深度学习与自然语言处理作业 1 报告

段沛东
ZY2303106
19231209@buaa.edu.cn

摘要

本文为深度学习与自然语言处理课程第一次作业的报告，主要由两部分内容组成。第一部分内容为使用给定的中文语料库（16 本金庸小说）验证了齐普夫定律；第二部分内容根据参考文献获取了计算语言平均信息熵的公式，并且以给定的中文语料库，分别使用以词和字为单位的一元、二元、三元模型计算了中文的平均信息熵。

问题描述

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。本作业的第一部分内容为使用给定的中文语料库来验证其是否符合齐普夫定律。

信息熵的概念最早由香农（1916-2001）于 1948 年借鉴热力学中的“热熵”的概念提出，旨在表示信息的不确定性。熵值越大，则信息的不确定程度越大。对于语言来说，不同的语言平均每个字符所含有的信息量不同，这可以通过语言的平均信息熵来反映。本作业的第二部分内容为使用给定的中文语料库来计算中文的平均信息熵，要求分别计算以词为单位和以字为单位两种情况下的信息熵。

齐普夫定律验证实验的结果与分析

本实验编写 python 程序实现了使用给定的中文语料库验证齐普夫定律，首先获取一个储存了语料库中各个.txt 文件绝对路径的列表，并且依次读取其中的内容，将读取到的内容拼接成一个总字符串，之后使用 re 库处理所得字符串，将汉字以外的符号全部剔除，使用结巴分词库对处理之后的字符串进行精确分割，得到分割后的中文词语组成的列表，最后还需要将列表中的词语与中文停词列表进行对比，删去除于停词列表中的词语，得到最终用于验证定律的词语列表。使用 collections 库的 Counter 方法统计列表中各个词语出现的频数，可视化结果如图 1 所示。可以看出在词语名次和词语频数组成的坐标系中绘制出的图像基本呈斜率为-1 的直线，由于采用的是对数坐标轴，可得词语名次和词语频数大致呈反比例函数关系，满足齐普夫定律。

