中文平均信息熵计算

刘峻池
2252298@qq.com

摘要

本实验以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理，即删除了停词、英文大小写字母、广告等冗余内容，并进行了两组分别以“词”和“字”为单位的实验，在实验中建立了一元、二元、三元语言模型，用以语料库进行词频统计与信息熵计算，并对于数据结果进行了图形化表示与分析。

一、介绍

中文信息熵是对汉字、词语或文本中信息量的度量。信息熵通常用于评估一段文本中的信息量大小。中文信息熵的计算基于信息论，它是基于概率论的一个数学模型，通过对中文语言的统计分析，可以计算出中文语言中的信息熵。

一元模型指的是以单个字符为单位来计算信息熵，比如对于一段文本，我们可以统计每个汉字出现的频率，并据此计算每个汉字的信息熵。二元模型和三元模型则是以连续的两个或三个字符为单位来计算信息熵，因为中文语言中的词语通常由多个汉字组成，这种方法可以更好地反映出中文语言的结构和规律。

中文信息熵的特点在于，中文语言中的汉字和词语数量较多，因此相对于英文等其他语言，中文的信息熵更高，包含更多的信息。中文信息熵的应用非常广泛，例如，在自然语言处理和文本挖掘领域，可以用中文信息熵来度量文本的信息量、预测下一个字符或词语等。此外，在网络安全和加密等领域，中文信息熵也被广泛应用于密码学中，例如对于密码强度的评估和生成等。

总的来说，中文信息熵是一种非常有用的工具，它可以帮助我们更好地理解中文语言中的信息结构和规律，并在实际应用中发挥重要作用。

本实验主要完成了以下任务：

(1) 读取语料库数据并进行预处理，包括删除非必要词汇、非法符号、广告等冗余内容。

(2) 分别建立了一元(unigram)、二元(bigram)以及三元(trigram)语言模型

(3) 利用上述语言模型对于语料库中的中文武侠小说进行信息熵计算

(4) 对于中文信息熵结果进行分析与总结

二、方法原理

1、数据预处理

本实验中所作的数据预处理内容有：

(1) 删除了各个小说中源网站的广告内容

(2) 删除了文章内的停词，即标点以及常见语气助词等无实意的字或词。

(3) 删除了文章内的英文内容，如大小写英文字母等

(4) 仅保留了常见汉字（包括简体与繁体）与常见标点符号

2、N-gram语言模型

N-Gram是一种基于统计语言模型的算法。其基本思想是将文本中的内容按照大小为N的滑动窗口进行操作，形成了长度为N的字节片段序列，这些字节片段被称为gram。

通过对所有gram的出现频度进行统计，并按照预先设定的阈值进行过滤，可以形成关键gram列表，这个列表就是文本的向量特征空间，其中每一种gram都是一个特征向量维度。

在语料库中，可以将其中的内容 $X=\{...X_{-2},X_{-1},X_0,X_1,X_2...\}$ 视为一个长度有限的平稳随机过程[1]，而为了计算语料库的整体信息熵，则需要计算每个词汇按照顺序出现的概率，此概率的计算表示如下：

$$P(X_1,X_2,...,X_n)=P(X_1)P(X_2\mid X_1)\cdots P(X_n\mid X_{n-1},...,X_1)$$

该模型基于这样一种假设，及第n个词的出现只与前n-1个词相关，与其它任何词都无关，整句出现的概率就是各个词汇出现概率的乘积。在实际应用中常常用到的为一元、二元、三元语言模型。

一元组模型表示如下：

$$P_1(X_1,X_2,...,X_n)=P_1(X_1)P_1(X_2)\cdots P_1(X_n)$$

二元组模型表示如下：

$$P_2(X_1,X_2,...,X_n)=P_2(X_1)P_2(X_2\mid X_1)\cdots P_2(X_n\mid X_{n-1})$$

三元组模型表示如下：

$$P_3(X_1,X_2,...,X_n)=P_3(X_1,X_2)P_3(X_3\mid X_2,X_1)\cdots P_3(X_n\mid X_{n-1},X_{n-2})$$

