基于 n-Gram 模型的中文平均信息熵计算

刘张辰宇
ZY2203703

Abstract

由 16 本武侠小说作为中文语料库，通过 python 的正则化匹配进行数据清洗之后，使用 py 自带的 jieba 分词库进行精准分词，分别对一元、二元、三元组计算中文平均信息熵，并将频数最高的前十个词可视化显示出来。

Introduction

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。公式如图：

$$H(X) = \sum_{x \epsilon X} P(x) \log\left(\frac{1}{P(X)}\right) = -\sum_{x \epsilon X} P(x) \log(P(x))$$

信息论之父克劳德·香农给出的信息熵的三个性质：

1. 单调性，发生概率越高的事件，其携带的信息量越低；
2. 非负性，信息熵可以看作为一种广度量，非负性是一种合理的必然；
3. 累加性，即多随机事件同时发生存在的总不确定性的量度是可以表示为各事件不确定性的量度的和，这也是广度量的一种体现。

本次实验首先通过栈 stack 进行深度优先搜索，确保拿到每一个.txt 语料文件，然后对所有的语料文件即.txt 文件通过正则匹配的方式，清除除了中文以外其他字符，并写入一个新的.txt 文件中。

在 1-Gram 模型中，通过 jieba 库中 lcut 函数分词后结合 Counter 函数，直接统计每个词的词频和总词数，再通过一元模型的信息熵公式，计算得到中文平均信息熵。

在 2-Gram 及 3-Gram 模型中，在进行分词之后，需要将词每两个组合或者每三个组合起来，得到新的语料库，再分别统计二元和三元词的词频及总词频算出先验概率再利用 split 函数和 Counter 后的语料库结合 dict 的表示键值方法，算出条件概率，得到中文平均信息熵。

Methodology

本次实验主要把中文分为三个语言模型分别计算平均信息熵，分为 1-gram、2-gram、3-gram。马尔科夫假设[1]：随意一个词出现的概率只与它前面出现的有限的 k 个词有关。K=0

即为 1-gram 模型。

M1: 1-Gram

当 k = 0 时，这个时候对应的模型叫做一元模型，即 wi 与它前面的 0 个词相关，即 wi 不与任何词相关，每一个词都是相互独立的，P(W)计算如下：

$$P(\omega 1\omega 2\ldots\omega n) = \prod_i P(\omega i)$$

M2: 2-Gram

当 k = 1 时，这个时候对应的模型叫做二元模型，即 wi 与它前面的 1 个词相关，P(W)计算如下：

$$P(\omega 1\omega 2\ldots\omega n) = \prod_i P(\omega 1)P(\omega 2|\omega 1)P(\omega 3|\omega 2)\ldots P(\omega n|\omega n-1)$$

M3: 3-Gram

当 k = 2 时，这个时候对应的模型叫做三元模型，即 wi 与它前面的 2 个词相关，P(W)计算如下：

$$P(\omega 1\omega 2\ldots\omega n) = \prod_i P(\omega 1)P(\omega 2|\omega 1)P(\omega 3|\omega 2,\omega 1)\ldots P(\omega n|\omega n-1,\omega n-2)$$

Experimental Studies

首先将 16 本小说进行数据清洗，去除掉其中除了中文汉字以外的字符，并将清洗之后的内容存入.txt 文件中。

对清洗之后的语料库，分别用三个语言模型计算中文的平均信息熵，

Table 1: Entropy of three models

	词库总词数	不同词的个数	信息熵
1-Gram	4291369	172243	12.178
2-Gram	4232111	1943313	6.929
3-Gram	4171332	3460538	2.300

