CSE 498 Natural Language Processing
Project 1 Fall 2017
Sachin Joshi

Section 2 Questions

• Show that it makes sense to set $C^*(w, v) = N_1/N$ for those unseen tokens ((w, v) with C(w, v) = 0)

$$\mathbf{C(w, v) = 0}$$
$$\mathbf{C^* = (C + 1) * (N_{C+1}/N_C)}$$
$$C^* = (0 + 1) * (N_{0+1}/N_0)$$
$$C^* = (1) * (N_1/N_0)$$
$$C^* = (N_1/N_0)$$
$$C^* = (N_1/N) \; (N \sim N_0)$$

• Calculate the probability mass reserved for the unseen tokens when GT smoothing is used, and compare the mass to the mass reserved when the Laplacian smoothing is used.

GT Smoothing:

$$\mathbf{C^* = (C + 1) * (N_{C+1}/N_C)}$$
$$\mathbf{C^*(w, v) = P_L(w, v) * N}$$
$$C = 0,$$
$$|V| = 21777 \text{ (number of unique unigrams)},$$
$$N = 597496 \text{ (number of bigrams)}$$
$$N_{C+1} = N_{0+1} = N_1 = 125462 \text{ (obtained from the file ff.txt)}$$
$$C^* = (0 + 1) * (125462/473640233)$$
$$\mathbf{C^* = 2.648888149 * 10^{-4}}$$
$$P_L(w, v) = C^*/N = (2.648888149 * 10^{-4})/597496$$
$$\mathbf{P_L(w, v) = 4.433315284 * 10^{-10}}$$
$$\mathbf{\text{Probability mass} = P_L(w, v) * |V|^2}$$
$$\mathbf{\text{Probability mass} = 4.433315284 * 10^{-10} * 21777^2}$$
$$\mathbf{\text{Probability mass} = 0.210244537}$$

Laplacian:

$$C(w, v) = 0$$
$$N \text{ (number of all tokens i.e., 2 consecutive tokens (bigrams))} = 597496$$
$$|V| = 21777$$
$$\mathbf{P_L(w, v) = (C(w, v) + 1)/(N + |V|)}$$
$$\mathbf{P_L(w,v) = (0 + 1)/(597496 + 21777) = 1.614796705 * 10^{-6}}$$

Probability mass = $P_L(w, v) * |V|^2$
Probability mass = $1.614796705 * 10^{-6} * 21777^2$
Probability mass = 765.7975223

Clearly the probability mass reserved for the unseen tokens is less when GT smoothing is applied as compared to probability mass when the Laplacian smoothing is applied.

Section 3.1 Data Exploration

• Find N_0, the number of zero frequency tokens (w, v).

Total number of unique unigrams (V) = 21777 (obtained from the file train_tokens.txt and does not include the words: <s> and </s>).

Total number of bigrams, N (obtained from the file train_tokens.txt) = 597496.

Number of zero frequency tokens (w, v) $N_0 = (|V|^2$ – Number of possible bigrams)
$N_0 = 21777^2 – 597496$
$N_0 = 474237729 – 597496 = 473640233$

• Find the probability mass reserved for the zero frequency tokens (w, v) in Laplacian and GT smoothings.

Laplacian:

C(w, v) = 0
N (number of all tokens i.e., 2 consecutive tokens (bigrams)) = 597496
|V| = 21777
$P_L(w, v) = (C(w, v) + 1)/(N + |V|)$
$P_L(w,v) = (0 + 1)/(597496 + 21777) = 1.614796705 * 10^{-6}$
Probability mass = $P_L(w, v) * |V|^2$
Probability mass = $(1.614796705 * 10^{-6}) * 21777^2$
Probability mass = 765.7975223

GT Smoothing:

$C^* = (C + 1) * (N_{C+1}/N_C)$
$C^*(w, v) = P_L(w, v) * N$
C = 0
$N_{C+1} = N_{0+1} = N_1$ = 125462 (obtained from the file ff.txt)
$C^* = (0 + 1) * (125462/(473640233))$
$C^* = 2.648888149 * 10^{-4}$
$P_L(w, v) = C^*/N = (2.648888149 * 10^{-4})/597496$
$P_L(w, v) = 4.433315284 * 10^{-10}$

Probability mass = $P_L(w, v) * |V|^2$
Probability mass = $4.433315284 * 10^{-10} * 21777^2$
Probability mass = 0.210244537

• Use the training data to find N_C for each C. Take the logarithm of both quantities and plot a figure.

The file ff.txt contains N_C for each C. The plot between log N_C and log C is shown below:

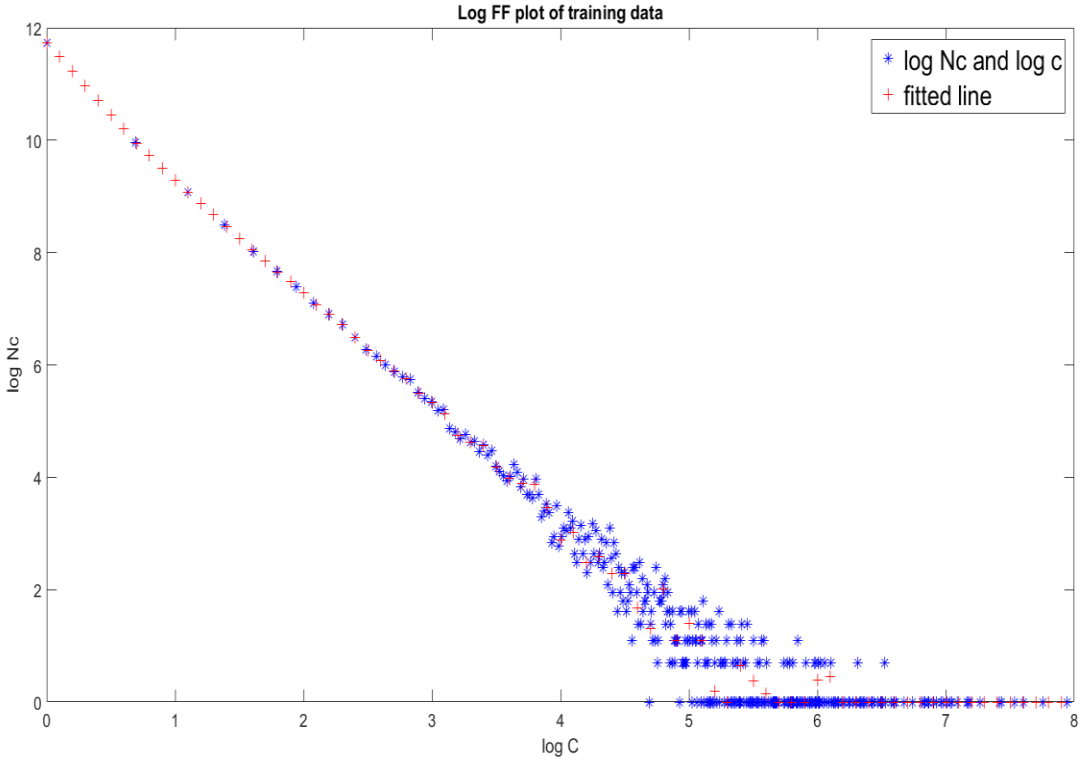

Figure 1: Plot of frequencies of frequencies in the log scale, with a line fitted to the points.