3、中文信息熵计算

中文信息熵是对汉字、词语或文本中信息量的度量。信息熵通常用于评估一段文本中的信息量大小。中文信息熵的计算基于信息论，它是基于概率论的一个数学模型，通过对中文语言的统计分析，可以计算出中文语言中的信息熵。在本实验中涉及三类中文信息熵计算公式表示如下[2]。

一元模型的信息熵计算公式为：

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

其中 $P(x)$ 可近似等于每个词在语料库中出现的频率。

二元模型的信息熵计算公式为：

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

其中联合概率 $P(x,y)$ 可近似等于每个二元词组在语料库中出现的频率，条件概率 $P(x\mid y)$ 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。

三元模型的信息熵计算公式为：

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

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。

三、实验结果与分析

本实验使用Python进行编程实现，以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理，即删除了停词、英文大小写字母、广告等冗余内容，并进行了两组实验，分别以“词”和“字”为单位进行一元、二元、三元语言模型的词频统计与信息熵计算，数据结果展示如下。

语料库名称	语料库字数	分词个数	一元信息熵	二元信息熵	三元信息熵
白马啸西风	38665	23660	9.6939	3.2055	1.2826
碧血剑	264214	142779	12.7813	3.7685	0.4903
飞狐外传	235563	128234	12.577	3.7997	0.5006
连城诀	121096	68018	11.763	3.425	0.6583
鹿鼎记	638632	356699	12.4071	4.7183	1.0059
三十三剑客图	34840	18793	12.4065	1.6778	0.0953
射雕英雄传	504901	281170	12.5238	4.4519	0.91
神雕侠侣	514133	284420	12.8306	4.4629	0.6834
书剑恩仇录	280096	151048	12.692	3.9374	0.5008
天龙八部	641450	354643	12.7915	4.5383	0.8555
侠客行	192798	107285	11.8919	3.7633	0.7755
笑傲江湖	503848	277483	12.4287	4.5694	0.8644
雪山飞狐	71720	39849	11.7493	2.8896	0.5019
倚天屠龙记	521179	286408	12.7154	4.4541	0.7788
鸳鸯刀	20651	12317	9.8125	2.6332	0.9722
越女剑	9971	6038	9.0832	2.2824	0.9917

表格1：以“词”为单位的实验结果

语料库名称	语料库字数	分词个数	一元信息熵	二元信息熵	三元信息熵
白马啸西风	38665	38665	8.6333	4.1374	1.604
碧血剑	264214	264214	9.7324	5.6538	1.8116
飞狐外传	235563	235563	9.5988	5.5488	1.8831
连城诀	121096	121096	9.3799	5.0218	1.6954
鹿鼎记	638632	638632	9.5351	5.9504	2.458
三十三剑客图	34840	34840	10.0047	4.2761	0.6578
射雕英雄传	504901	504901	9.5916	5.9184	2.2995

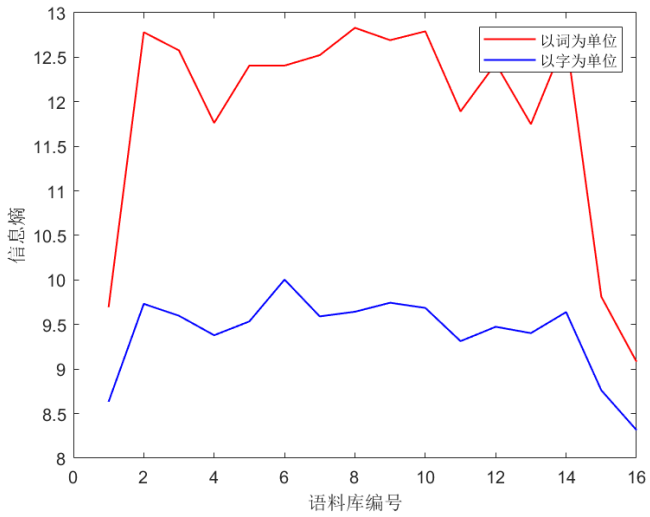
神雕侠侣	514133	514133	9.644	6.0383,	2.3229
书剑恩仇录	280096	280096	9.7447	5.5859	1.8711
天龙八部	641450	354643	9.6865	6.0369	2.3815
侠客行	192798	107285	9.3136	5.2986	1.8647
笑傲江湖	503848	277483	9.475	5.8221	2.3700
雪山飞狐	71720	39849	9.4037	4.7598	1.3662
倚天屠龙记	521179	286408	9.6402	5.9458	2.3094
鸳鸯刀	20651	12317	8.7623	3.7575	1.2387
越女剑	9971	6038	8.3157	3.245	1.1659

