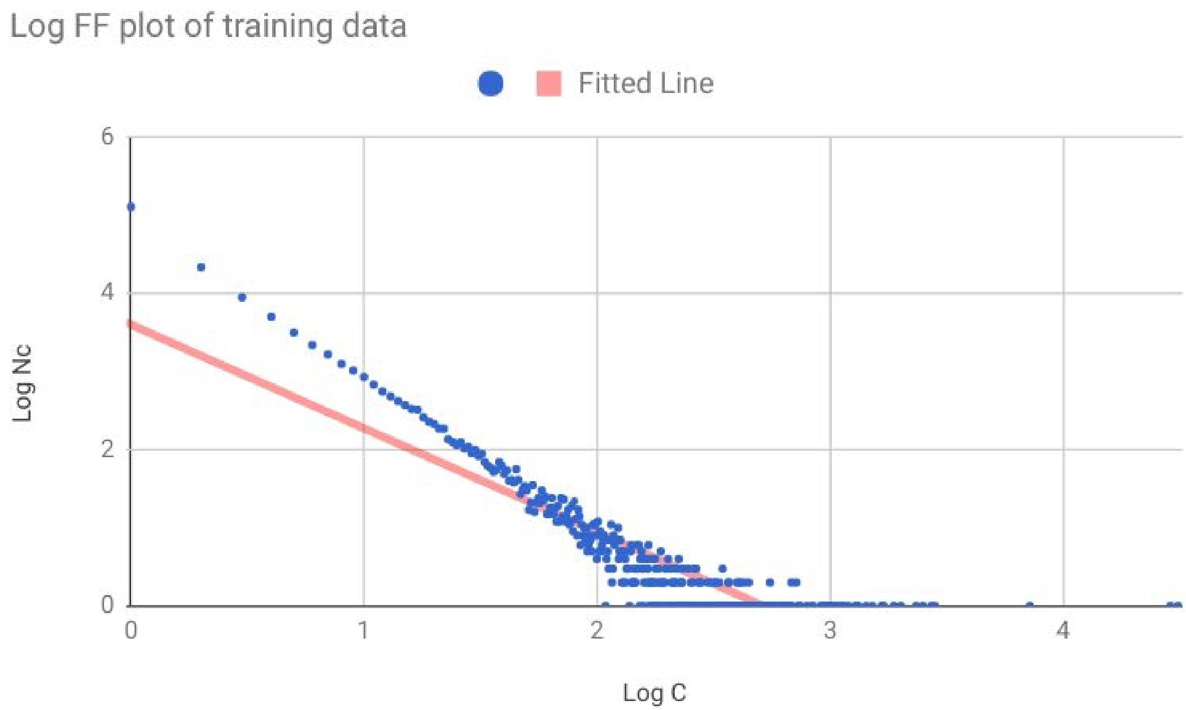
Billy DeLucia

CSE498 Spell Checker Project Report



Question 2:

A) $C^* = (C + 1) (N_{c+1} / N_c)$ => $C^*_0 = (0 + 1) (N_1 / N_0) = N_1 / N_0$

If we use the conventional calculation for C* we can then derive the N_1 / N estimation.

$\sum_{x:\ count(x) = 0} P(x) = \sum_1^{N0} [C^*_0 / N] = \sum_1^{N0} [(N_1 / N_0) / N] = \sum_1^{N0} [N_1 / (N * N_0)]$

$= N_1 / (N * N_0) * \sum_1^{N0}[1] = (N_1 / (N * N_0)) * N_0 = N_1 / N$

B) The zero mass M_0 is calculated by multiplying the probability given to a zero count token $P(X_0)$ and the number of zero count tokens N_0

For Good Turing smoothing this is calculated as such:

$P(X_0) = C^*_0 / N$

$M_0 = N_0 * (C^*_0 / N)$

For Laplacian smoothing:

$P(X_0) = 1 / (N + |V|)$

$M_0 = N_0 * (1 / (N + |V|))$

Question 3.1:

A) The number of unseen bigrams N_0 can be found by squaring the number of unique unigrams $|V_u|^2$ then subtracting the number of seen bigrams N.

$N_0 = |V_u|^2 - N$

I found the following

$|V_u| = 21,779$ $|V_u|^2 = 474,324,841$

N = 179,093

$N_0 = 474,145,748$

B) Good Turing:

$M_0 = N_0 * (C^*_0 / |V_u|^2)$ $C^*_0 = 0.18295146395320908$

$M_0 = 474,145,748 * (0.18295146395320908 / 474,324,841) = 0.18295146395320908$

$M_0 = \sim 18.3\%$

Laplacian:

$M_0 = N_0 * (1 / (N + |V_u|^2))$

$M_0 = 474,145,748 * (1 / (179,093 + 474,324,841)) = 0.998168130850942$

$M_0 = \sim 99.8\%$