Name: Shi Cun Course: CSE 498-013(NLP)

A1:

Firstly, we assume that the frenquency of each N-grams word is c*, ($c^* = (c+1)\frac{N_{x+1}}{N_x}$, N_x means the the number of n-grams appears x times). So $c^*(0) = \frac{N_1}{N_0}$, the probility of each N_0 is P_{GT}, P_{GT}(x=0) = $c^*(0) \div N = \frac{N_1}{N_0 * N}$. Therefore, the count of unseen tokens c*(w,v) = $P_{GT} * N_0 = \frac{N_1}{N}$.

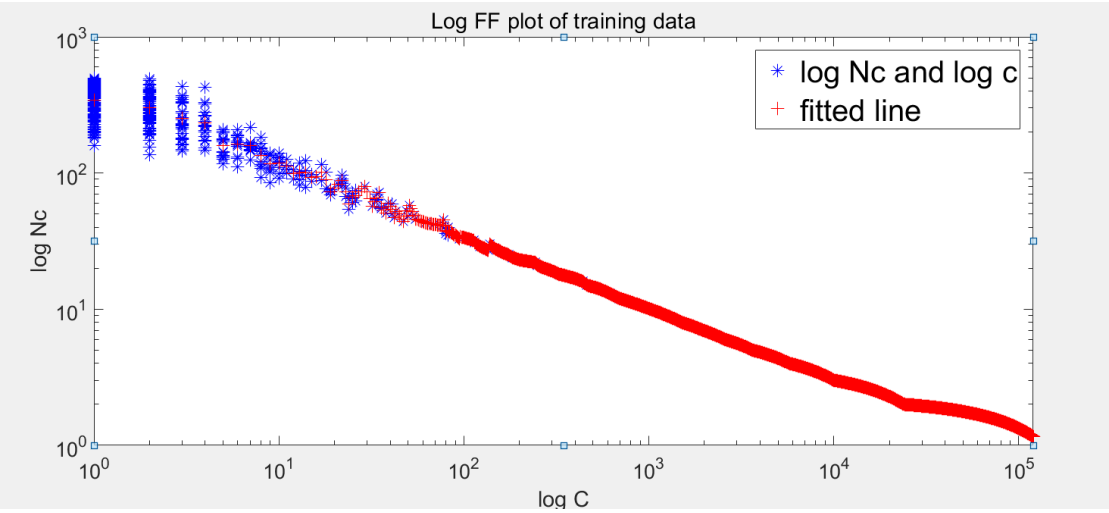
A2:

GT smoothing : c*(w,v) = $P_{GT} * N_0 = N_0 * \frac{c^*}{N} = \frac{N_1}{N_0} * \frac{N_0}{N} = \frac{N_1}{N}$

Laplacian smoothing : $c^*(w,v) = P_L(w,v) \times N = N * (c+1) \div (N + |V|)$

Because c = 0;

So $c^*(w,v) = \frac{N}{N+|V|}$



This figure is created by ff.txt, we can see that logNc and logC roughly following $\log N_c = a * \log C + b$.