中文信息熵分析

ZY2207304-范想明

一、背景

信息熵是英国数学家克劳德·香农于 1948 年提出的一个表征符号系统中单位符号平均信息量的指标，并给出了一个十分简洁的公式：

$$H(X)=-\sum_{i=1}^{n}P(x_i)\log_2 P(x_i) \tag{1}$$

式中：$P(x_i)$——某符号系统中符号 x_i 出现的概率

假设一个句子中由特定的字或词（x_i，$i\in 1,2\ldots,n$）组成，则其概率分布为：

$$P(x_1,x_2,...,x_n)=P(x_1)P(x_2\mid x_1)...P(x_n\mid x_1,x_2,...,x_{n-1}) \tag{2}$$

二、研究方法

以分析单个符号（字或词）为例，一般其分布构建简化为 n 元语言模型，即将每个单元的概率只与前（n-1）个单元有关，其中一元模型可表示为：

$$P(x_1,x_2,...,x_n)=P(x_1)P(x_2)...P(x_n) \tag{3}$$

以此类推，二元、三元模型可分别表示为：

$$P(x_1,x_2,...,x_n)=P(x_1)P(x_2\mid x_1)...P(x_n\mid x_{n-1}) \tag{4}$$

$$P(x_1,x_2,...,x_n)=P(x_1,x_2)P(x_2\mid x_2,x_1)...P(x_n\mid x_{n-2},x_{n-1}) \tag{5}$$

因此，根据上述简化模型，一元、二元和三元模型下，其信息熵可分别表示为：

$$H(X)=-\sum_{i=1}^{n}P(x_i)\log_2 P(x_i) \tag{6}$$

$$H(X\mid Y)=-\sum_{i=1}^{n}P(x_i\mid y_i)\log_2 P(x_i\mid y_i) \tag{7}$$

$$H(X\mid Y,Z)=-\sum_{i=1}^{n}P(x_i,y_i,z_i)\log_2 P(x_i\mid y_i,z_i) \tag{8}$$

其中联合概率和条件概率的计算根据《数理统计》中的公式进行计算即可。

三、实验

本测试中以十六部小说《白马啸西风》、《碧血剑》、《飞狐外传》、《连城诀》、《鹿鼎记》、《三十三剑客图》、《射雕英雄传》、《神雕侠侣》、《书剑恩仇录》、《天龙八部》、《侠客行》、《笑傲江湖》、《雪山飞狐》、《倚天屠龙记》、《鸳鸯刀》、《越女剑》为语料进行处理和分析，处理阶段将所有文本中非中文内容剔除，而在分词阶段，利用 jieba 库对处理后的语料进行分词，如遇到 cn_stopwords 中的分词符则将其去除，防止大量的分词符污染数据。

最终的计算结果如下图所示：

```
对中文基于单个字的一元模型的信息熵为: 9.85621112894599
对中文基于单个词的一元模型的信息熵为: 13.746148516657124
对中文基于单个词的二元模型的信息熵为: 6.408944789193165
对中文基于单个词的三元模型的信息熵为: 1.1107175812279841
```