# 汉字和英文字母的信息熵

范逍宇 2019013273

## 一、信息熵的定义。

之所以选择这个自选实验，是因为我在生活中经常听到有人说中文是世界上信息密度最大的文字之一，联合国决议的不同语言的翻译版之中，中文版总是最薄的。例如，下面是联合国主页（www.un.org）在今天（2022年4月22日）的标语，可以看到确实是中文版本最短。


当然，以上只是道听途说的传闻，没有进行严谨的定量研究，但是，这也引发了我的好奇，中文到底是不是信息密度很大的文字呢？在学完信息熵的理论之后，我终于有机会对这个猜想进行验证。

1948 年，英国数学家克劳德·香农在《A mathematical theory of communication》[1]一文中给出了信息熵的定义，用来表征符号系统中单位符号的平均信息量，具体来说，信息熵的计算公式如下：

$$\Omega = -\sum_{i=1}^{n} p_i \log_2 p_i$$

其中 $p_i$ 是该符号系统中某个符号 i 出现的频率。

有了这个公式，我们就可以计算汉字和英文字母的信息熵了。

## 二、计算汉字和英文字母的信息熵。

### 1.数据集。

要计算汉字或英文字母的信息熵，首先需要找到汉字和英文的语料库，而且这个语料库最好规模较大，能够体现出该语言在各种场景下的使用，这样才能准确地计算每个汉字或字母在中文或英文中出现的频率，从而准确地计算信息熵。

这里，使用目前 NLP 领域常用的中文语料库——nlp_chinese_corpus（https://github.com/brightmart/nlp_chinese_corpus，在github收藏量达到 7k+）中的中英文翻译部分语料作为本次实验的语料库。其中既有对于专业术语的翻译，也有对于日常口语的翻译，覆盖了语言在不同场合下的用法，而且这样保证了中英文语料在语义方面是等价的，排除了语义不同可能造成的影响。

在该语料库中，一共有 520 万对中英文互译的句子，其中一个例子如下：

```
{"english": "In Italy, there is no real public pressure for a new, fairer tax system.",
"chinese": "在意大利，公众不会真的向政府施压，要求实行新的、更公平的税收制度。"}
```

其中每个中英文对中，中文句子平均有36个字，英文句子平均有19个单词。

经统计，在训练集中，一共有158,587,384个汉字，487,287,397个英语字母。

由于语料库太大，因此上交作业的压缩包文件中不包含语料库的数据，如果要运行实验代码，需要从 https://drive.google.com/file/d/1EX8eE5YWBxCaohBO8Fh4e2j3b9C2bTVQ/view 下载语料库数据，并放在 /data 文件夹下。

