作业一、中文平均信息熵

崔多
1479518308@qq.com

一 摘要

在自然语言处理领域中，信息熵也被用来衡量语言中的信息量或信息密度。对于汉字来说，每个汉字都有其特定的出现概率，我们可以通过统计汉字出现的频率来计算汉字的信息熵。中文分词的方式有多种，其中基于字的分词和基于词语的分词是比较常见的。基于字的分词将文本按照每个汉字进行切分，而基于词语的分词则将文本切分成一个个常见的词语。

本文将分别使用基于字和基于词语的分词方式，计算同一段中文文本的信息熵，并对比两种方式的结果。

二 研究内容简介

信息熵是是信息论的基础，可以理解为某种特定信息的出现概率，描述了信息员各种可能时间发生的不确定性，既系统的混乱程度。

对于语言而言，信息熵可以表明单个字符所包含的信息量。本文所计算的汉字信息熵，即是通过给定的量化中文文本，计算每个字符出现的概率，经过概率和熵值的计算，进而获得汉字信息熵。

三 实验方法

3.1 熵，信息熵

熵是所有可能结果的信息量的总和，表示的是期望的稳定性，熵值越小，期望越稳定，当熵为 0 时该事件为必然事件，熵越大表示该事件的可能性越难以估量。

信息熵是信息理论中一个重要的概念。它是信息量的度量，表示随机变量中信息的平均度量。在信息论中，信息熵的大小表示了一个信息源的不确定度、信息量以及传输的效率。信息熵的单位是比特，包括基于二进制、十进制等多种不同的计量方式。 一个信息源中的信息熵越高，说明这个信息源的不确定度越大，信息量越大，传输的效率就越低。相反，信息熵越低，说明这个信息源的不确定度越小，信息量越小，传输的效率就越高。

假设$\mathrm{X}=\{\ldots X_{-2}，\ X_{-1}，\ X_0，\ X_1，\ X_{-2}\ldots\}$是有限域上的离散随机事件，设 P（$X_i$）表示事件$X_i$的发生概率。X 的熵值如式 3.1 所示。

$$\mathrm{H(X)}=\mathrm{H(P)}=\lim_{n\to\infty}-E_P\,log\,P(X_0|X_{-1},X_{-2},\ldots) \tag{3.1}$$

对数的底通常选用 2 和 e，也可以选用其他数字，底数为 2 时，单位为 bit，此时熵值公式可表示为式 3.2。

$$\mathrm{H(P)}=\lim_{n\to\infty}-E_P\,log\,P(X_0|X_{-1},X_{-2},\ldots)=\lim_{n\to\infty}-\frac{1}{n}E_P\,log\,P(X_1X_2\ldots X_n) \tag{3.2}$$

当事件的随机过程满足平稳性和遍历性条件时，以下关系成立:

$$\mathrm{H(P)} = \lim_{n\to\infty} -\frac{1}{n} log\, P(X_1 X_2 \ldots X_n) \tag{3.3}$$

3.1 中文信息熵

中文信息熵是对一段中文文本的信息量大小的度量，可以用来衡量文本的复杂度和难度。信息熵计算公式为：

$$\mathrm{H(X)} = -\sum P(X_i)\log P(X_i) \tag{3.4}$$

信息熵的计算流程为：

1）统计文本中每个汉字或词组出现的次数，并计算其出现的概率。

2）对于每个汉字的出现概率，根据公式 3.4，计算$-P(X_i)\log P(X_i)$的值，然后将所有值相加，得到信息熵 H（X）的值。

四 实验结果

4.1 以字为单位的信息熵

1）数据预处理：

通过遍历文件夹读取文本文件，去除英文，根据 cn_punctuation.txt 去除特殊符号等非中文的干扰信息，获得全部汉字共计 7261689 个，并以列表形式储存。

2）计算中文信息熵：

经计算，资料库中共有 5765 个不同的汉字，统计文本中每个汉字的出现次数，获得的出现次数最多的汉字如表 4.1 所示。

表 4. 1 汉字及出现频次

汉字	出现次数	汉字	出现次数
一	139398	了	111928
不	134151	道	111058
的	121672	人	84306
是	112709	他	73576

根据公式 3.4，本文计算汉字的出现概率以及计算信息熵，获得的 17 篇文献的信息熵为 9.5438。

4.2 以词为单位的信息熵

1）分词与数据预处理：

为了防止不同语句之间的错误拼接，通过先分词，再筛选汉字的方式进行预处理。

汉字的分词根据 cn_stopwords.txt 作为停词词库进行分词，共获得 6188792 个词组。

遍历全部分词去除英文，标点及特殊符号等非中文的干扰信息，获得全部根据停词词库所获得的词表，共计词语 4687640 个。

2）计算中文信息熵：

经计算，资料库中共有 163304 个不同的词语，统计文本中每个词语的出现次数，获得的出现次数最多的汉字如表 4.1 所示。

表 4. 2 词语及出现频次

汉字	出现次数	汉字	出现次数
的	114643	道	60528
了	103506	你	56402
他	64314	我	56233
是	60995	在	41666

根据公式 3.4，本文计算汉字的出现概率以及计算信息熵，获得的 17 篇文献的信息熵为 13.0444。

4.2 不同停词方式的中文信息熵

本文对比了一汉字，停词库，jieba 库以及 snownlp 等常用停词库，进行不同停词方式下的汉字信息熵计算，计算结果如表 4.3 所示。

表 4. 3 不同停词方式下的汉字信息熵

停词方式	中文信息熵	文本分词数	不同词语数
汉字	9.5438	7261689	5765
停词词库	13.0444	4687640	163304
Jieba 库	13.0244	4698386	163659
thulac 库	11.9030	4982105	181800

五 结论

中文的信息熵通常偏大，单个汉字在不同词组中的表达意义多有不同，每个汉字通常有多种不同的含义，而不同的词组可以通过不同的汉字组合产生相似或完全不同的含义。因此，如果仅仅以汉字为单位进行信息熵的计算，可能会忽略这些复杂的语言现象，导致信息熵计算的结果偏小。

因此，在计算中文的信息熵时，需要考虑分词模型。分词是影响中文信息熵计算的因素之一。中文分词具有一定的主观性和灵活性，因此不同的分词方式可能会对信息熵的计算产生不同的影响。一般来说，采用合理的分词方式可以更准确地反映中文语言的复杂性和丰富性，从而更准确地计算中文的信息熵

目前，已有较为完善的停词词库，如 jieba，thulac，snownlp 等等，对于中文信息上的研究已经较为简便，此外，还可以通过多元信息熵，以滑动窗口的模型方式，对中文的信息熵有更加深刻的研究。