深度学习与自然语言处理报告

19376252 张一辉
tbgsection@qq.com

摘要

本文是深度学习与自然语言处理的第一次实验报告，实验内容为以字和词计算中文的信息熵。实验使用若干本小说作文实验数据，去除停用词，分别以单个汉字和词语作为单位，建立一元，二元，三元信息熵模型进行计算并且将所有小说作为总和进行计算用以估计中文的信息熵。最终计算结果为字的一元、二元和三元模型的信息熵为 9.94，7.04，3.49，词的一元、二元和三元模型的信息熵为 13.58，6.51，1.17。

引言

信息熵是信息理论中的一个重要概念，用于描述一个随机变量的不确定性或者信息量大小。简单来说，信息熵可以理解为一个信源所产生信息的平均不确定度。香农是信息论的奠基人之一，他提出了信息熵这一重要概念，香农认为[1]，一个消息的信息量应该与它的不确定度成正比，信息熵就是一个随机变量中不确定度的量度。信息熵越大，代表这个随机变量越难以预测，包含的信息量也就越大。与英文或其他语言的信息熵概念相似，中文的信息熵是用来衡量中文文本信息量的一个指标。在中文中，一个汉字或一个词语的出现往往具有一定的概率，而文本中所有汉字或词语的出现概率分布就构成了中文文本的概率分布函数，通过计算这个概率分布函数的信息熵，可以衡量中文文本的信息量大小和不确定性程度。中文的信息熵可以用于文本分类、信息检索、语言模型等多个领域，对于中文自然语言处理和文本分析具有重要的意义。

研究方法

M1：信息熵

依据 Boltzmann's H-theorem，香农把随机变量X的熵值 H=定义如下，其值域为$\{x_1, \ldots, x_n\}$:

$$H(X) = E[I(X)] = E[-ln(P(X)]$$

其中，P为X的概率质量函数（probability mass function），E为期望函数，而$I(X)$是X的信息量（又称为自信息）。$I(X)$本身是个随机变数。

当取自有限的样本时，熵的公式可以表示为：

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

其中$H(X)$表示随机变量X的信息熵，$P(x)$表示X取某个值的概率，log_2表示以2为底的对数。

通过以上计算得到的信息熵的单位通常是比特（bit），表示信息的量度。如果使用以e为底的对数，则信息熵的单位是纳特（nat）

M2: 联合信息熵

针对于联合分布的随机变 $(X,Y)\sim P(X,Y)$在两变量相互独立的情况下, 联合信息熵为

$$\begin{aligned} H(X\mid Y) &= -\sum_{y\in Y} P(y)\log\left(P(x\mid y)\right) \\ &= -\sum_{y\in Y} P(y)\sum_{x\in X} P(x)\log\left(P(x\mid y)\right) \\ &= -\sum_{y\in Y}\sum_{x\in X} P(x)P(y)\log\left(P(x\mid y)\right) \\ &= -\sum_{y\in Y}\sum_{x\in X} P(x,y)\log\left(P(x\mid y)\right) \end{aligned}$$

在计算信息熵时，可以将文本分成多个部分来计算信息熵。在多元模型中，文本被划分成连续的多个字符或符号组成的元组，每个元组被视为一个符号，它们的出现概率也被计算在内，最终通过信息熵公式计算整个文本的信息熵。多元模型可以根据需要将文本分成任意长度的元组，例如二元模型、三元模型等等。在处理自然语言文本时，多元模型可以更好地反映语言中的多元组合规律，比如在处理语音识别、机器翻译等任务时，多元模型可以更好地表达语言中的上下文信息。根据联合信息熵的定义得到

二元模型计算公式

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

三元模型计算公式

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

实验内容

算法实现

首先对小说进行数据处理，首先读取 inf 中的数据，作为目录，然后依次读取若干本小说的内容，先将其中的小说网站广告文本进行删除。对于词的模型，使用 jieba 库进行分词处理。根据 cn_stopwords.txt 文件中的停词，将小说内容中的停词去除，最终只保留中文文本。

对于处理过后的文本，使用字典进行词频统计，二元三元模型，将多个词组合为 tuple 作为字典的关键字。

一元模型中，使用频率来作为概率直接计算信息熵，在二元模型中，需要计算联合概率和条件概率。二元词组出现频率作为联合概率$P(x,y)$，每个二元词组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也就是第一个词在一元模型中的频率作为条件概率$P(x|y)$。同理在三元模型中，三元词组出现频率作为联合概率$P(x,y,z)$，每个三元词组出现的频数与以该三元词组的第一个词为词首的三元词组的频数的比值也就是前两个词在二元模型中的频率作为条件概率$P(x|y,z)$。

