3 Language Models – MLE & Smoothing

1. (a) The probability $P(x)$ of an event x smoothed using Absolute Discount (AD) is:

$$P_{AD}(X) = \begin{cases} \frac{count(X)-\delta}{N} & if\ count(X) > 0 \\ \\ \frac{(B-N_0)\delta/N_0}{N} & otherwise \end{cases}$$

Derive an Absolute Discount smoothing formula for the conditional probability of a trigram $P(z|xy)$, such that when no counts of order n are available, conditional probability of $(n-1)$-gram is recursively used, also with AD smoothing.

(b) The probability $P(x)$ of an event x smoothed using Linear Discount (LD) is:

$$P_{LD}(X) = \begin{cases} \frac{(1-\alpha)count(X)}{N} & if\ count(X) > 0 \\ \\ \frac{\alpha}{N_0} & otherwise \end{cases}$$

Derive a Linear Discount smoothing formula for the conditional probability of a trigram $P(z|xy)$, such that when no counts of order n are available, conditional probability of $(n-1)$-gram is recursively used, also with LD smoothing.

2. (a) Complete the python program `mle.py` to estimate via MLE the parameters of a character trigram model, and write them to a file.

(b) Complete the program `generate.py` to generate a random sequence of characters consistent with the loaded trigram model.

(c) Run the program `smooth.py` and enter different input sentences. Discuss why some sentences have zero probability. Modify the program `smooth.py` to perform a simple smoothing via Lidstone's or Laplace's Law. Discuss the values chosen for N and B.

3. (a) Extend the program mle.py to estimate the coefficients for a linear Interpolation smoothing. Write the coefficients into the first line of the model file, followed by the trigram parameters.

Linear Interpolation: $P(z|xy) = \lambda_1 P(z) + \lambda_2 P(z|y) + \lambda_3 P(z|xy)$

Coefficient estimation via deleted interpolation:

λ_1=λ_2=λ_3=0
foreach trigram xyz with $count(xyz) > 0$
  depending on the maximum of the following three values:
    case $\frac{count(z)-1}{N-1}$: increment λ_1 by $count(xyz)$
    case $\frac{count(yz)-1}{count(y)-1}$: increment λ_2 by $count(xyz)$
    case $\frac{count(xyz)-1}{count(xy)-1}$: increment λ_3 by $count(xyz)$
normalize λ_1,λ_2,λ_3

(b) Extend the program `smooth.py` to load the Linear Interpolation coefficients in the first line of the file, and use them to smooth the trigram probabilitites. Compare the results with the smoothing obtained in the previous exercise.