北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理
第一次大作业

平均信息熵计算

院（系）名称　自动化科学与电气工程学院

学　生　学　号　ZY2103810

学　生　姓　名　游虎杰

2022 年 4 月

一、问题描述

分别以字和词为单位，计算给定金庸小说内中文的平均信息熵。

二、问题表达

所需要用到的金庸小说共 16 本，分别如下：

1、白马啸西风

2、碧血剑

3、飞狐外传

4、连城诀

5、鹿鼎记

6、三十三剑客图

7、射雕英雄传

8、神雕侠侣

9、书剑恩仇录

10、天龙八部

11、侠客行

12、笑傲江湖

13、雪山飞狐

14、倚天屠龙记

15、鸳鸯刀

16、越女剑

三、具体算法实现

信息熵计算公式为

$$H(X) = -\sum_{x \epsilon X} p(x) \log p(x)$$

其中 X 为某一本金庸小说的内容，x 为单个字或词，具体计算通过 python 实现。在

python 中，利用 read 函数读取文本文件中的内容，返回为字符串，并依次统计每个字或词的出现次数，以此计算出现频率，进而计算信息熵。以词为单位计算信息熵时，需要对文章内容进行分词，本实验通过 python 中 jieba 模块的 lcut 函数实现分词。

四、运行结果

分别以字和词为单位，对各小说计算信息熵，结果如表 1 所示：

表 1 信息熵计算结果

小说名称	信息熵（保留四位有效数字）	
	以字为单位	以词为单位
白马啸西风	8.541	9.380
碧血剑	9.449	11.66
飞狐外传	9.307	11.44
连城诀	9.099	10.76
鹿鼎记	9.229	11.17
三十三剑客图	9.674	11.63
射雕英雄传	9.355	11.46
神雕侠侣	9.327	11.47
书剑恩仇录	9.474	11.66
天龙八部	9.357	11.49
侠客行	9.088	10.91
笑傲江湖	9.201	11.29
雪山飞狐	9.156	10.86
倚天屠龙记	9.366	11.57
鸳鸯刀	8.720	9.673
越女剑	8.468	9.213

五、个人总结和体会

通过这次作业，我对信息熵的理解有了进一步的加深，对信息熵的计算过程有了更为熟练的掌握。同时，在编写代码的过程中，我对 python 的应用水平有了进一步提升。

六、作业代码

https://github.com/youlll/DP-NLP1.git