中文平均信息熵的计算

一 实验要求

1. 阅读文章 *Entropy_of_English_PeterBrown*。

2. 计算中文（分别以词和字为单位的）平均信息熵。

二 实验原理

2.1 信息熵

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。

香农给出了满足上述三个条件的随机变量不确定性度量函数，即信息熵

$$H(X) = -\sum_{x\in\chi} p(x)\log p(x)$$

2.2 N-gram 模型

N-Gram 是一种基于统计语言模型的算法。它的基本思想是将文本里面的内容按照字节进行大小为 N 的滑动窗口操作，形成了长度是 N 的字节片段序列。每一个字节片段称为 gram，对所有 gram 的出现频度进行统计，并且按照事先设定好的阈值进行过滤，形成关键 gram 列表，也就是这个文本的向量特征空间，列表中的每一种 gram 就是一个特征向量维度。

该模型基于这样一种假设，第 N 个词的出现只与前面 N-1 个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。这些概率可以通过直接从语料中统计 N 个词同时出现的次数得到。

$$p(w_1,w_2,\ldots,w_m) = p(w_1) * p(w_2|\ w_1) * p(w_3|\ w_1,w_2)\ldots p(w_m|\ w_1,\ldots,w_{m-1})$$

利用马尔科夫链的假设，即当前这个词仅仅跟前面几个有限的词相关，因此也就不必追溯到最开始的那个词，这样便可以大幅缩减上述算式的长度。当 n=1, 一个一元模型（unigram model)即为：

$$P(w_1,w_2,\cdots,w_m) = \prod_{i=1}^{m} P(w_i)$$

当 n=2, 一个二元模型（bigram model)即为：

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i | \ w_{i-1})$$

当 n=3, 一个三元模型（trigram model)即为：

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i | \ w_{i-2} w_{i-1})$$

三 实验过程

3.1 语料库及预处理

采用金庸的 16 本小说作为中文的语料库。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。包括删除非中文字符，删除标点符号等

为了删除标点符号，定义了标点表。除了常用的逗号、句号等之外，添加了部分数据集中的特殊符号，例如：• ．— * - ~ ﹏ |等。

分词采用 jieba 库。jieba 是 python 中的一个中文分词库，在本实验中以精确模式进行分词。

3.2 信息熵的计算

一元模型的信息熵计算公式在 2.1 节中已经给出，二元模型和三元模型的信息熵计算采用条件熵公式。条件熵 $H(X | \ Y)$ 表示在已知随机变量 X 的条件下随机变量 Y 的不确定性。随机变量 Y,Z 给定的条件下随机变量 X 的条件熵如下：

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

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

四 实验结果

4.1 以词为单位

采用 jieba 分词后，模型的输出便是以词为单位的中文平均信息熵。程序输出如下：

表 1 以词为单位的中文平均信息熵

语料库字数	一元分词个数	二元词个数	三元词个数
7420081	4430767	4430751	4430735
平均词长	一元平均信息熵	二元平均信息熵	三元平均信息熵
1.67	12.01312 比特/词	6.8915 比特/词	2.41661 比特/词

4.2 以字为单位

未经过分词，逐字输入模型，得到的结果为以字为单位的平均信息熵。程序的输出结果如下：

表 2 以字为单位的平均信息熵

语料库字数	一元字数	二元字数	三元字数
7420081	7420081	4430751	4430735
语料库字符串数量	一元平均信息熵	二元平均信息熵	三元平均信息熵
16	9.50211 比特/字	6.68521 比特/字	3.94294 比特/字

五 实验总结与收获

对比一元、二元、三元三种语言模型得到的结果可以发现，随着 N 取值逐渐增大，考虑到的前后文关系越多，通过分词后得到的文本中词组的分布就越简单，文本的信息熵则越小。这是因为通过前后文关系减小了不确定性。因为 N-gram 模型都是采用马尔可夫假设近似估计，阶数越高，考虑到的前后文的语义信息越多，能更好的评价这种语言的信息熵。从这个角度而言，三元模型对中文信息熵的估计要相比于一元和二元模型更准确。

参考文献

1. *Entropy_of_English_PeterBrown*
2. *CSDN 博客：https://blog.csdn.net/weixin_42663984/article/details/115718241?spm=1001.2101.3001.6650.3&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-3.pc_relevant_default&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-3.pc_relevant_default&utm_relevant_index=6*
3. *知乎专栏：https://zhuanlan.zhihu.com/p/32829048*
4. *《数学之美》吴军 2014-11 ISBN: 9787115373557*