任务2：语言模型

预备知识

在自然语言处理领域，语言模型（Language Model）用于计算和衡量一句话的概率，比如「我喜欢吃西瓜」这句话出现的概率大于「我喜欢吃西部」，后者不太自然。而语言模型最常用的算法是N-Gram算法，即将一句话的概率拆分为句子中每个词的概率:

$P(W) = P(w_1w_2...w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)...P(w_n|w_1w_2...w_{n-1})$

而为了计算简单与避免稀疏性等原因，常常采用马尔科夫假设，即当前词的出现只与它前k-1个词有关：

$P(W) = P(w_1w_2...w_n) = \prod P(w_i|w_{i-k+1}w_{i-k+2}...w_{i-1})$

上式我们称之为k-1阶马尔科夫链，或者k元语言模型。例如，1阶马尔科夫链 / 2元语言模型：

$P(W) = P(w_1w_2...w_n) = P(w_1) \prod P(w_i|w_{i-1})$

更多资料，参见：Stanford CS124 Stanford CS224d

任务说明

在训练集上训练一个语言模型，并在测试集上测试其性能。

建议方法

```
1.统计语言模型
2.神经语言模型
```

分别对应两份参考资料

评价指标

评价语言模型一般使用困惑度 (perplexity)： $perplexity = 2^{(-1/n)\sum \log_2 P(w_i|w_{i-k+1}w_{i-k+2}...w_{i-1})}$

数字越低代表句子概率越大，结果越好。

数据说明

数据集：Penn Treebank (PTB)

training set：42068 sentences
development set：3370 sentences
test set：3761 sentences

注：vocab文件为数据集中按词频排序的词表，每行为<word, frequency> pair，截取前10k个，语料中剩余的低频词已被统一替换成 '<unk>'，数字被统一替换成 'N'

提示：若出于内存空间不足或其他原因，可以进一步缩小词表，将低频词全替换为统一标识如 'UNK'等。

提交内容

1. 实现代码
2. 实验报告，包括但不限于：
   2.1 完成思路及过程
   2.2 实验结果及分析
   2.3 存在问题及可能改进方向