深度学习与自然语言处理第一次作业——中文信息熵

姓名：郑梓岳 学号：19231191

一. 作业内容

首先阅读文章：Entropy_of_English_PeterBrown，参考该文章来计算中文（分别以词和字为单位）的平均信息熵。

二. 信息熵

信息熵的概念最早由香农（1916-2001）于 1948 年借鉴热力学中的“热熵”的概念提出，旨在表示信息的不确定性。熵值越大，则信息的不确定程度越大。其数学公式可以表示为：

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

信息熵有三个性质：

（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。

三. N 元语言模型

3.1 语言模型

对于自然语言相关的问题，比如机器翻译，最重要的问题就是文本的序列有时候不是符合我们人类的使用习惯，语言模型就是用于评估文本序列符合人类语言使用习惯程度的模型。当前的语言模型是以统计学为基础的统计语言模型，统计语言模型是基于预先人为收集的大规模语料数据，以真实的人类语言为标准，预测文本序列在语料库中可能出现的概率，并以此概率去判断文本是否“合法”，是否能被人所理解。

3.2 N-Gram 语言模型

语言模型就是用来计算一个句子的概率的模型，也就是判断一句话是否是人话的概率。给定一个句子（词语序列）：

$$S = W_1, W_2, \ldots, W_K$$

它的概率可以表示为：

$$P(S) = P(W_1, W_2, \ldots, W_K) = p(W_1)P(W_2|W_1)\ldots P(W_K|W_1, W_2, \ldots, W_{K-1})$$

其中$P(W_1)$表示第一个词W_1出现的概率；$P(W_2|W_1)$是在已知第一个词的前提下，第二个词出现的概率；以此类推。

显然，当句子长度过长时（比如该作业的数据库为金庸的十六本小说，一本有几万字），$P(W_K|W_1, W_2, \ldots, W_{K-1})$的可能性太多，无法估算。俄国数学家马尔可夫假设任意一个词W_i出现的概率只同它前面的词W_{i-1}有关，使$P(S)$变为：

$$P(S) = P(W_1)P(W_2|W_1)P(W_3|W_2)\ldots P(W_i|W_{i-1})\ldots P(W_n|W_{n-1})$$

其对应的统计语言模型就是二元模型。

也可以假设一个词由前面 N-1 个词决定，即 N 元模型。当 N=1 时，每个词出现的概率与其他词无关，为一元模型，对应 S 的概率变为:

$$P(S) = P(W_1)P(W_2)P(W_3)\ldots P(W_i)\ldots P(W_n)$$

当 N=3 时，每个词出现的概率与其前两个词相关，为三元模型，对应 S 的概率变为:

$$P(S) = P(W_1)P(W_2|W_1)P(W_3|W_1, W_2)\ldots P(W_i|W_{i-2}W_{i-1})\ldots P(W_n|W_{n-2}, W_{n-1})$$

四． 计算小说平均信息熵过程

4.1 中文数据预处理

本次作业的数据是金庸的十六本小说，包括白马啸西风，碧血剑，飞狐外传，连城诀，鹿鼎记，三十三剑客图，射雕英雄传，神雕侠侣，书剑恩仇录，天龙八部，侠客行，笑傲江湖，雪山飞狐，倚天屠龙记，鸳鸯刀，越女剑。

但是，由于下载资源来自于网络，其中包含不少非小说内容，比如“本书来自 www.cr173.com 免费 txt 小说下载站更多更新免费电子书请关注 www.cr173.com”等，因此要对预料进行清洗，即删除无关内容。

其次，为了进行分词，需要将清洗完毕的预料进行符号过滤，即删除一切符号和空格。

4.2 分词

分词的目的是采用不同语言模型计算文本的信息熵。本次作业采用了 3 种模型，包括一元模型（以字为单位），二元模型，三元模型。

分词的方法为，遍历一次文本，通过所设定的模型，将文本总字符串分成多个子字符串。

4.3 词频统计

将上述分词后的子字符串分别进行词频统计，方便进行概率的计算。

4.4 计算中文信息熵

通过计算的每个分词的频率，带入信息熵公式，计算出 4 种模型下的中文信息熵。

五． 实验结果

1. 实验结果展示

实验的部分输出如下图所示：

具体实验结果如下表所示：

作品名称	一元信息熵	二元信息熵	三元信息熵
三十三剑客图	9.6682	4.8354	0.9830
书剑恩仇录	9.4602	5.7830	2.3938
侠客行	9.1525	5.5912	2.3692
倚天屠龙记	9.3933	6.0211	2.8052
天龙八部	9.4034	6.1251	2.9499
射雕英雄传	9.4345	6.0635	2.7625
白马啸西风	8.9091	4.5836	1.6211
碧血剑	9.4458	5.8693	2.3534

作品名称	一元信息熵	二元信息熵	三元信息熵
神雕侠侣	9.3726	6.0630	2.8463
笑傲江湖	9.2063	5.8976	2.8709
越女剑	8.8239	3.6424	0.9105
连城诀	9.1710	5.3988	2.1570
雪山飞狐	9.2014	5.1636	1.7585
飞狐外传	9.3079	5.7540	2.4070
鸳鸯刀	9.0334	4.2154	1.1176
鹿鼎记	9.2811	5.9931	2.9532

2. 实验结果分析

对比 1-gram、2-gram、3-gram 三种语言模型得到的结果可以看到，随着 N 取值变大，文本的信息熵则越小，这是因为 N 取值越大，通过分词后得到的文本中词组的分布就越简单，N 越大使得固定的词数量越多，固定的词能减少由字或者短词打乱文章的机会，使得文章变得更加有序，减少了由字组成词和组成句的不确定性，也即减少了文本的信息熵，符合实际认知。