北京航空航天大學
BEIHANG UNIVERSITY

自然语言处理作业一

平均信息熵的计算

学院名称	自动化科学与电气工程学院
学生学号	ZY2103809
学生姓名	王海腾
指导教师	秦曾昌

2022 年 4 月

一、任务描述

1、 阅读文章《An Estimate of an Upper Bound for the Entropy of English》

2、 分别以词和字为单位计算平均信息熵

二、实验原理

1、 信息熵

信息熵，是 1948 年 C.E.Shannon（香农）从热力学中借用过来提出的概念，解决了对信息的量化度量问题。C. E. Shannon 在 1948 年发表的论文“通信的数学理论（A Mathematical Theory of Communication ）”中指出，任何信息都存在冗余，冗余大小与信息中每个符号（数字、字母或单词）的出现概率或者说不确定性有关。Shannon 借鉴了热力学的概念，把信息中排除了冗余后的平均信息量称为“信息熵”，并给出了计算信息熵的数学表达式。

1） 一元模型信息熵

$$H(x) = -\sum_{x\in\chi} P(x)\, log\, P(x)$$

其中$P(x)$可近似为每个字或词在语料库中出现的概率

2） 二元模型信息熵

$$H(X|Y) = -\sum_{x\in\chi} P(x,y)\, log\, P(x|y)$$

其中联合概率$P(x,y)$可近似为每个二元字组或词组在语料库中出现的频率，条件概率$P(x|y)$可近似等于每个二元词组在语料库中出现的频数与以该二元词组的第一个词为词首的二元词组的频数的比值。

3） 三元模型信息熵

$$H(X|Y,Y) = -\sum_{x\in\chi} P(x,y,z)\, log\, P(x|y,z)$$

其中联合概率$P(x,y,z)$可近似等于每个三元词组在语料库中出现的频率，条件概率$P(x|y,z)$可近似等于每个三元词组在语料库中出现的频数与以该三元词组的前两个词为词首的三元词组的频数的比值。

2、 统计语言模型

假定 S 表示某个有意义的句子，由一连串特定顺序排列的词ω_1，ω_2，ω_3，…. ω_n 组成。其中 n 代表句子的长度。现在我们想知道 S 在文本中出现的可能性，即：

$$p(s) = p(\omega_1, w_2, w_3, w_4 \cdots w_n)$$

利用条件概率公式：

$$p(\omega_1, w_2, w_3, w_4 \cdots w_n) = p(\omega_1)p(\omega_1|\omega_2)\cdots p(\omega_n|\omega_1, \omega_2, \cdots \omega_n)$$

当计算$p(\omega_1)$，仅存在一个参数；计算$p(\omega_1|\omega_2)$，存在两个参数，此后越往后累积的，计算难度加大。所以马尔可夫提出一种假设：假设ω_i出现的概率只与前面 N-1 个词相关，当 N=2 时，就是二元模型，N=3 就是三元模型，本次实验分别使用一元模型、二元模型，三元模型来统计语料库字数，分词个数，平均词长，模型长度，基于模型的中文信息熵和运行时间。

三、实验过程与结果分析

1、 数据预处理

1） 删除无用字符

数据集为金庸先生的 16 本小说，其中包含了大量乱码与无用或重复的中英文符号，因此需要对该实验数据集进行预处理。删除所有的隐藏符号、非中文字符、标点符号。

删除的字符如下图所示：

```
r1 = u'[a-zA-Z0-9!"#$%&\'()*+,-./:;<=>?「」@，。?★、…【】《》？“”‘’！[\\]^_`{|}~]+'
```

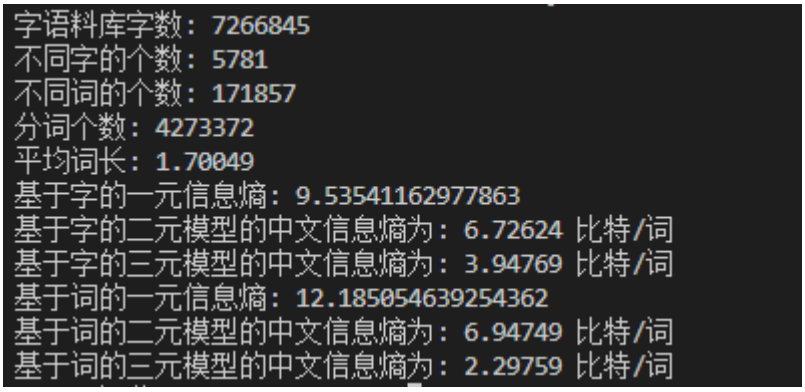
2） Jieba 分词

此外，由于实验要求分别以字和词为基准对信息熵进行计算，在以词为基准进行计算时，因此还需要对文本进行分词处理，本实验采用不含重复的精确模式进行分词。

本实验将 16 本小说合并在一起进行处理，通过上面两种数据处理方式，分别形成以字和词为组成的语料库。

2、 实验结果

划分方式		字	词
分词个数		7266845	4273372
平均词长		1	1.70049
不同字词个数		5781	171857
信息熵	一元模型	9.5354	12.1851
	二元模型	6.7262	6.9475
	三元模型	3.9477	2.2976


字语料库字数：7266845
不同字的个数：5781
不同词的个数：171857
分词个数：4273372
平均词长：1.70049
基于字的一元信息熵：9.53541162977863
基于字的二元模型的中文信息熵为：6.72624 比特/词
基于字的三元模型的中文信息熵为：3.94769 比特/词
基于词的一元信息熵：12.185054639254362
基于词的二元模型的中文信息熵为：6.94749 比特/词
基于词的三元模型的中文信息熵为：2.29759 比特/词