北京航空航天大學

BEIHANG UNIVERSITY

深度学习与自然语言处理作业1

院（系）名称 自动化科学与电气工程学院

学生姓名 潘翔

学生学号 ZY2103707

2022 年 4 月

一、作业内容

首先阅读文章：Entropy_of_English_PeterBrown 链接：https://docs.qq.com/pdf/DUUR2Z1FrYUVqU0ts 作业内容：参考上面的文章来计算中文(分别以词和字为单位)的平均信息熵。

二、原理

信息熵是 1948 年 C.E.Shannon(香农)从热力学借用过来的高年，解决了对信息的量化度量问题。信息是个很抽象的概念。人们常说信息很多，或者信息较少，但却很难说清除信息到底有多少。比如一本五十万字的中文书到底有多少信息量。

一般而言，当一种信息出现概率更高的时候，表明它被传播得更广泛，或者说，被引用得程度更高。我们可以认为，从信息传播得角度来看，信息熵可以表示信息得价值。这样就有一个衡量信息价值高低得标准，可以做出关于知识流通问题得更多推论。

其计算公式可以表示为

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

在该次试验中，$P(x)$可以近似等于每个词在语料库中出现的频率。

而对于二元和三元模型，联合分布得随机变量

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

联合分布用于计算文献中和课程中提到的二元模型(bigram)和三元模型(trigram)

则二元模型的计算公式为

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

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

三元模型的计算公式为

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

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。

三、分析过程与结果

3.1 处理数据

在提供的数据库中，有大量不能计入的文本符号，以老师提供的鹿鼎记为例。

对数据库进行一系列的预处理。小说开头存在“本书来自…”的广告用语，需要对齐进行删除处理。对数据库中文本的熵进行计算，直接将所有的非中文文本进行删除，包括标点符号，特殊符号比如#$%^&*等，回车符号“\n”，英文空格” ”，中文半角空格”\u3000”，最后得到的结果是纯中文文字的小说，不包括其他的内容。

使用《三十三剑客图》作为样本，经过处理后得到的结果，在 debug 中查看如下所示

对于以字为单位进行处理，直接对得到的结果按照字符进行分析，对于词处理，使用 jieba 库的 jieba.cut 方法对数据进行分词处理，便于后续熵的计算。

3.2 试验方法和结果

进行数据库预处理

试验结果

直接对字进行处理得到的结果如下所示

	小说名称	小说字数	一元模型(比特/字)	二元模型(比特/字)	三元模型(比特/字)
1	白马啸西风	59291	8.872	4.531	1.634
2	碧血剑	416581	9.46	5.865	2.348
3	飞狐外传	375501	9.313	5.753	2.405
4	连城诀	194534	9.174	5.398	2.156
5	鹿鼎记	1024246	9.295	5.987	2.948
6	三十三剑客图	53443	9.675	4.835	0.981
7	射雕英雄传	772746	9.452	6.054	2.754
8	神雕侠侣	827742	9.351	6.016	2.835
9	书剑恩仇录	435816	9.472	5.778	2.389
10	天龙八部	1021425	9.407	6.125	2.948
11	侠客行	309779	9.154	5.591	2.369
12	笑傲江湖	824733	9.208	5.897	2.87
13	雪山飞狐	117167	9.172	5.119	1.773
14	倚天屠龙记	818398	9.395	6.021	2.804
15	鸳鸯刀	30444	8.98	4.183	1.151
16	越女剑	13649	8.823	3.642	0.911
17	16本小说总体	7295495	9.539	6.723	3.944

利用 jieba 库进行分词之后，按照词进行计算得到结果如下所示

	小说名称	小说词数	平均词长	一元模型(比特/词)	二元模型(比特/词)	三元模型(比特/词)
1	白马啸西风	37020	1.602	10.07	3.972	0.878
2	碧血剑	242920	1.715	11.748	4.998	0.992
3	飞狐外传	221216	1.697	11.528	4.996	1.053
4	连城诀	117249	1.659	11.043	4.692	0.955
5	鹿鼎记	605041	1.693	11.448	5.753	1.62
6	三十三剑客图	31331	1.706	11.68	2.948	0.273
7	射雕英雄传	456735	1.692	11.807	5.467	1.289
8	神雕侠侣	494887	1.673	11.584	5.493	1.452
9	书剑恩仇录	253409	1.72	11.72	5.027	1.038
10	天龙八部	604370	1.69	11.737	5.677	1.48
11	侠客行	183261	1.69	11.189	4.958	1.127
12	笑傲江湖	482378	1.71	11.398	5.62	1.527
13	雪山飞狐	71064	1.649	10.913	4.134	0.844
14	倚天屠龙记	474956	1.723	11.761	5.518	1.328
15	鸳鸯刀	18342	1.66	10.243	3.155	0.586
16	越女剑	8047	1.696	10.069	2.529	0.327
17	16本小说总体	4302226	1.696	12.168	6.94	2.312

四、 结果分析

1. 对于字和词两种计算方法比较可以看出，利用词进行计算信息熵较小，利用字计算的信息熵较大。
2. 无论是用字还是词进行计算都可以看出，信息熵的大小与文本总字数有关，文本总字数越小，信息熵越小，显然越多的字的混乱程度大概率是大于较少字数的文本的，所以这个结果也表明字数较多的文本混乱程度较大。
3. 从一元模型到二元模型再到三元模型可以看出信息熵的绝对数值是下降的，可以预见对于多元模型，信息熵的绝对数值会进一步下降，所以三元模型足够刻画信息熵问题。
4. 根据计算结果可以看出，中文信息熵的大小远远大于英文，回答了上课时提出的问题，中文包含的信息量大于英文。

五、 参考资料

https://blog.csdn.net/weixin_42663984/article/details/115718241
https://blog.csdn.net/qq_37098526/article/details/88633403