表格2：以“字”为单位的实验结果

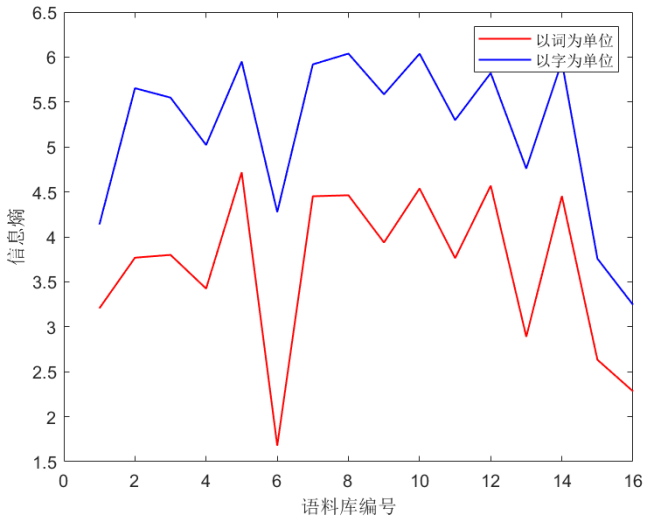
由上述实验结果可以得出：

1、N-gram模型中的N对语料库的信息熵有着至关重要的影响。当N更大时，对下一个词出现的约束性信息更多，词组间考虑的前后文关系越详尽，词组分布越简单，同时信息熵会较小；而当n更小的时候，在训练语料库中出现的次数更多，但是约束信息更少，信息熵会较大。

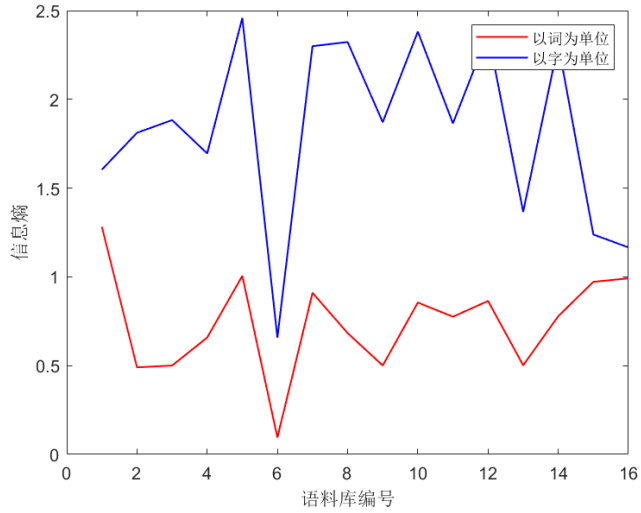
2、以“词”为单位时要比以“字”为单位时的分词个数更多，因为字是词的组成部分；在一元模型中，以“词”为单位时的信息熵要比以“字”为单位时的更大，而在二元、三元模型中以“词”为单位时的信息熵要比以“字”为单位时的更大。分析得在统计中文信息熵时，在一元模型中以“字”为单位统计信息熵更小，而在N≥2时，以词为单位统计信息熵更小。



图像1：以“词”和“字”为单位的一元模型信息熵对比图



图像2：以“词”和“字”为单位的二元模型信息熵对比图



图像3：以“词”和“字”为单位的三元模型信息熵对比图

四、结论

本实验以作家金庸的16本中文小说作为语料库，分别对每本小说进行了数据预处理，即删除了停词、英文大小写字母、广告等冗余内容，并进行了两组分别以“词”和“字”为单位的实验，在实验中建立了建立了一元、二元、三元语言模型，用以语料库进行词频统计与信息熵计算，并对于数据结果进行了图形化表示与分析。

在本次实验中我不禁更深刻的了解了一些自然语言模型，同时还锻炼了编写代码的能力，了解了jieba等新的Python第三方库,此次实验使我受益匪浅。

参考文献

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

[2] 中文信息熵的计算 https://blog.csdn.net/qq_37098526/article/details/88633403