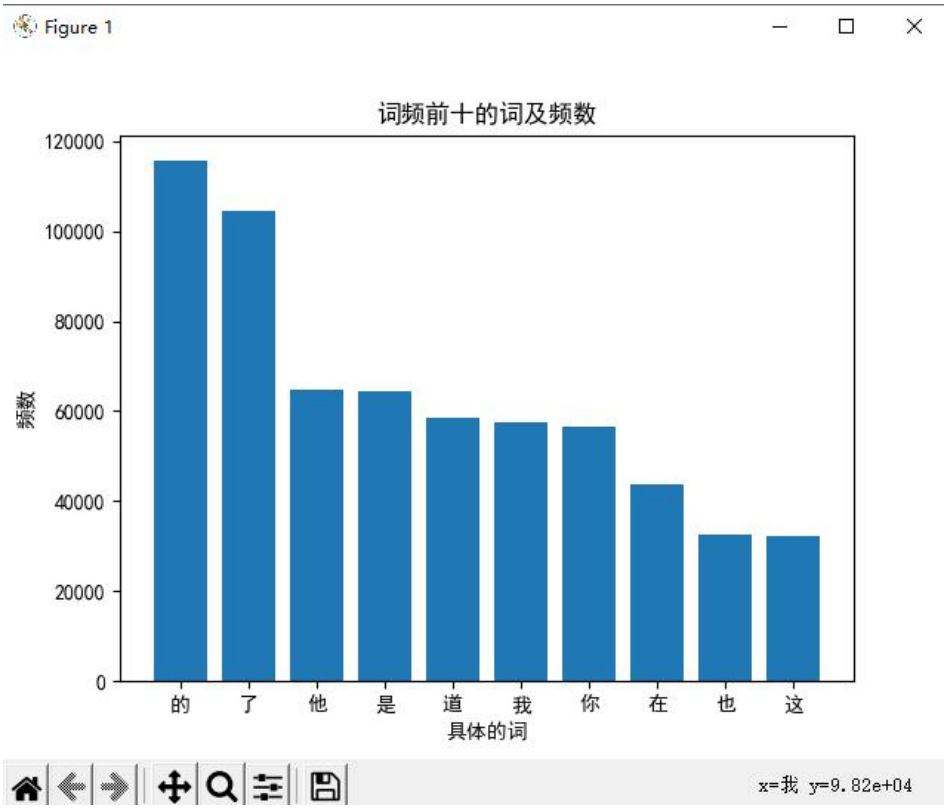

Figure 1：1-Gram 中频次前十的词

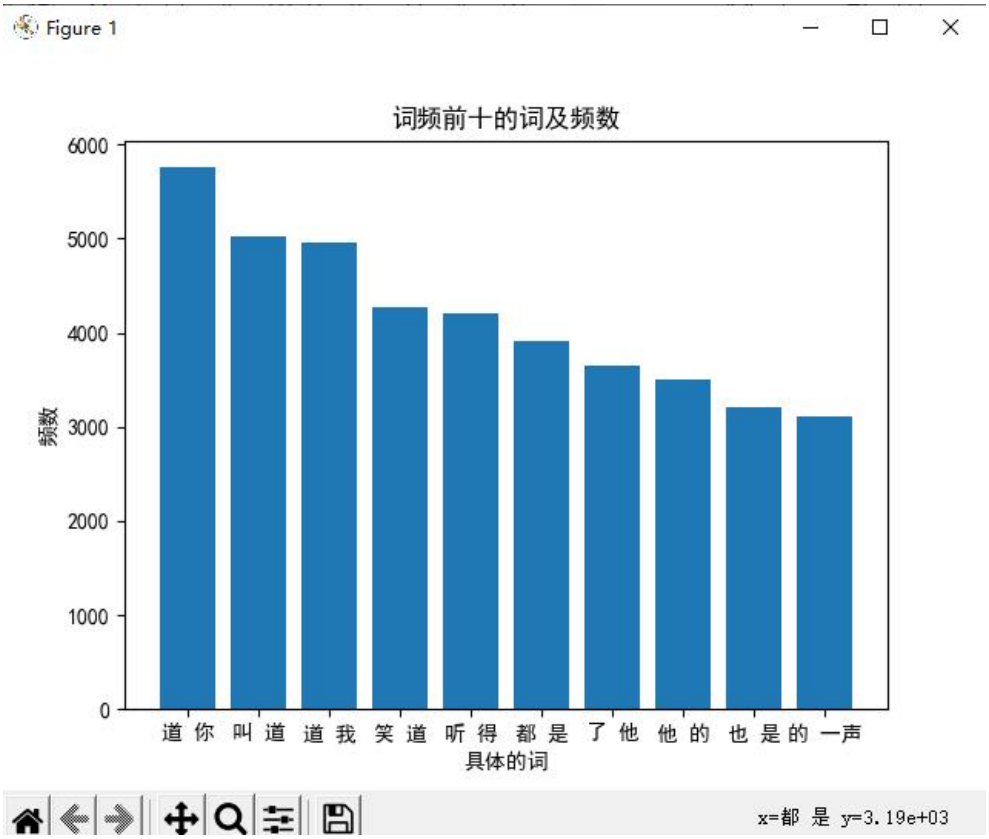

Figure 2：2-Gram 中频次前十的词

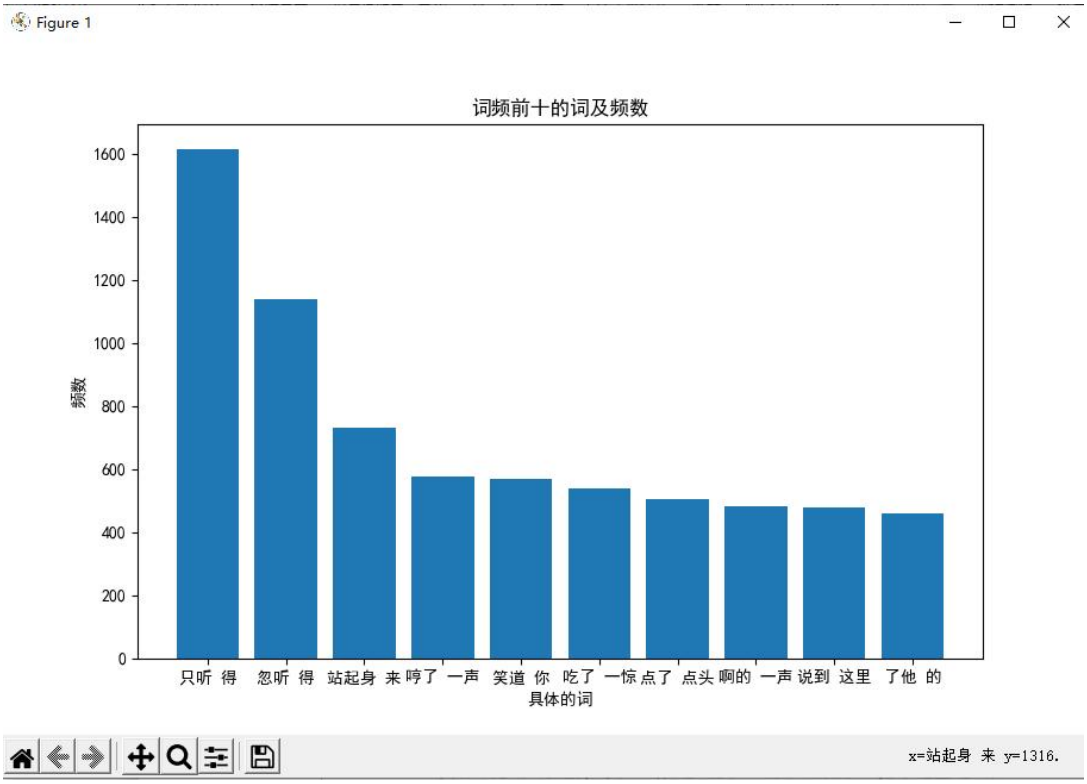

Figure 3：3-Gram 中频次前十的词

Conclusions

三元模型的中文平均信息熵最小，一元模型的中文平均信息熵最大。符合常理：一元模型字较少，携带的信息量小，不确定性高，故信息熵最大；三元模型字较多，甚至可以猜测出句子意思，携带的信息量大，不确定性低，故信息熵最小。

对比三种模型可以看出，N 取值越大，其对应的信息熵越小，因为考虑的前后文越多，词组合的种类越少，即文章变得越有序。

References

[1]Brown P F, Della Pietra S A, Della Pietra V J, et al. An estimate of an upper bound for the entropy of English[J]. Computational Linguistics, 1992, 18(1): 31-40.