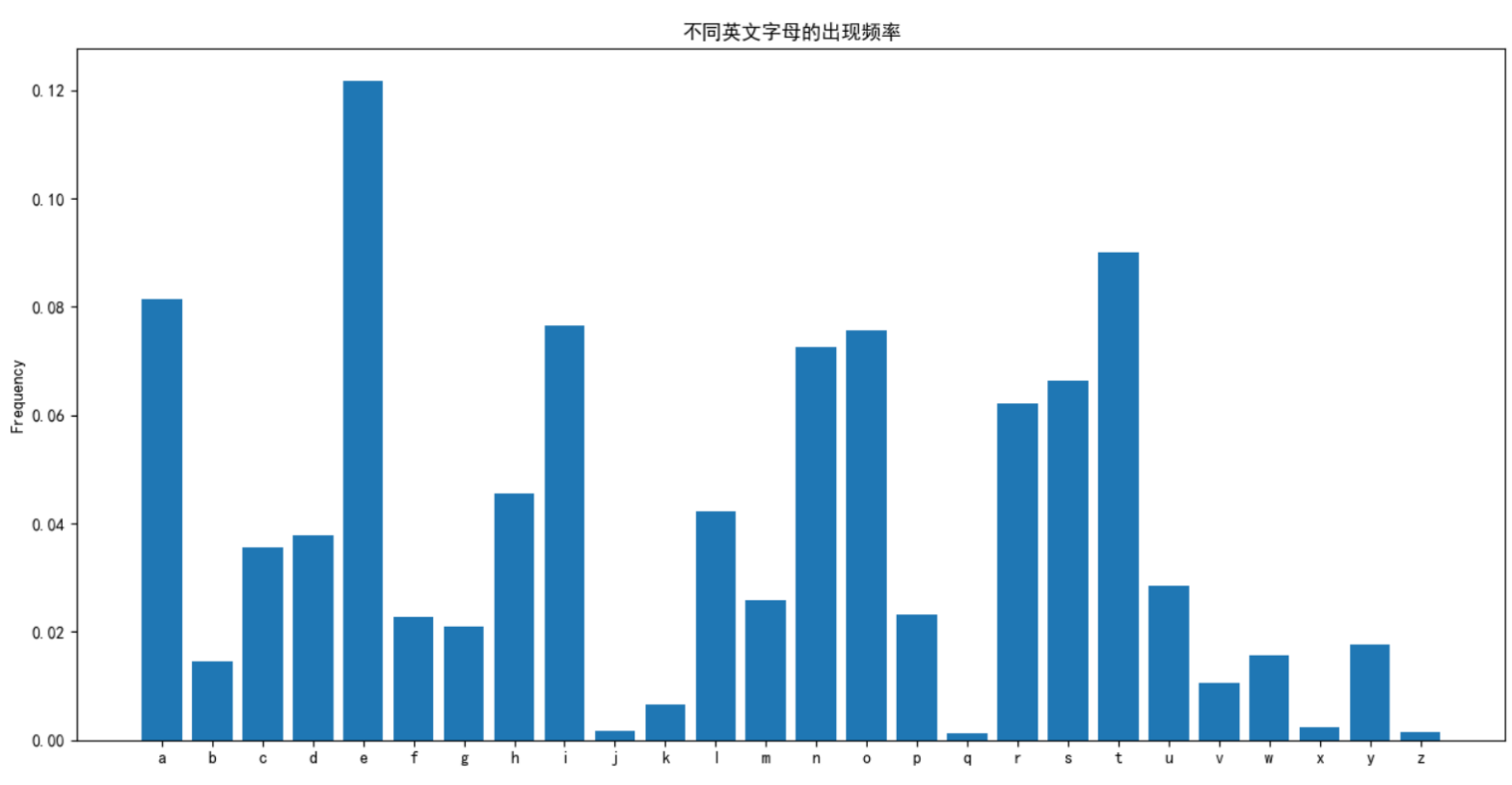
## 2.计算信息熵。

计算过程比较简单，根据信息熵的定义，我们首先需要统计语料库中每个汉字或英语字母出现的频率，然后代入信息熵的计算公式即可。

下面是每个英语字母出现的频率：（不区分大小写）

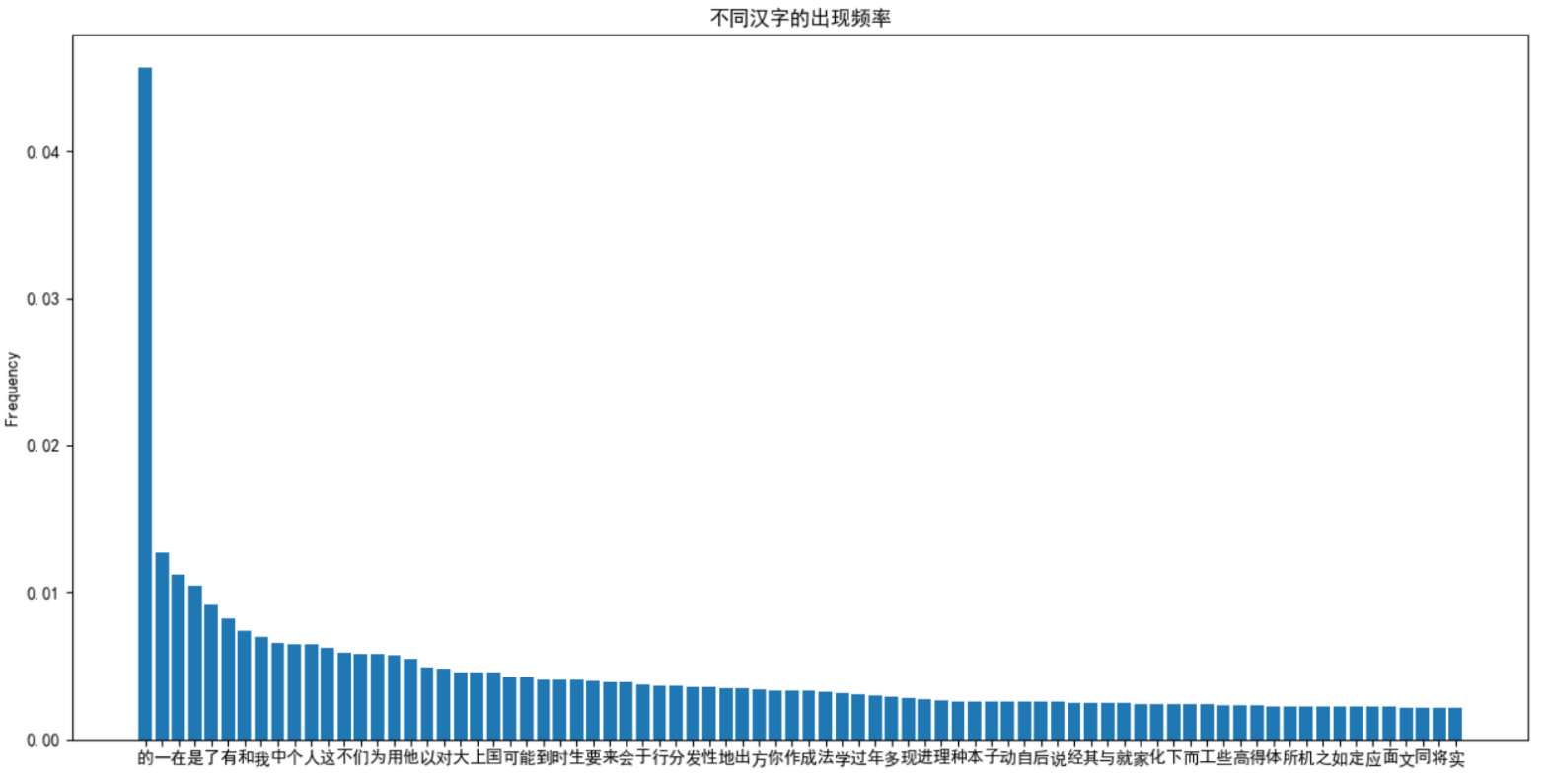
```
a : 0.0815  b : 0.0145  c : 0.0355  d : 0.0378  e : 0.1216  f : 0.0228  g : 0.021
h : 0.0454  i : 0.0766  j : 0.0018  k : 0.0065  l : 0.0421  m : 0.0257  n : 0.0725
o : 0.0757  p : 0.0231  q : 0.0012  r : 0.0622  s : 0.0664  t : 0.0900
u : 0.0284  v : 0.0105  w : 0.0156  x : 0.0024  y : 0.0177  z : 0.0014
```

不同英文字母出现频率的统计图为




可以看到，不同英文字母的出现频率有较大差异，e 的出现频率最高，q 的出现频率最低，它们的出现频率相差了 100 倍，根据信息熵的计算公式，在字符种类相同的情况下，这将会导致信息熵的降低。

下面是不同汉字的出现频率：（为了便于画图，只取了出现频率最大的前 80 个汉字）




可以看到，汉字的出现频率同样由较大差异，“的”出现的频率最大，而有的字在语料库中只出现了 1 次。但除了“的”之外，其他常用汉字的出现频率相差没有英文字母那样悬殊。

信息熵的计算结果如下：（单位为比特）

```
中文汉字的信息熵为：9.6685
英文字母的信息熵为：4.1839
```

可以看到，汉字的信息熵明显大于英文字母的信息熵，这实际上说明，一个汉字蕴含的信息量大于一个英文字母蕴含的信息量。

## 三、相关工作和分析思考。

事实上，很早之前就有科学家对各种语言的信息熵进行了计算，例如香农本人就设计过计算英语字母信息熵的实验。目前一般认为英文字母的信息熵在 3.9 左右，这与本实验的结果非常接近。

1976 年，Wong [2]等人从汉字、部首、音节等角度计算了汉字的信息熵，得出的结论如下：（三行分别表示汉字、部首、音节的信息熵）

| | Entropy (bits) |
|---|---|
| characters | 9.63 |
| radicals | 6.41 |
| Mandarin syllables | 8.68 |