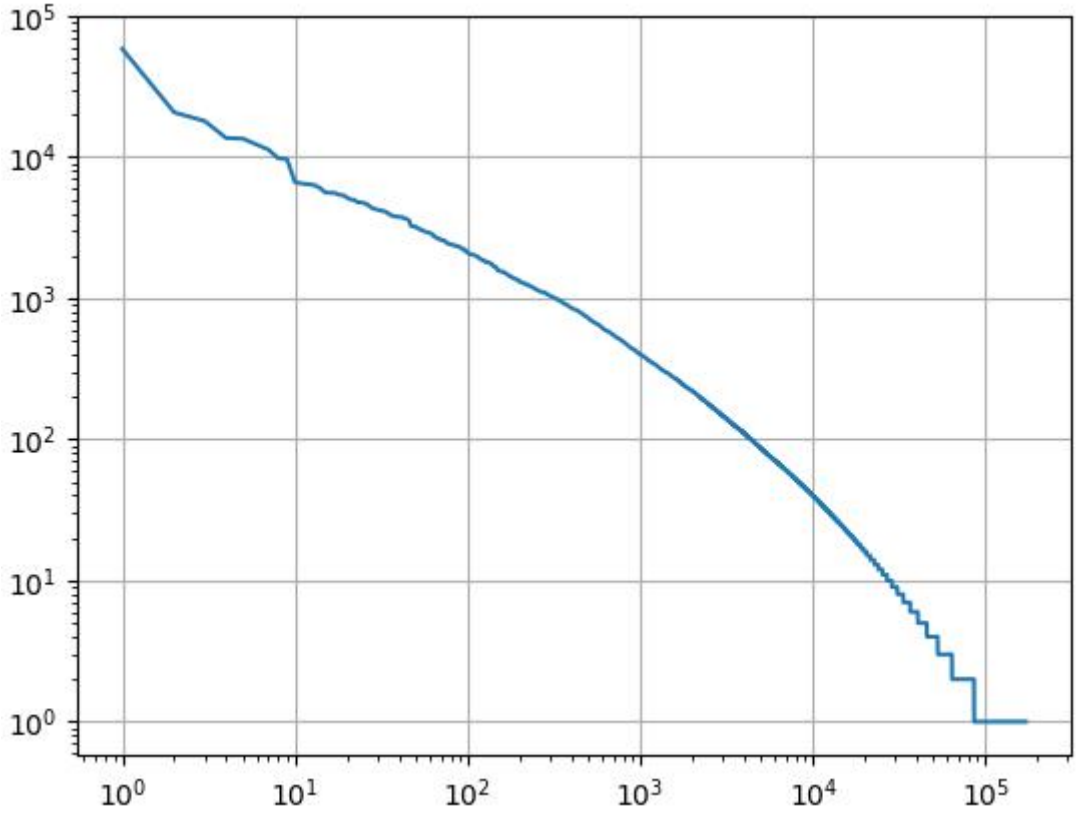

图 1 名次和频数关系图（以词为单位分割）

除了以词为单位分割得到列表，本实验还尝试了直接以字为单位分割得到列表，并且使用其验证齐普夫定律，结果如图 2 所示。

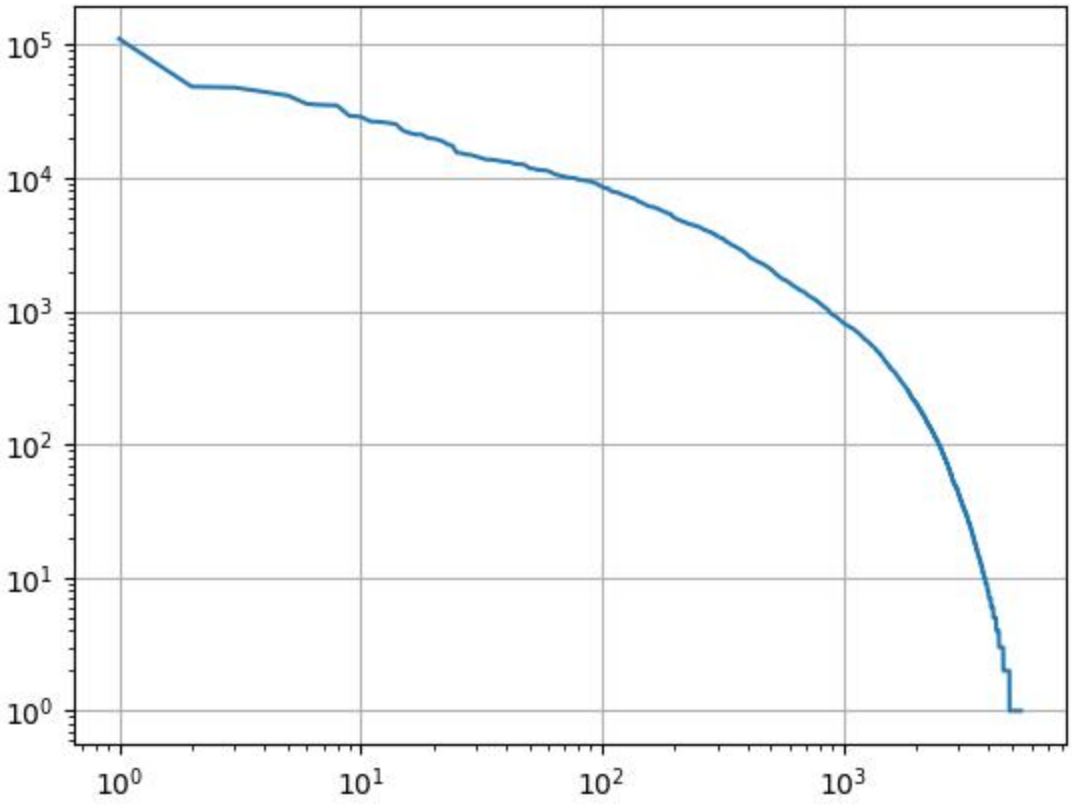

图 2 名次和频数关系图（以字为单位分割）

可以看出绘制出的图像弯曲程度明显增加，即满足齐普夫定律的程度低，这能够反映出中文是以词语为基本单位而不是字，所以在自然语言处理的研究中应该以词语作为研究中文语言的基本单位。

中文平均信息熵计算实验的结果与分析

根据文献[1]，我们可以得到信息熵的计算公式如下：

$$H(X) = -\sum_{x \epsilon X} p(x) log p(x)$$

对于语言而言，X表示的是该门语言中所有词语的集合，$p(x)$则为x出现的概率，显然，$p(x)$的分布是未知的，我们无法求得精确的分布，只能使用语言模型去估计其分布，估计的近似程度越高，所求得的信息熵也就越准确。本实验采用了一元模型、二元模型和三元模型分别作为$p(x)$的估计，并且对每种模型都和齐普夫定律验证实验中所做的一样根据以字和词为单位分割分别进行了讨论。语料库的处理方法和上一实验相同，代码运行的结果如图 3 所示：

```
以字为单位使用一元模型计算中文平均信息熵的结果为：  16.977155972200404 比特/词
以词为单位使用一元模型计算中文平均信息熵的结果为：  20.146342671363964 比特/词
以字为单位使用二元模型计算中文平均信息熵的结果为：  7.034742242410081 比特/词
以词为单位使用二元模型计算中文平均信息熵的结果为：  6.503044542435616 比特/词
以字为单位使用三元模型计算中文平均信息熵的结果为：  3.5021734402171893 比特/词
以词为单位使用三元模型计算中文平均信息熵的结果为：  1.1515360782306 63 比特/词
```

图 3 中文平均信息熵计算结果

可以看出随着 N-gram 模型 N 取值的增加，即考虑前后文关系的长度增大，平均信息熵呈减小趋势，这种趋势在以词单位的分割上表现的更为明显。考虑可能的原因为随着 N 的增加，得到的词语表意更加准确，因此更加清晰地分出了高频词语和低频词语，使得平均信息熵降低。

参考文献

[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.