数据	单位	一元模型信息熵	二元模型信息熵	三元模型信息熵
白马啸西风	字	9.215534064	4.089770663	1.210862213
	词	11.12675812	2.912274028	0.370373725
碧血剑	字	9.742677774	5.68115645	1.791198225
	词	12.88504353	3.96214504	0.430754846
飞狐外传	字	9.622388134	5.575071395	1.86224783
	词	12.62570653	4.040617992	0.460855889
连城诀	字	9.513325287	5.0897432	1.635239694
	词	12.20665339	3.5889817	0.368526835
鹿鼎记	字	9.648360073	6.028930564	2.406060346
	词	12.63224504	4.992705021	0.838050026
三十三剑客图	字	10.00488254	4.283369148	0.648845169
	词	12.53445375	1.808929425	0.090861795
射雕英雄传	字	9.737210579	5.971596791	2.195657722
	词	13.03556051	4.592478038	0.537740386
神雕侠侣	字	9.671899588	6.075427115	2.306325262
	词	12.74049196	4.683850597	0.636332042
书剑恩仇录	字	9.746843059	5.606040785	1.859098323
	词	12.72735664	4.136925337	0.497420902
天龙八部	字	9.780061057	6.116624423	2.346231806
	词	13.01890385	4.838749625	0.663325731
侠客行	字	9.434686273	5.379789251	1.814752404
	词	12.28736023	3.992604985	0.512543379
笑傲江湖	字	9.507843339	5.863099455	2.358041223
	词	12.52345978	4.838958475	0.795512539
雪山飞狐	字	9.496108838	4.803896539	1.299899326
	词	12.05652502	3.065747927	0.290698042
倚天屠龙记	字	9.701321808	5.987764861	2.273393645
	词	12.89063608	4.684704451	0.644051491
鸳鸯刀	字	9.209841965	3.65516177	0.895289609
	词	11.09772869	2.156200112	0.2431723
越女剑	字	8.779834855	3.108865843	0.843588321
	词	10.46726936	1.738085301	0.243693724
所有小说	字	9.939989535	7.040655571	3.494664992
	词	13.58132256	6.516666497	1.175154281

表 1: 计算结果

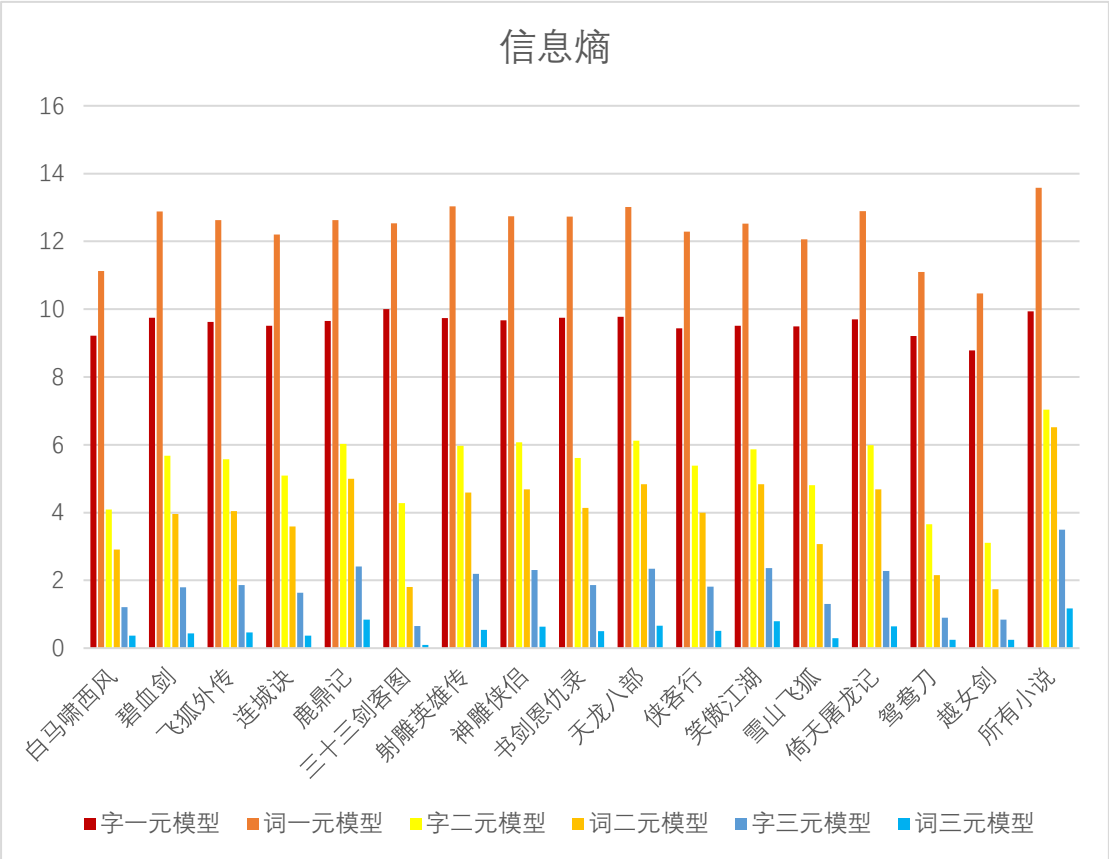

图 1：计算得到的信息熵

结论

经过计算我们得到了中文的一元信息熵为 9.4bit,参考国家语言文字研究所给出的 9.6bit 的结果，二者十分接近，本次实验计算结果是有效的。相比于英文 1.75bit 的信息熵，中文的信息熵远远高于英文，含有的信息量更多，同样的内容表述，需要的中文长度小于英文，这与我们实际生活中的感觉一直。

在计算结果中可以看出，无论是字还是词，一元模型的熵大于二元模型，而二元模型大于三元模型。随着元数增加，字词组的排列变得有序，不确定度下降，减小了本文的信息熵也减少。

References

[1]Shannon, C. E . A Mathematical Theory of Communication[J]. Bell Systems Technical Journal, 1948, 27(4):623-656.

[2]陈原.现代汉语定量分析[M].上海.上海教育出版社,1989:267

[3] Mori S , Yamaji O . An Estimate of an Upper Bound for the Entropy of Japanese[J]. Ipsj Journal, 1997, 38:2191-2199.