可以看到，汉字的信息熵的计算结果与本实验的计算结果极其接近，另外也可以看出，汉字、部首、音节的信息熵都高于英文字母的信息熵。

2002 年，Behr 等人使用PPM压缩算法来衡量不同语言的信息熵，他们研究了《圣经》的不同语言版本的原始大小和压缩后的大小，实验结果如下：

| Language | Original Size (bytes) | Ratio (Original) | Compressed Size (bytes) | Ratio (Compressed) |
|---|---|---|---|---|
| English | 1936473 | 1 | 390846 | 1 |
| Spanish | 1804756 | 0.932 | 384681 | 0.962 |
| French | 1896459 | 0.979 | 393805 | 1.01 |
| Chinese | 884860 | 0.457 | 337505 | 0.864 |
| Korean | 1259920 | 0.651 | 346478 | 0.886 |
| Arabic | 1875204 | 0.968 | 418443 | 1.071 |
| Japanese | 1519224 | 0.785 | 452337 | 1.157 |
| Russian | 1506920 | 0.778 | 376162 | 0.962 |

Table 2: Results for the Bible using the PPMD compression algorithm. The ratio is the ratio of the size in the language divided by the size in English. We expect the ratio of the compressed sizes to be close to 1.

可以看出，中文版本的《圣经》在上述语言中的原始大小最小，而各种语言版本的圣经在经过 PPM 压缩算法之后的大小十分接近，可以认为压缩之后的大小实际体现了《圣经》本身包含的信息量，即体现了压缩算法能有效地取出各种语言中的冗余信息，而中文版的压缩比最小，说明中文的信息冗余量最少，可以间接地反映出中文的信息熵更大。另外，作者又采用了其他的压缩算法和其他文本，均得到了相同的结论。

当然，有人可能觉得将英文字母与汉字进行比较不太公平，因为汉字显然从笔画来说比英文字母更加复杂，公平起见，应该比较英文单词和中文词的信息熵。香农估计每个单词的信息熵为 11.82 bit（也有研究认为大约是 9.7 bit [4]）。但是，问题在于对英文按单词划分是自然地，但是如何对中文进行词划分呢？（之前在《人工智能导论》中使用python 的 jieba 库进行中文分词，但感觉效果很差）事实上，找到经过人工划分的准确的中文分词大规模语料库可能是解决这个问题的关键。

另外，可以看到无论采用哪种说法，都可以得到单词信息熵大于字母信息熵的结论，我们可以定性地进行分析：假设字符集中每个字符出现的频率相同，那么字符集中字符的种类增加，会使字符的信息熵增大，证明如下：

$$\Omega = -\sum_{i=1}^{n} p_i \log_2 p_i = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$$

可以看到，每个字符出现的频率相同的情况下，字符的信息熵会随着字符集中字符种类的增大而增大，而英文单词的个数远大于英语字母的个数，这是英文单词的信息熵大于英语字母的信息熵的一部分原因。

在字符种类一定的情况下，可以考虑字符频率分布对信息熵的影响。因为 $-p\log_2 p$ 是一个上凸函数，根据琴生不等式：

$$\Omega = -\sum_{i=1}^{n} p_i \log_2 p_i \leq -n\overline{p} \log_2 \overline{p}$$

