成 绩 _____

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理第一次作业

中文平均信息熵的计算

院（系）名称	自动化科学与电气工程学院
专 业 名 称	自动化
学 号	SY2103106
姓 名	段晓玥
指 导 教 师	秦曾昌

2022 年 4 月 7 日

一、任务描述

1. 阅读文章《An Estimate of an Upper Bound for the Entropy of English》；

2. 参考以上文章，分别以字和词为单位，在给定数据集上计算中文的平均信息熵。

二、实验原理

1. 信息熵的计算

“信息熵”的概念由香农在 1948 年首次引入到信息论之中，被用于衡量一个离散随机变量的不确定性。设X是一个离散的随机变量，则其信息熵被定义为：

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

熵值越大，表明信息的不确定程度越大。当 log 底数取为 2 的时候，对应信息熵单位为 bits。

2. 中文 jieba 分词

jieba 是基于 Python 的中文分词工具，包括“全模式”、“精确模式”、“搜索引擎模式”三种分词模式，默认为精确模式。其中精确模式分出来的词汇，前后不会有重复；而另外两种模式分词会有重复（如“中国科学院”可能被分为“中国”、“科学”、“学院”、“科学院”、“中国科学院”5 个词）。通过 jieba，可以将中文连续的文本自动分为多个中文词汇。本实验中采用不含重复的精确模式进行分词。

三、实验过程与结果

1. 数据集预处理

本次实验数据集为金庸的 16 本小说（存储于.txt 文档中）以及包含各小说题目的一个.txt 文档。为了更准确地计算中文平均信息熵，将所有.txt 文本中的乱码、无用的中英文符号进行去除。去除的符号主要包括：

```
char_to_be_replaced = "\n
`1234567890-=/*-~!@#$%^&*()_+qwertyuiop[]\\QWERTYUIOP{}|asdfghjkl;\'ASDFGHJKL:\"
zxcvbnm,./ZXCVBNM<>?~！@#￥%……&*（）——+【】：；“‘”《》？，。、★「」『』～ "
□ａｎｔｉ－ｃｌｉｍａｘ＋．／０１２３４５６７８９＜＝＞@ＡＢＣＤＥＦＧＨＩＪＫ
ＬＭＮＯＰＱＲＳＴＶＷＸＹＺ［＼］ｂｄｅｆｇｈｊｋｏｐｒｓｕｖｗｙｚ￣\u3000\x1a"
```

去除上述符号过后，剩余的文本只包含中文纯文字。将去除上述符号后的文本作为语料库X，分别以每个中文汉字或 jieba 分词后的每个中文词汇为单位，便

可基于字或词计算得到中文平均信息熵。具体地，设每个中文汉字或 jieba 分词后所得的每个中文词汇为x，且有$x \in X$，统计每个中文汉字或中文词汇在语料库X中出现的总次数为$t(x)$，语料库X包含的中文汉字总字数或中文词汇总数为N，则有x在语料库中出现的概率为$p(x) = \frac{t(x)}{N}$，代入$H(X) = -\sum_{x \in X} p(x)\log(p(x))$即可计算得到语料库的中文平均信息熵。

2. 实验结果

基于字和词计算中文平均信息熵实验结果如下表所示。

小说题目	总字数	分词总个数	平均词长	平均信息熵（基于字，bits）	平均信息熵（基于词，bits）	运行时间（s）
鹿鼎记	1019819	600568	1.698	9.28	11.45	7.41
鸳鸯刀	29272	17155	1.706	9.03	10.48	0.19
飞狐外传	375085	220795	1.699	9.31	11.53	2.42
雪山飞狐	113337	67217	1.686	9.20	11.11	0.74
连城诀	194487	117200	1.659	9.17	11.04	1.23
越女剑	13649	8047	1.696	8.82	10.07	0.08
笑傲江湖	824552	482192	1.710	9.21	11.40	5.31
神雕侠侣	810147	477261	1.697	9.37	11.73	5.75
碧血剑	415726	242049	1.718	9.45	11.74	2.64
白马啸西风	57422	35136	1.634	8.91	10.24	0.38
射雕英雄传	766608	450311	1.702	9.44	11.82	5.01
天龙八部	1021083	603994	1.691	9.40	11.73	6.64
倚天屠龙记	818206	474760	1.723	9.39	11.76	5.09
侠客行	309711	183192	1.691	9.15	11.19	2.03
书剑恩仇录	435615	253082	1.721	9.46	11.71	2.62
三十三剑客图	53285	31175	1.709	9.67	11.68	0.34
全部语料库	**7258068**	**4264158**	**1.702**	**9.53**	**12.18**	**47.52**