Natural Language Processing - Project 1
CSE 398/498-013
Name: Shihao Jing
Email: shj316@lehigh.edu

2 Questions
Answer the following two questions in your report
• Show that it makes sense to set C ∗ (w, v) = N1/ N for those unseen tokens ((w, v) with C(w, v) = 0).

The probability mass for unseen tokens is $\sum P(w,v) = \frac{N_1}{N_0 N} N_0 = \frac{N_1}{N}$.

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

Laplacian smoothing:

$$P(w,v) = \frac{C(w,v)+1}{N+|V|}$$

for $C(w,v) = 0$, $p(w,v) = \frac{1}{N+|V|}$

$$\sum P(w,v) = \frac{N_0}{N+|V|}$$

GT smoothing:

$$C_0 = \frac{N_1}{N_0}$$

$$p(w,v) = \frac{C_0}{N} = \frac{N_1}{N_0 N}$$

$$\sum P(w,v) = \frac{N_1}{N_0 N} N_0 = \frac{N_1}{N}$$

For data given, $V = 21779, N = 691075, N_0 = 474145749, N_1 = 126432$

Laplacian smoothing:

$$\sum P(w,v) = \frac{N_0}{N+|V|} = \frac{474145749}{691074 + 21779^2} = 0.998$$

GT smoothing:

$$\sum P(w,v) = \frac{N_1}{N} = \frac{126432}{691074} = 0.18$$

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