其中 $\overline{p}$ 表示 $p_i$ 的平均数，而由于字符种类一定，因此：

$$\overline{p} = \frac{1}{n} \sum_{i=1}^{n} p_i = \frac{1}{n}$$

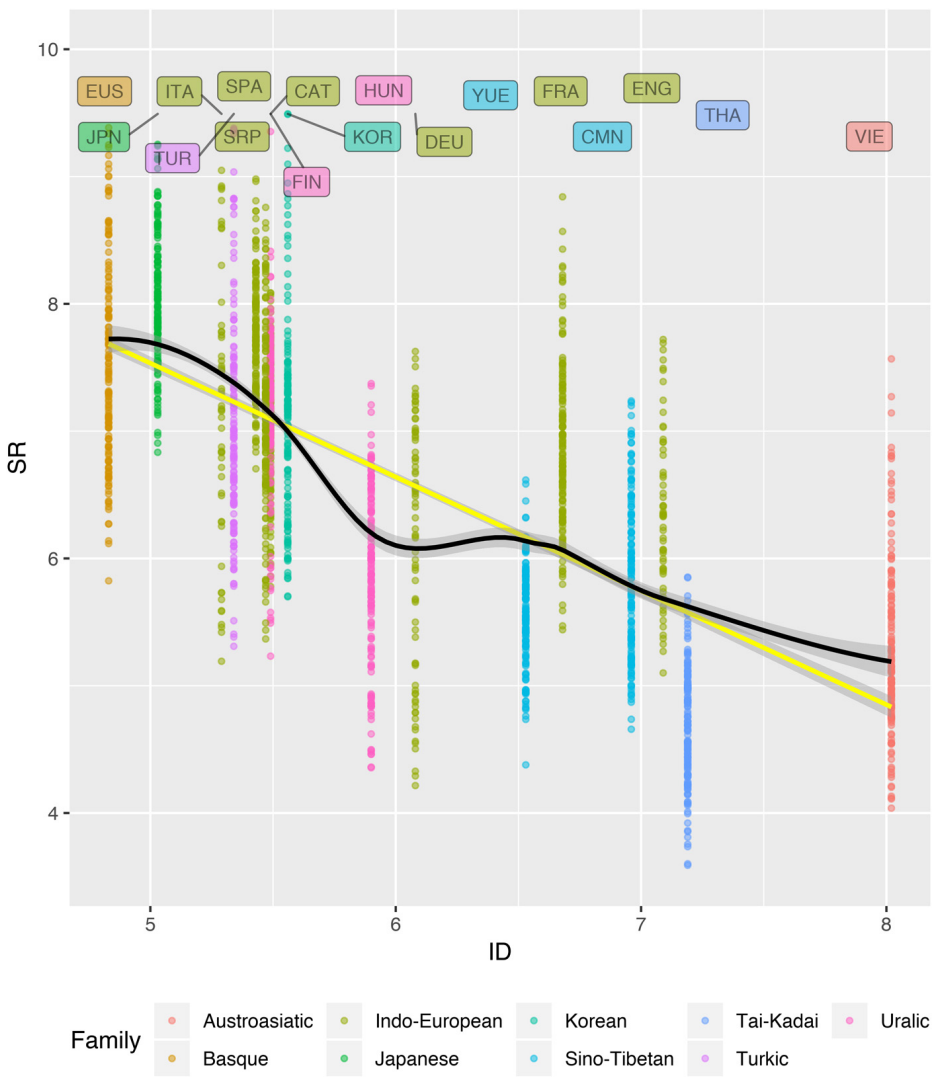
即：

$$\Omega \leq \log_2 n$$

当且仅当 $p_1 = p_2 = \ldots = p_n$ 时等号成立，即字符频率分布得越均匀，字符的信息熵越大，字符分布得越不均匀，字符的信息熵越小。这也是容易理解的，考虑极端的情况，如果某个字符的频率分布为 0，其他字符频率均匀分布，这实际上退化成了字符种类数为 n-1 的情况，根据前面的结论，信息熵自然会降低。

文字可能总是会存在上述的问题使得不同文字无法进行统一公平的比较，而语音也许天然不存在这样的问题，不同语音可以以音节为单位进行划分，事实上，我们可以公平地比较不同语言每个音节的信息熵。

最近也有研究 [5]比较了各个语言的语音输出速率，得出了很有意思的结论，即信息密度大（单音节信息熵大）的语言往往语速很慢，信息密度小的语言平均语速很快，而最终语音的带宽，即每秒传递的信息量大致相同，大约为 39.15bit/s。从下图也可以看出，每个音节的信息熵与语速在一定程度上成负相关，相关系数为 -0.71。



总的来说，经过本次实验，我更加充分地理解了信息熵的概念，同时通过查阅资料了解了其他研究者在这方面取得的成果，此实验拓宽了我的知识面，增强了我的实践能力，再次感谢老师和助教的悉心指导。

## 四、参考文献。


[1] Shannon, Claude Elwood. "A mathematical theory of communication." *The Bell system technical journal* 27.3 (1948): 379-423.

[2] Wong, K., and R. Poon. "A comment on the entropy of the Chinese language." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.6 (1976): 583-585.

[3] Behr Jr, Frederic H., et al. "Estimating and comparing entropy across written natural languages using PPM compression." (2002).

[4] Grignetti, Mario C. "A note on the entropy of words in printed English." *Information and Control* 7.3 (1964): 304-306.

[5] Coupé, Christophe, et al. "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche." *Science advances* 5.9 (2019): eaaw2594.