自然语言处理 小作业一：语言模型 PPL 计算

* 姓名: 管仁阳　学号:519021911058　邮箱: guanrenyang@sjtu.edu.cn

1　公式推导

$$PPL = (\prod_{k=1}^{K} \frac{1}{P(\omega_k|W_{k-n+1}^{k-1})})^{\frac{1}{K}}$$

$$\log_{10} PPL = -\frac{1}{K}\sum_{k=1}^{K} \log_{10} P(\omega_k|W_{k-n+1}^{k-1})$$

其中 $P(\omega_k|W_{k-n+1}^{k-1})$ 使用回退法进行 smoothing:

$$P(\omega_k|W_{k-n+1}^{k-1}) = \begin{cases} P(\omega_k|W_{k-n+1}^{k-1}) & \text{if } W_{k-n+1}^{k} \text{ exists.} \\ P(\omega_k|W_{k-n+2}^{k-1}) \times back\_off(W_{k-n+1}^{k-1}) & \text{otherwise..} \end{cases}$$

2　PPL 计算

本题中的模型为 tri-gram 模型，取 $n = 3$；每条序列长度为 13，即 $k = 1, 2, \cdots, 13$：学号本身 12 位，加上序列结束符号 $< /s >$。

$$\log_{10} PPL = -\frac{1}{13}[\log_{10} P(\omega_2|\omega_1) + \sum_{k=3}^{13} \log_{10} P(\omega_k|\omega_{k-2}, \omega_{k-1})]$$

tri-gram 模型的回退法概率计算式如下:

$$P(\omega_k|W_{k-2}^{k-1}) = \begin{cases} P(\omega_k|\omega_{k-2}, \omega_{k-1}) & \text{if } W_{k-2}^{k} \ \exists. \\ P(\omega_k|\omega_{k-1}) \times back\_off(\omega_{k-2}, \omega_{k-1}) & W_{k-2}^{k-1} \ \exists \text{ and } W_{k-2}^{k} \ \neg\exists.. \\ P(\omega_k) \times back\_off(\omega_{k-1}) \times back\_off(\omega_{k-2}, \omega_{k-1}) & \text{Otherwise.} \end{cases}$$

由于篇幅限制，以下只列出了第一个序列的具体计算式

(1). **021033210023:** $\log_{10} PPL = 1.2166224866923079$, $PPL = 16.467303381304372$

$$\begin{aligned} \log_{10} PPL = -\frac{1}{13}[&(-1.610146 - 0.6270623) + (-0.8515801 + 0) + (-0.2787536) + \\ &(-0.4080372 - 0.9792967) + (-0.3646991) + (-1.879542 + 0.01406545) + \\ &(-1.359456 - 1.127541 + 0) + (-0.39794 + 0) + (-0.4080372 - 0.9792967) + \\ &(-0.7403627) + (-0.8515801 - 0.007770127) + \\ &(-0.9650043 - 1.218526 + 0.06262255) + (-0.8381492 + 0)] = 8.201149451928574 \end{aligned}$$

$$PPL = 158909350.2561911$$

(2). **019033910051:** $\log_{10} PPL = 1.1572288455384616$, $PPL = 14.362460442062682$
(3). **120033910006:** $\log_{10} PPL = 1.1695236418461539$, $PPL = 14.774869098909809$
(4). **120033910013:** $\log_{10} PPL = 1.0059370446153848$, $PPL = 10.13764419662622$