DL-NLP 第一次作业

姓名：赵久昂 学号：SY2106313

一、背景介绍

1.1 信息熵

信息熵是一个数学上颇为抽象的概念，在这里不妨把信息熵理解成某种特定信息的出现概率（离散随机事件的出现概率）。一个系统越是有序，信息熵就越低；反之，一个系统越是混乱，信息熵就越高。信息熵也可以说是系统有序化程度的一个度量。

1.2 语言模型

语言模型（Language Model, LM）的作用是估计不同语句在对话中出现的概率，并且LM适用于许多不同的自然语言处理应用程序。概率语言建模的目标是计算单词序列的语句出现的概率，$P(W) = P(w_1, w_2, w_3 \dots w_n)$，比如一句话由4个单词组成：则第五个单词出现的概率为：$P(w_5|w_1, w_2, w_3, w_4)$。

1.3 马尔科夫特征的定义

如果随机过程的未来状态的条件概率分布（以过去和现在状态为条件）仅取决于前k个状态，而不取决于其之前的所有状态，则称随机过程具有马尔可夫特征。 具有此特征的过程称为 Markov 过程。对于语言模型而言，只需要给出前k个词，可以估计下一个词的概率。

1.4 N-gram 模型

基于马尔可夫假设，N-gram 可以被定义出来。N-gram模型中，一个单词的出现仅与前面 N-1 个单词的出现相关。其中比较常见的是：

unigram 模型：$P(w_1 w_2 \dots w_n) = P(w_1) * P(w_2) * \dots P(w_n)$

bigram 模型：$P(w_n|w_1 w_2 \dots w_{n-1}) = P(w_n|w_{n-1})$

trigram 模型：$P(w_n|w_1 w_2 \dots w_{n-1}) = P(w_n|w_{n-1}, w_{n-2})$

1.5 N-gram模型的信息熵计算

根据大数定理，在数据足够大的情况下，词或二元词组或三元词组出现的概率大致等于其出现的频率。

unigram模型：$H(x) = -\sum_x P(x) log P(x)$，其中P(x)可近似等于每个词在语料库中出现的频率。

bigram模型：$H(x) = -\sum_{x,y} P(x,y) log P(x|y)$，其中联合概率P(x,y)可近似等于每个二元词组在语料库中出现的频率，条件概率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。

trigram模型：$H(x|y,z) = -\sum_{x,y,z} P(x,y,z) log P(x|y,z)$，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。

二、实验过程

- 实验环境
  - 处理器 2GHz 四核 Intel Core i5
  - 内存 16GB 3733 MHz LPDDR4X
  - Python 3.9
  - jieba 0.42.1
- 实验数据

  包括《神雕侠侣》、《笑傲江湖》等小说文本
- 实验代码
  - file_reader.py 读取文件的程序，从 inf.txt 中读取文件名，根据文件名再读取文件
  - processor.py 计算信息熵的核心程序，通过 file_reader 中的文件读取程序，读取所需要分析的文本文件

    `get_split_words()`是用来获取分词的函数，可以分别以单词或字为单位进行分词；

    其中`get_tf() get_bigram_tf() get_trigram_tf()`分别是计算一元模型、二元模型、三元模型的词频概率函数，返回值的类型是 dict，其中 key 为词（一元、二元、三元模型），value 为出现的次数。这三个函数在计算不同模型的信息熵时可以复用；

    `deal_unigram_entropy()`是计算一元模型信息熵的函数；

    `deal_bigram_entropy()`是计算二元模型信息熵的函数；

    `deal_trigram_entropy()`是计算三元模型信息熵的函数；
  - result.csv 和 result.txt 是计算信息熵的结果
  - result_dealer.py 是辅助生成 csv 结果的处理程序

三、实验结果

3.1 以单词为单位（jieba分词）

	unigram信息熵（比特/词）	bigram信息熵（比特/词）	trigram信息熵（比特/词）
白马啸西风	8.773415031	4.138128512	1.508676464
碧血剑	10.36345844	5.242692949	1.781588812
飞狐外传	10.20084687	5.163360069	1.836992293
连城诀	9.705543793	4.802763432	1.677505731
鹿鼎记	9.947962595	5.500947898	2.390113456
三十三剑客图	10.28798966	4.014612674	0.810202737
射雕英雄传	10.27816105	5.48518298	2.172751281
神雕侠侣	10.2743398	5.594703445	2.200186237
书剑恩仇录	10.27482916	5.22793403	1.871915881
天龙八部	10.23742854	5.54510398	2.285695169
侠客行	9.838637678	4.9346472	1.83444377
笑傲江湖	10.046642	5.46083263	2.290682333
雪山飞狐	9.81141657	4.453765496	1.351748995
倚天屠龙记	10.33255276	5.498096713	2.156801993
鸳鸯刀	8.90430126	3.596918157	1.132563676
越女剑	8.595582525	3.160859072	0.923157603

3.2 以字为单位

	unigram信息熵（比特/字）	bigram信息熵（比特/字）	trigram信息熵（比特/字）
白马啸西风	8.240228592	4.235945214	1.85920083
碧血剑	9.031029381	5.523719232	2.575269575
飞狐外传	8.907123167	5.390154949	2.592634474
连城诀	8.719280815	5.005453756	2.319599289
鹿鼎记	8.807657778	5.433554751	2.994002216
三十三剑客图	9.196565683	4.904299948	1.384907691
射雕英雄传	8.949406322	5.554137012	2.90169747
神雕侠侣	8.945241119	5.656242396	2.991144442
书剑恩仇录	9.009944952	5.415865608	2.609972598
天龙八部	8.939654175	5.583326578	3.013320266
侠客行	8.731471331	5.133635428	2.483569731
笑傲江湖	8.812343905	5.429717721	2.942797532
雪山飞狐	8.779887412	4.874408351	1.987181195
倚天屠龙记	8.968928105	5.56423498	2.919892557
鸳鸯刀	8.380225968	3.990032379	1.424869748
越女剑	8.17845427	3.53525215	1.182165123