494
total	9.952138076	13.12662424	7.021648272	6.373260272	3.493406322	1.391015863

Conclusions

C1: 验证 Zipf's Law

由小说的词排名和词频的双对数曲线可知，二者的对数曲线近似为一次负相关关系，符合齐普夫定律中描述的反比关系。

C2: 计算平均信息熵

根据表格数据可知，当 *N* 值变大时，基于字和词的统计结果均出现信息熵变小的趋势，这是因为 *N* 值越大，涵盖的信息量越多，信息的不确定性变小，信息熵越小。

References

[1] Brown P F , Pietra V J D , Mercer R L ,et al.An estimate of an upper bound for the entropy of English[J].Computational Linguistics, 1992.DOI:10.5555/146680.146685.